A HYBRID APPROACH TO BIMODAL SPEECH RECOGNITION

Christoph Bregler, Stephen M. Omohundro, Yochai Konig

Computer Science Division
University of California
Berkeley, CA 94720

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

{bregler,om,konig}@icsi.berkeley.edu

ABSTRACT

We explore multimodal recognition by combining visual lipreading with acoustic speech recognition. We show that combining visual and acoustic speech information improves the recognition performance significantly, especially in noisy environments. This is achieved 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. Our focus in this paper is 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)".

1. INTRODUCTION

Current state-of-the-art acoustic speech recognition systems perform reasonably well only in very controlled lab environments. Once they are applied to real world domains with signal distortions due to background noise, bad acoustic channel characteristics, or crosstalk, the recognition performance decreases drastically. Recognizing a large vocabulary reliably is not yet possible, and most working systems which are used outside of the research lab environment are constrained to a very small vocabulary size, or are speaker dependent.

One way to increase robustness against acoustic signal distortion is to employ noise reduction methods, or noise invariant preprocessing. Another approach is to consider several speech modalities jointly, especially the visual modality, like lipmovements.

One of the earliest successful attempts to improve speech recognition by combining acoustic recognition and lipreading was done by Petajan in 1984 [17]. Further more recent experiments include Mase and Pentland [13], Yuhas et al.[22], Stork et al.[20], Goldschen [7], Silsbee [19], and Bregler et al. [3]. All approaches attempt to show that computer lipreading is able to improve speech recognition, especially in noisy environments. The largest vocabulary set used so far in lipreading experiments was reported by Alan Goldschen, who trained Hidden Markov Models (HMM) on a medium sized speaker dependent subset of the TIMIT database. Other systems have worked on phoneme classification, isolated words, or small continuous word recognition problems. Reported recognition improvements are difficult to interpret and compare, because they are highly dependent on the complexity of the selected task, how advanced the underlying acoustic system is, and how many simplifications were made for the visual task (ranging from reflective lipmarkers or special lipstick to unlabeled full views of the subject moving and tilting her head in front of the camera).

We are interested in developing a system which is capable of performing a state-of-the-art task using speech technologies which produce state-of-the-art results on acoustic recognition in clean and noisy environments, and vision technique that don't require constraints like markers, fixed positioning, or special lighting, scaling, or orientation conditions.

We study environments including the inside of a moving car, and an office environment with overlapping cross talk from a officemate. Our bimodal hybrid speech recognition system consists of a new visual learning and tracking technique, a channel noise robust acoustic front end (RASTA-PLP), a connectionist phone probability estimator (MLP) and a HMM speech recognizer. Both the acoustic front end and the MLP/HMM speech recognizer have been demonstrated to produce state-of-the art results on standard acoustic databases (like ARPA resource management) and are capable of scaling up to the current largest tasks in the speech community (the ARPA Wall Street Journal evaluation on 5,000 word continuous speaker independent recognition). 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. Details about the other parts of our hybrid system are reported in [8, 2].

For our experiments we collected two different visual-acoustic databases, a continuous word recognition spelling task and an open vocabulary spontaneous dialog task, the Berkeley Restaurant Project (BeRP)[9]. We show significant improvement using lipreading over the acoustic baseline system on the first database. Experiments with the second database are still in progress.

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 speaker's face.

1058-6393/95 $4.00

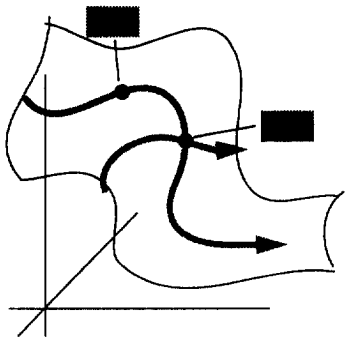

Figure 1: Lipconfiguration space as smooth nonlinear manifold

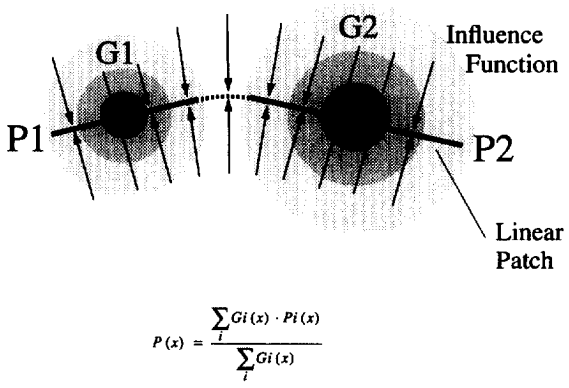

Figure 2: Local linear patches glued together to a nonlinear surface.

