Using Surface-Learning to improve Speech Recognition with Lipreading

Christoph Bregler, Stephen Omohundro, Yochai Konig, and Nelson Morgan
{bregler,om,konig,morgan}@icsi.berkeley.edu

Computer Science Division
University of California
Berkeley, CA 94720

Int. Computer Science Institute
1947 Center Street
Berkeley, CA 94704

Abstract

We explore multimodal recognition by combining visual lipreading with acoustic speech recognition. We show that combining the visual and acoustic clues of speech improves the recognition performance significantly especially in noisy environment. We achieve this with a hybrid speech recognition architecture, consisting of a new visual learning and tracking mechanism, a channel robust acoustic front end, a connectionist phone classifier, and a HMM based sentence classifier. We focus in this paper on the visual subsystem based on "surface-learning" and active vision models. Our bimodal hybrid speech recognition system has already been applied to a multi-speaker spelling task, and work is in progress to apply it to a speaker independent spontaneous speech task, the "Berkeley Restaurant Project (BeRP)".

Summary

1 Introduction

Most efforts aiming toward robust speech recognition focus on methods that reduce the distortion of signals. Signal distortion 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, such as positional information about the visible articulators (lipmovements, tongue and teeth positions). It is well known that human speech perception is inherently bi-modal [11, 6].

Extending automated speech recognition to the visual modality has been investigated for quite some time. As popular non-connectionist approaches the work of Petajan, Bischoff, Bodoff, and Brooke [14], Mase and Pentland [10] should be mentioned. Recently Goldschen [7] 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 [18], who used static images for vowel discrimination. Wolff, Prasad, Stork, and Hennecke [16] are using a modified TDNN for isolated word segments, and Bregler, Hild, Manke, and Waibel [3] used a MS-TDNN extension for continuous word recognition.

We are interested in 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.

Our bimodal hybrid speech recognition system consists of a new visual learning and tracking technique, a channel invariant acoustic front end (RASTA-PLP), and a MLP/HMM speech recognizer. We focus in this paper on the visual subsystem, a new learning paradigm called

"surface learning" [4] and its application to an "active vision" tracking technique. Further we show how this new technique is integrated with an existing state-of-the-art acoustic recognition system. Finally we report performance measurements of the new combined system applied to our bimodal databases.

2 Visual Lip Processing

The first crucial task that has to be solved in our system is coding and tracking the configuration of lip positions of the talkers face.

2.1 Learning the Space of Lips

Our goal is to reduce the high dimensional sampled image data to low dimensional "lip-feature" vectors without losing relevant information. Knowing that lip positions are produced by some underlying muscle apparatus, the dimensionality of our "lip-configuration-space" should not be higher than the number of free parameters in our vocal tract. (The ultimate goal is to induce a mapping from our raw images to a space with dimensionality of exactly that parametric model.) As an example, imagine that each possible $n \cdot m$ image can be represented as a point in a $n \cdot m$ dimensional space. Take one specific lip-shape and change gradually the amount of mouth-opening. The corresponding point in the $n \cdot m$-dimensional space will move along a 1-dimensional curve embedded in the $n \cdot m$-dimensional image space. All possible modifications of the lip shape together will span a low dimensional nonlinear surface (or manifold) embedded in the high dimensional image space.

We use a new learning technique, which we call "Surface-Learning" [4] to induce this low dimensional subspace from high dimensional data. Once we have learned such a nonlinear surface, we can perform various different queries on it. The most important case for lip recognition is the nearest-point query. Given a new lip-image, we want to find the closest point on the surface in order to find the best matching "legal" point. Another interesting query is the completion query. Values of certain dimensions are unspecified (hidden areas in the image), so we can intersect the subspace of the unspecified dimensions with the learned surface and determine the specific value or range of the unknown dimensions. In section 2.3 we present another useful surface operation, the "interpolation task".

The surface learning approach itself starts from the observation that if the data points were drawn from a *linear* surface, then a principle components analysis could be used to discover the dimension of the linear space and to find the best-fit linear space of that dimension. The largest principle vectors would span the space and there would be a precipitous drop in the principle values at the dimension of the surface. A principle components analysis will no longer work, however, when the surface is nonlinear because even a 1-dimensional curve could be embedded so as to span all the dimensions of the space.

