SEE ME, HEAR ME: INTEGRATING AUTOMATIC SPEECH RECOGNITION AND LIP-READING

Paul Duchnowski[1] *Uwe Meier*[1] *Alex Waibel*[1,2]

[1]University of Karlsruhe, Karlsruhe, Germany
[2]Carnegie Mellon University, Pittsburgh PA, USA

ABSTRACT

We present recent work on integration of visual information (automatic lip-reading) with acoustic speech for better overall speech recognition. A Multi-State Time Delay Neural Network performs the recognition of spelled letter sequences taking advantage of lip images from a standard camera. The problems addressed include efficient but effective representation of the visual information and optimum manner of combining the two modalities when rendering a decision. We show results for several alternatives to direct gray level image as the visual evidence. These are: Principal Components, Linear Discriminants, and DFT coefficients. Dimensionality of the input is decreased by a factor of 12 while maintaining recognition rates. Combination of the visual and acoustic information is performed at three different levels of abstraction. Results suggest that integration of higher order input features works best. On a continuous spelling task, visual-alone recognition of 45-55%, when combined with acoustic data, lowers audio-alone error rates by 30-40%.

1. INTRODUCTION

Natural human-computer interaction cannot be achieved as long as keyboards remain the primary input modality. It is the aim of multiple concerted research projects in our labs at University of Karlsruhe and Carnegie Mellon University to develop interfaces that will take advantage of all communication modalities normally and effortlessly used by people. Integration of all such information would make not only interaction with computers more satisfying for the user but would also enable the machine to better comprehend the user's intent and instructions. We are pursuing machine understanding of speech, lip motion, gesture, eye gaze, hand writing, face recognition and tracking and sound localization. Overviews of some of these projects can be found in [14].

In this paper we concentrate on the problem of integrating acoustic and visual information for better speech recognition. It is well known that hearing-impaired listeners and those listening in adverse acoustic environments (noise, reverberation, multiple speakers) rely heavily on the visual input to disambiguate among acoustically confusable speech elements. The usefulness of lip movement information stems in large part from its rough complementariness to the acoustic signal: the former is most reliable for distinguishing the place of articulation, the latter conveys most robustly manner and voicing information (e.g. [13]).

Automatic speech recognition (ASR) systems' performance is, if anything, even more sensitive to degradation of the acoustic input. Therefore, it is only natural to try to supplement the acoustic data with lip movement information. Related work on this concept has been reported by other researchers in [4, 9, 10, 11, 12, 15]. These studies clearly indicated that combined audio-visual speech recognition was feasible. However, most of the experiments have relied on such simplifications as head-mounted cameras, reflective markers placed on the speaker's lips, or manual extraction of the relevant part of the face image. Our goal is to translate the concept to a practical system dealing with more complex tasks and employing robust and *non-invasive* capture and pre-processing of the visual information. Here we report primarily on two aspects of this effort: designing an efficient but rich representation of the visual input and developing a method to optimally combine the acoustic and visual evidence when making the final identification decision.

The audio-visual ASR system under development in our laboratory was first described in [2]. It is designed to recognize continuously spelled names and nonsense letter sequences of arbitrary length using the German alphabet. The task is thus equivalent to contiuous recognition with a small but highly confusable vocabulary.

2. SYSTEM DESCRIPTION AND DEVELOPMENT

2.1. Fundamental Design

In the basic set-up, we record, in parallel, the acoustic speech and the corresponding series of mouth images of the speaker. The acoustic signal is sampled at 16kHz with 14-bit resolution. A fairly standard front-end then computes 16 Melscale Fourier coefficients on Hamming-windowed speech segments at a 10 msec frame rate. This component remains invariant for all experiments described below.