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 muscular apparatus, the dimensionality of our "lip-configuration-space" should not be higher than the number of free parameters in the vocal tract. (The ultimate goal is to induce a mapping from our raw images to a space with a dimensionality exactly equal to that of the 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 (see Figure 1).

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 ranges 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 principal components analysis could be used to discover the dimension of the linear subspace and to find the best-fit linear space of that dimension. The largest principal vectors would span the space and there would be a precipitous drop in the principal values at the dimension of the surface. A principal 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 principal values can be used to determine the most likely dimension of the surface and that number of the largest principal components span its tangent space [15]. 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 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. Figure 2 illustrates this "gluing" process.

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 is used for dimension reduction and input coding for our recognition system. (Similar coding using linear subspaces instead of nonlinear subspaces were reported by [18, 11]).

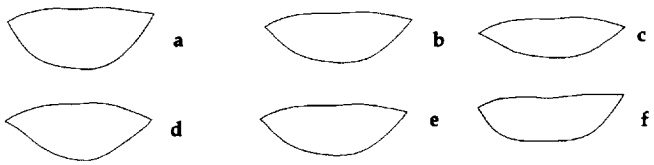

Figure 3: Examples of lip boundaries.

- The other feature space we investigated is the so called "lip-boundary-shape" space. We use a snake tracking technique [10] 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 an 80 dimensional vector. Figure 3 shows some example boundaries. Based on these 80 dimensional vectors we learned the space of "legal" lip boundaries and use it for the tracking algorithm described below.

2.2. ACTIVE MODELS FOR TRACKING

In order to find and scale the graylevel matrix around the lips, we need to have a robust tracking technique. Popular approaches for tracking objects are "snakes" [10] and "deformable templates" [23]. 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.

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. A local minima represents the correctly aligned snake to the object 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 eyes, noses, or full faces. 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.

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 4a). 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

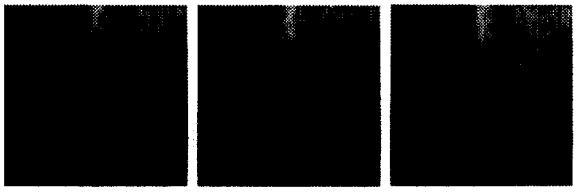

Figure 4: a) Initial crude estimate of the contour b) An intermediate step in the relaxation c) The final contour.

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 4b shows the result after a few steps and figure 4c 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 + 9 acoustic dimensions obtained from RASTA-PLP adds up to a 19-dimensional bimodal vector). This requires us to interpolate the 30 visual frames per second to 100 frames per second.

Currently we use two different interpolation approaches: Linear interpolation and surface-based nonlinear interpolation. Some lip shapes change drastically within less than 30 msec which causes "poor" linear interpolated shapes. (e.g. plosives like /b/ and /p/, see Figure 5a). 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 don't generate incorrect shapes (Figure 5b). A more detailed performance analysis of this interpolation technique and its application to other vision domains is discussed in [4].

We are also planning to quantify the interpolation performance with images taken by a high-speed camera (100 frames per second)[1].

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

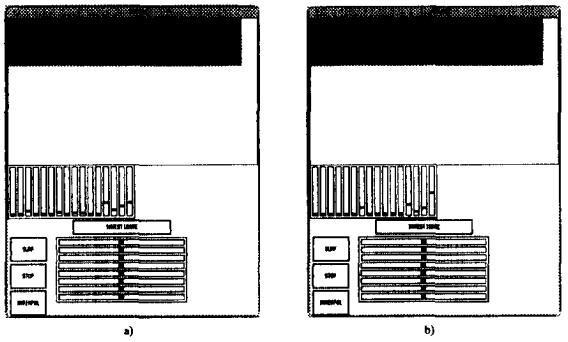

Figure 5: Lip images: a) linear interpolation b) nonlinear interpolation.

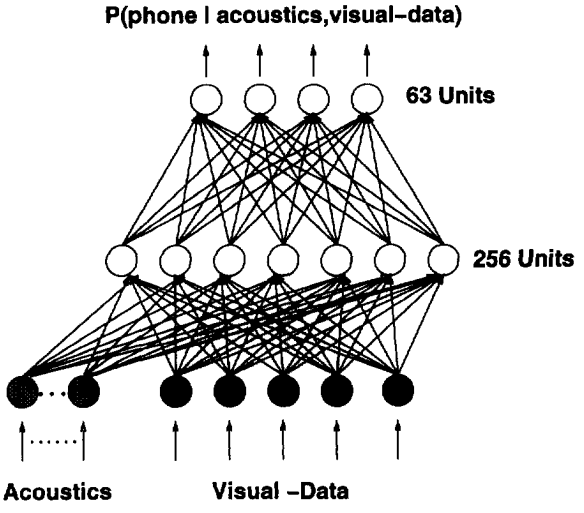

Figure 6: Connectionist architecture.