If a nonlinear surface is smooth, however, then each local piece looks more and more linear under magnification. If we consider only those data points which lie within a local region, then to a good approximation they come from a linear surface patch. The principle values can be used to determine the most likely dimension of the surface and that number of the largest principle components span its tangent space [12]. The key idea behind our representation is to "glue" these local patches together using a partition of unity.

We are exploring several implementations, but all the results reported here come from a representation based on the "nearest point" query. The surface is represented as a mapping from the embedding space to itself which takes each point to the nearest surface point. K-means clustering is used to determine a initial set of "prototype centers" from the data points. A principle components analysis is performed on a specified number of the nearest neighbors of each prototype. These "local PCA" results are used to estimate the dimension of the surface and to find the best linear projection in the neighborhood of prototype i. The influence of these local models is determined by Gaussians centered on the prototype location with a variance

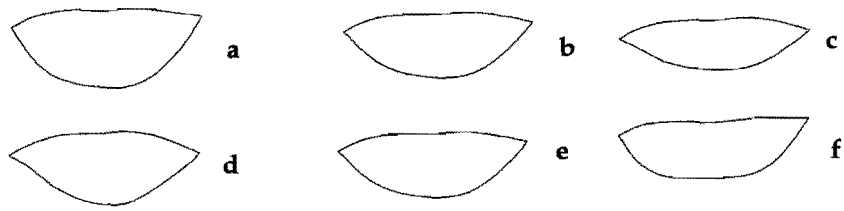


Figure 1: **Examples of lip boundaries.**

determined by the local sample density. The projection onto the surface is determined by forming a partition of unity from these Gaussians and using it to form a convex linear combination of the local linear projections:

$$P(x) = \frac{\sum_i G_i(x) P_i(x)}{\sum_i G_i(x)} \tag{1}$$

This initial model is then refined to minimize the mean squared error between the training samples and the nearest surface point using EM optimization and gradient descent.

We induced surfaces in two different lip feature spaces.

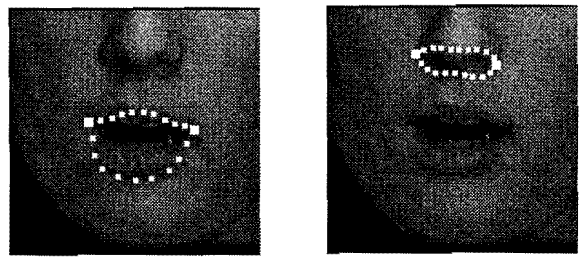
- The most straight forward space is the graylevel space. A 16 x 24 graylevel matrix centered around the lips is treated as an 384 dimensional vector. A learned low dimensional surface embedded in the high dimensional graylevel space will be used for dimension reduction and input coding to our recognition system. (Ultimately we want to use the surface learning paradigm for the full recognition task, but for now we limit its usage just to the task of extracting relevant features from the lips, i.e. projecting new lip images into the surface coordinates).
- The other feature space we investigated is the so called "lip-boundary-shape" space. We used a snake tracking technique [9] to "label" the lip boundaries in the training set (see below). Along the boundaries we evenly distributed 40 points. The x-y coordinates of the points formed a 80 dimensional vector. Figure 1 shows some example boundaries. Based on these 80 dimensional vectors we learned the space of "legal" lip boundaries and use it for a special tracking algorithm (see below).

2.2 Active Models for Tracking

In order to find and scale the graylevel matrix around our lips, we need to have a robust tracking technique. Popular approaches for tracking objects are "snakes" [9] and "deformable templates" [19]. Both of these approaches minimize an "energy function" which is a sum of an internal model energy and an energy measuring the match to external image features.