The visual evidence is obtained by "frame-grabbing" the output of a conventional camcorder camera at 30 frames/sec, with 8-bit gray level resolution. In our first work [2], pictures of the lip region were manually extracted from the image. Currently, speakers are asked to position themselves such that their lips appear within a rectangle that is simultaneously shown on the screen of a workstation. However, no special markers, restraints or position indicators are used. The 144x80 pixel images in that rectangle constitute the video input available for further processing. In the most straightforward approach the original images are low-pass filtered and downsampled to a size of 24x16 pixels. Rudimentary histogram fitting normalizes the pixel values to lie in the interval [-1,1]. The resulting sequence of 384-dimensional "normalized pixel vectors" (one vector for each lip image frame) is then used as the input to the recognition algorithm.

In the basic system a modular Multi-State Time Delay Neural Network (MS-TDNN) [5] is used to perform the recognition. Figure 1 schematically shows the architecture. Through the first three layers (input-hidden-phoneme/viseme) the acoustic and visual inputs are processed separately. The third layer produces activations for 62 phoneme or 42 viseme[1] states for acoustic and visual

[1]A viseme, the rough visual correlate of a phoneme, is the smallest visually distinguishable unit of speech.

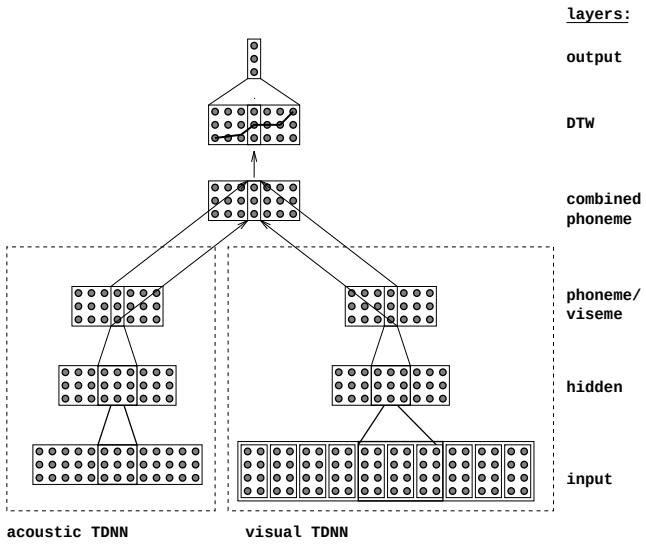

Figure 1. Original recognition network architecture (Net-P).

data, respectively. Weighted sums of the phone and corresponding viseme activations are entered in the combined layer and a one stage DTW algorithm finds the optimal path through the phone states that decodes the recognized letter sequence. The weights in the parallel networks are trained by backpropagation. There are 15 hidden units in both subnets. The combination weights are computed dynamically during recognition to reflect the estimated reliability of each modality. These "entropy weights" [2], λ_A for the acoustic side and λ_V for the visual are given by:

$$\begin{aligned} \lambda_A &= b + \frac{S_V - S_A}{\Delta S_{max-over-data}} \\ \lambda_V &= 1 - \lambda_A \end{aligned} \tag{1}$$

The entropy quantities S_A and S_V are computed for the acoustic and visual phone/viseme activations by normalizing these to sum to one 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 particular modality. The bias b pre-skews the weights to favor one of the modalities.

2.2. Visual Data Representation

Unlike for acoustic speech data, there are no generally agreed-upon parameterization strategies for the visual lip images. 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 have been investigating several alternate visual data representations consistent with this strategy.

The dimensionality of the normalized pixel vector is quite high, especially when compared with the acoustic input vector. 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. Equalizing the visual and acoustic input vector dimensions also seems advantageous for low-level integration of the information (see Sec. 2.3). Storage savings would also result.

2.2.1. Principal Components

A well-known method that can accomplish this is Principal Component Analysis [7] (also known as Karhunen-Loeve expansion). In this approach the original vectors are projected onto the eigenvectors of their covariance matrix and only the coefficients corresponding to the largest N eigenvalues are retained ($N < 384$). This preserves most of the variance in the original data. By treating the images simply as data vectors, we could arbitrarily reduce the dimensionality of the data. Visual examination suggested that images essentially indistinguishable from the originals could be reconstructed from as few as 16 principal components (PCs). Pilot experiments suggested, however, that recognition performance reached a plateau for input of between 30 and 60 PCs. In all subsequent experiments with this method, 32 PCs were used as the visual input.