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 has already been 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 6.

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 7 shows the mutual information between the acoustic feature vectors and the visual feature vectors with various temporal offsets. The X-

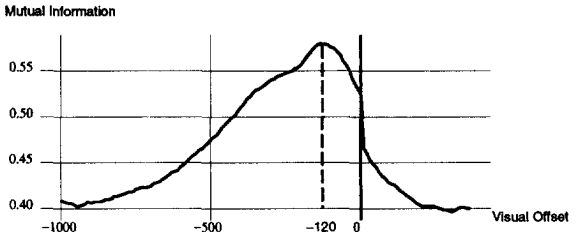

Figure 7: Cross-modal mutual information measurements. The X-axis shows the the visual to acoustic offset and the Y-axis shows the cross-modal mutual information

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 maximally 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.

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 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.

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

Task	Acoustic	Eigenlips	Delta-Lips	Err.Red.
clean	11.0 %	10.1 %	11.3 %	-
20db SNR	33.5 %	28.9 %	26.0 %	22.4 %
10db SNR	56.1 %	51.7 %	48.0 %	14.4 %
15db SNR crosstalk	67.3 %	51.7 %	46.0 %	31.6 %

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

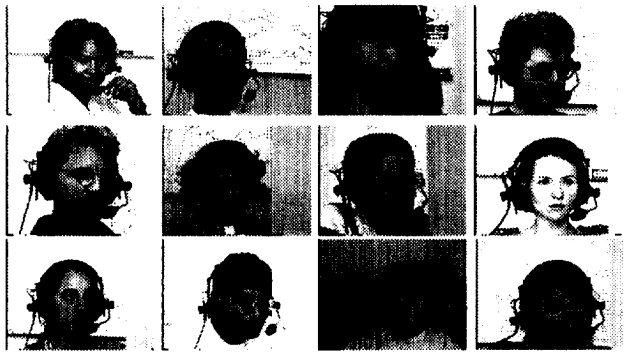

Figure 8: BeRP bimodal datacollection

4.2. BERKELEY RESTAURANT PROJECT

The Berkeley Restaurant Project (BeRP) [21, 9] is a spontaneous speech understanding and dialog system serving as a restaurant guide for people who want to go out to eat in the Berkeley area. It was developed at ICSI and used as a testbed for ideas in speech recognition, natural language research and related topics. Currently the user interacts with the system using 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.

Our bimodal database consists of subjects with various ethnic and national backgrounds, representing a realistic mix of the current population in the United States. No special attempts were made to reduce office background noise, or exact head/lip positioning, in order to provide a realistic human computer interaction scenario. Figure 8 shows some example frames.

Experiments are in progress to train a larger system using the techniques reported here and techniques reported in [9]. Currently we are also investigating the utility of full face-finding approaches [12].

Acknowledgments We would like to thank Jerry Feldman, Nelson Morgan, Alex Waibel, and the ICSI Speech Group for their support and helpful discussions. This research was funded by the Advanced Research Project Agency, under contract #N0000 1493 C0249 and ICSI. Parts of the database was collected with funds from Land Baden Wuerttemberg (Landesschwerpunkt Neuroinformatik) in Alex Waibel's research group.

5. 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 Advances in Neural Information Processing Systems 6. Morgan Kaufmann Publishers, 1994.

[4] C. Bregler and S. Omohundro, *Nonlinear Image Interpolation using Surface Learning* to appear in Advances in Neural Information Processing Systems 7. Morgan Kaufmann Publishers, 1995.

[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] D. Jurafksy, Chuck Wooters, Gary Tajchman, J. Segal, A. Stolcke, E. Fosler, N. Morgan, *The Berkeley Restaurant Project* Proc. Int. Conf. of Spoken Language Processing (ICSLP), Tokyo, Japan, 1994.

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

[11] M. Kirby, F. Weisser, and G. Dangelmayr, *A Model Problem in Represetation of Digital Image Sequences*, in Pattern Recgonition, Vol 26, No. 1, 1993.

[12] T. Leung, *Face Recognition using Stochastic Search* Personal communication.

[13] 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.

[14] 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.

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

[16] 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.

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

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

[19] P. L. Silsbee *Computer Lipreading for Improved Accuracy in Automatic Speech Recognition* Ph.D. Dissertation, University of Texas at Austin, May 1993.

[20] G.J. Wolff, K.V. Prasad, D.G. Stork, and M.Hennecke *Lipreading by Neural Networks: Visual Preprocessing, Learning and Sensory Integration.* in Advances in Neural Information Processing Systems 6. Morgan Kaufmann Publishers, 1994.

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

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

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