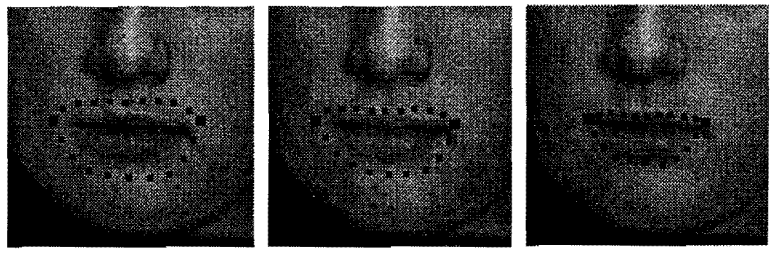
For example, to use the "snake" approach for lip tracking, we form the internal energy from the first and second derivatives of the coordinates along the snake, preferring smoother snakes to less smooth ones. The external energy is formed from an estimate of the negative image gradient along the snake. Figure 2a shows a snake which has correctly relaxed onto a lip contour. This energy function is not very specific to lips, however. The internal energy just causes the snake to be a controlled continuity spline. The "lip- snakes" sometimes relax onto undesirable local minima like that shown in Figure 2b. Models based on deformable templates allow a researcher to more strongly constrain the shape space (typically with hand-coded quadratic linking polynomials), but are difficult to use for representing fine grain lip features.

Our approach is to replace the internal energy described above by a quantity computed from the distance to the learned surface of lip boundary shapes.



a b

Figure 2: Snakes for finding the lip contours a) A correctly placed snake b) A snake which has gotten stuck in a local minimum of the simple energy function.



a b c

Figure 3: a) Initial crude estimate of the contour b) An intermediate step in the relaxation c) The final contour.

Because the training images are initially "labeled" with the conventional snake algorithm, incorrectly aligned snakes were removed from the database by hand. Our experiments show that a 5-dimensional surface in the 80-dimensional boundary space (40 x-y points along the boundary) is sufficient to describe the contours with single pixel accuracy in the image.

The tracking algorithm starts with a crude initial estimate of the lip position and size. It chooses the closest model in the lip surface and maps the corresponding resized contour back onto the estimated image position (Figure 3a). The external image energy is taken to be the cumulative magnitude of graylevel gradient estimates along the current contour. This term has maximum value when the curve is aligned exactly on the lip boundary. We perform gradient ascent in the contour space, but constrain the contour to lie in the learned lip surface. This is achieved by reprojecting the contour onto the lip surface after each gradient step. The surface thereby acts as the analog of the internal energy in the snake and deformable template approaches. Figure 3b shows the result after a few steps and figure 3c shows the final contour. The image gradient is estimated using an image filter whose width is gradually reduced as the search proceeds.

The lip contours in successive images in the video sequence are found by starting with the relaxed contour from the previous image and performing gradient ascent with the altered external image energies.

Empirically, surface-based tracking is far more robust than the "knowledge-free" approaches.

2.3 Nonlinear Interpolation and Sensor Fusion

Based on the tracking algorithm and the dimension reduction with the learned gray-level surface, we can produce 30 visual feature vectors per second (speed of our camera). The acoustic front end (RASTA-PLP) produces 100 feature vectors per second.

As input for the recognition system we want to generate combined visual acoustic feature vectors with 100 frames per second (10 visual dimensions obtained from our graylevel surface +

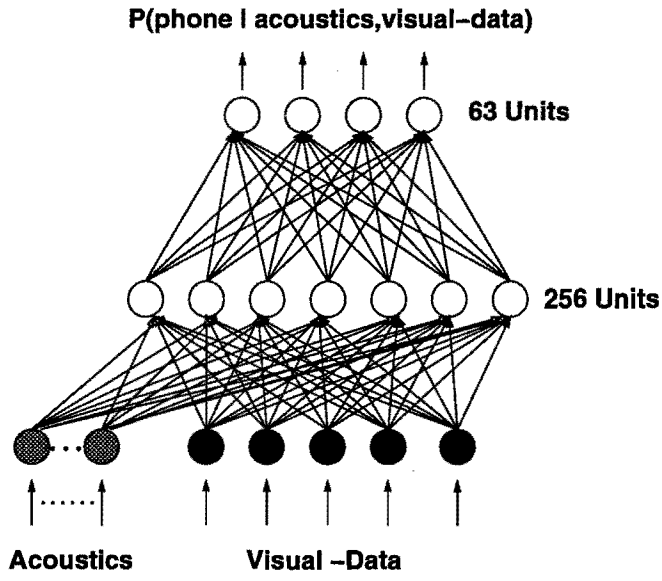


Figure 4: Connectionist architecture.