2.2.2. Linear Discriminant Analysis

PCA is most suitable to *description* of data with a reduced set of parameters. On the other hand, the goal of recognition is rather the *discrimination* among several input classes (e.g. phones). A related method of data rate reduction that is better geared towards this goal is Linear Discriminant Analysis (LDA) [3]. Here the original data is also projected onto a set of vectors and only the most significant of the resulting coefficients are used further. However, the projection vectors, calculated as the eigenvectors of so-called scatter matrices, maximize the separability of different input classes in the reduced-dimensional representation. In order to determine the projection vectors one has to assign each training data vector to a class. In our case we labelled the data vectors by one of 62 phones. Again, preliminary experiments led us to further use of 32 LDA coefficients.

2.2.3. Fourier Transfrom

It is known that almost all typical images are uniquely specified by the magnitude of their Fourier Transform [8]. This parameterization is also potentially resistant against translation of the input image and offers several methods of reducing the data count by grouping the DFT coefficients. We obtained most extensive and promising results for the "ring" grouping where each parameter $\hat{m}_i$ is calculated from the DFT magnitude $M(k_1, k_2)$ by:

$$\hat{m}_i = \sum_{k_1, k_2 \in \mathcal{R}_i} M(k_1, k_2) \tag{2}$$

where $\mathcal{R}_i$ contains k_1, k_2 such that $\rho_{i-1} \leq \sqrt{k_1^2 + k_2^2} \leq \rho_i$ with the ring radii ρ_i increasing logarithmically. This parameterization is thus roughly equivalent to computing the energies of the outputs of a bank of bandpass filters. We computed a total of 29 DFT-ring parameters. It should be noted that the DFT was calculated from the original 144x80 picture frame and not from the downsampled one.

All parameters are normalized by a histogram computation similar to that of the gray levels to lie between -1 and 1, before being used as input to the network.

2.3. Combination Alternatives

The combination of acoustic and visual information at 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 phones. Unlike phones, visemes have proven much more difficult to define consistently except for a few fairly constant sets. Also, little research has been done on non-English visemes. Combination of phones and visemes

Visual Input	Parameter Count	Word Accuracy (%)	
		data set	
		mum1-2	mum9-10
Gray Levels	384	55	44
PCs	32	52	45
LDA	32	53	52
DFT Ring	29	50	38

Table 1. Visual-only recognition rates for different data representations.

further prevents the recognizer from taking advantage of lower level correlations between acoustic and visual events such as inter-modal timing relationships. There is evidence that humans integrate the bi-modal inputs to take advantage of such cues [1, 13].

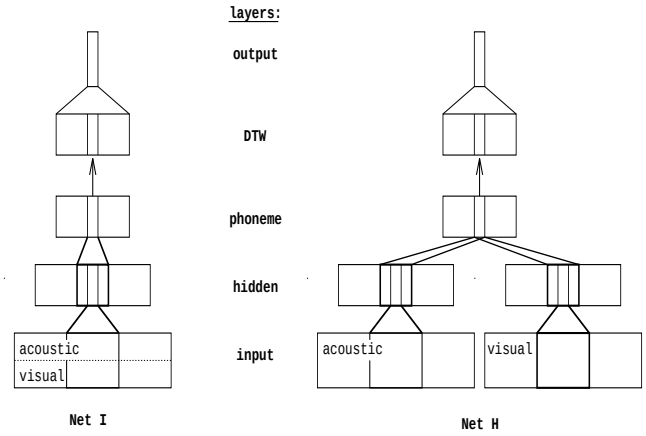

Figure 2. Alternate structures for acoustic and visual combination.

Two combination alternatives are illustrated in Figure 2. The acoustic and visual data vectors can simply be concatenated to form one input vector to a single MS-TDNN (both types of inputs are normalized to the same range) resulting in effective combination at the input layer. Alternately, we can elect to integrate the higher level features by combination at the hidden layer. In both cases all weights are trained by backpropagation. We will refer to the nets combining the modalities at the phone, hidden, and input level as Net-P, Net-H, and Net-I, respectively.

3. PERFORMANCE

We have tested the various recognizer versions on audio-visual speech data from one (male) speaker. In the standard paradigm 200 letter sequences (average length of about 6 letters) were recorded in two sessions on the same day. This data set was then stored digitally for further processing. 170 of the sequences were used as training data to set the weights of the network, 15 constituted the cross-validation set and 15 the test set. The covariance and scatter matrices required by PCA and LDA were computed using only the training set.

3.1. Experimental Results

Under optimal lighting and recording conditions we have observed visual-only word accuracy[2] as high as 72% which, when combined with the acoustic side, leads to error rate reductions of 60%. It is still difficult consistently to reproduce the recording environment necessary for this level of performance. We therefore report detailed results on more representative data.

[2] In our case "word" refers to a single pronounced letter.

Table 1 gives word accuracies for two different data sets when only the visual part of Net-P is used. The scores are generally quite comparable. It bears stressing that the non-gray level parameterizations achieve this recognition while reducing the data rate by a factor of 12.

Figure 3 shows the results of combining acoustic and visual information with Net-P for the tested parameterizations. Data set mum1-2 was used. 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. Error rate reduction varied from 30% for clean data to 37% for noisy speech.[3] On data set mum9-10, for which not all combination conditions have yet been tested, noisy error reduction of an average 40% has been observed.

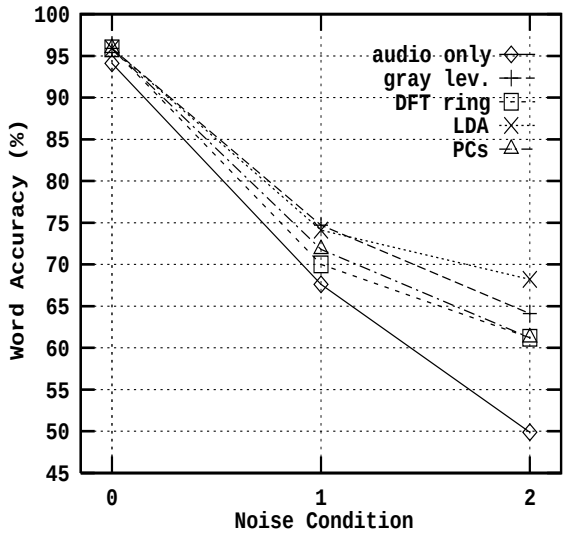

Figure 3. Net-P combination results in quiet (Noise = 0) and two noise conditions.

We compare the performance for the three alternative network structures in Figure 4. Only the results for gray level and LDA input are shown. Scores for PC were also obtained but were significantly and uniformly lower. For Net-I the gray level and LDA scores are essentially the same and are comparable to those of Net-P although perceptibly worse for clean speech (in fact, the audio-visual score is worse than the audio-alone for this condition). Net-H shows the gray level input significantly outperforming LDA and, in fact, turning in marginally best scores of all tested inputs and net architectures, except at the highest noise level.

3.2. Discussion

The results, taken together, indicate that of the tested visual input representations, the gray levels and LDA gave very similar performance under most conditions. Conversely, PC and DFT-Ring usually gave lower scores. Thus, with proper choice of transformation we can significantly (factor of 12) reduce the dimensionality of the input without sacrificing performance. Note that the reduction is done without any heuristic feature extraction. A better test of whether the smaller parameter count benefits generalization might be speaker-independent recognition which we intend to test.