9 acoustic dimensions obtained from RASTA-PLP adds up to a 19 dimensional bimodal vector). This requires to interpolate the 30 visual frames per second to 100 frames per second. Currently we linearly interpolate these additional vectors. But some lip shapes change drasticly within less than 30 msec which causes "poor" linear interpolated shapes. (e.g. plosives like /b/ and /p/). Given the learned surface of correct lip shapes, we can perform nonlinear interpolation however. If we take two points in the shape space and interpolate along the surface instead going along a straight line (linear interpolation) we never generate incorrect shapes. Studies of this new approach are in progress[1]. We are especially planning to quantify the interpolation performance with real images taken by a high-speed camera (200 frames per second).

3 Bimodal Recognition

Given the bimodal feature vectors we train a multi layer perceptron (MLP) to estimate the following phonetic 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 Hidden Markov Models (HMM) for complete words. (This MLP/HMM system was already successful applied to large continuous acoustic speech recognition [2].)

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 4.

The large window is necessary, because some lip movements start much earlier than the corresponding acoustic output. To confirm this, we looked at cross-modal mutual information measurements. Figure 5 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

[1] In collaboration with Michael Cohen, UC Santa Cruz

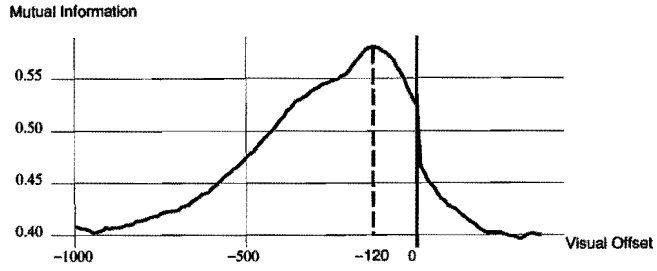


Figure 5: Cross-modal mutual information measurements. The X-axis shows the the visual to acoustic offset and the Y-axis shows the cross-modal mutual information

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)

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.

4 Application

4.1 Spelling-Task

The first experiment is based on a German multi-speaker spelling task database[2]. The training set (2 female, 4 male speakers) consists of 2955 connected letters. For cross-validation we have used an additional 364 letters. An 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.

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 of crosstalk.

Table 1 summarizes all simulation results. On clean speech we did not get a significant improvement. 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.

4.2 Berkeley Restaurant Project

The Berkeley Restaurant Project (BeRP) [17] is a spontaneous speech understanding and dialog system serving as a restaurant guide for people who want to go out for lunch or dinner in

[2]The database was collected in Alex Waibel's research group [3]

the Berkeley area. It was developed at ICSI and used as a testbed for various ideas in speech recognition, natural language research and all kinds of related topics. Currently the user interacts with the system over a head-mounted microphone stating queries like "I would like to eat Korean food not far from campus", and the system responds with suggestions or further questions.

After collecting enough bimodal data with an additional camera, we plan to train a special bimodal MLP/HMM version for the BeRP system in order to perform more robust recognition in the context of office background noise and cross-talk. Other tasks like determing when a speaker actually starts to speak are possible applications. Once the lips start moving, we can activate the whole system to perform recognition.

5 Summary

We demonstrated a new visual learning and processing technique for tracking and recognizing human lips. Based on the combination of this new visual subsystem and an 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.

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

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] C. Bregler, Y. Konig *"Eigenlips" for Robust Speech Recognition,* In Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing, Adelaide.

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

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] S. Omohundro, *Fundamentals of Geometric Learning*, University of Illinois at Urbana-Champaign Technical Report UIUCDCS-R-88-1408.

[13] S. Omohundro, *Bumptrees for Efficient Function, Constraint, and Classification Learning*, In Lippmann, Moody, and Touretzky (ed.), *Advances in Neural Information Processing Systems 3.* San Mateo, CA: Morgan Kaufmann.

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

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

[16] 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.

[17] Charles Clayton Wooters *Lexical Modeling in a Speaker Independent Speech Understanding System* Ph.D. Thesis, U.C. Berkeley, ICSI TR-93-068.

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

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