[3] The bias parameter b in Eq. 2 was set to the empirical optimum for each parameterization and noise condition. In practice it would be guided by an automatic noise level estimation algorithm which we intend to include in the system. Setting b to its optimum for clean speech still leads to uniformly improved combined scores under noise.

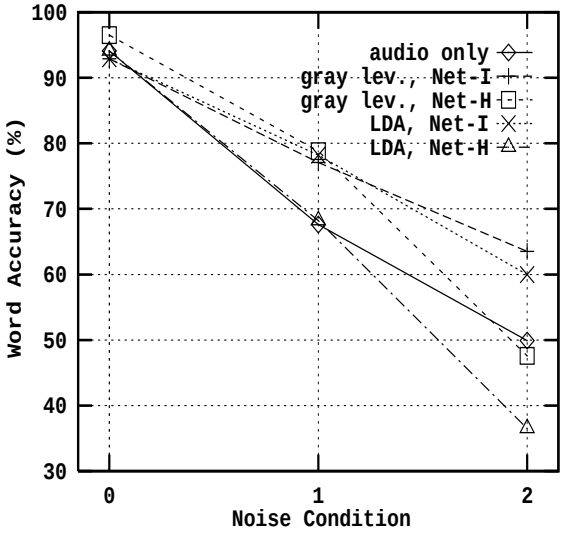

Figure 4. Combination results for Net-I and Net-H.

Comparison of different net structures yields more equivocal conclusions. All three are clearly capable of improving recognition with the addition of visual information. However, Net-P combination of the modalities *always* yields a better score than either modality alone which is not true of the other two structures. On the other hand, neither Net-I nor Net-H have been optimized at this time (for instance, the number of hidden units, 15 was inherited from Net-P). It is surprising the Net-I performed almost equivalently for both gray levels and LDA input. It was expected that making the number of visual parameters closer to the number of acoustic ones would improve the learning of input weights. It is possible that the LDA transformation obscures those cues most useful for low-level cross-modal correlation.

The results obtained with gray level input and Net-H are especially interesting. This structure is effectively combining higher level acoustic and visual features which it itself determines (i.e., they are not prescribed externally). It may thus be theoretically capable of learning such cross-modal relationships as the time interval between lip opening and onset of voicing or nasality. This suggests a closer examination of activation patterns in the hidden layer. The poor performance of Net-H under Noise Condition 2 stems probably from lack of noisy acoustic data in the training set. The network thus has no way of learning, at the integration level, what constitutes a noisy and thus ignorable feature. Net-P, on the other hand, dynamically adjusts the weights given to each modality depending on its reliability. A hybrid approach of pre-learned combination weights that can be modified during recognition might prove fruitful.

4. WORK IN PROGRESS

The goal of our project is the creation of a seamless multimodal communication interface. This necessitates liberating the user from as many limitations on movement as possible. The present system, however, is still fairly restrictive (though non-invasive). It is also not sufficiently robust against relatively small changes in the visual input.

We have identified three main sources of confusion: changes in lighting and position and size of the lips within the frame. By using adaptive histogram normalization of the image we have successfully made the system independent of reasonable variations in illumination including illumination gradients. The problem of size normalization is probably the most severe, with size changes of 10-20% causing severe performance degradations. We have found that to some extent these effects can be mitigated by training the networks on larger data corpuses, recorded under different conditions or artificially enlarged, shifted, etc. The training effort, however, becomes impractical.

A more elegant collective solution to these challenges involves automatic control of the camera to track the face of the speaker in a room and an algorithm to accurately locate the lips within the face image [6]. Prototypes of both the face-tracker and lip-locator have already been developed in our laboratory and we're in the process of integrating them into a single system. Together they allow for reasonable movement and enable size- and location-normalization of the input lip image.

ACKNOWLEDGEMENTS

This work is sponsored by the state of Baden-Württemberg, Germany (Landesschwerpunkt Neuroinformatik). Partial support was also provided by the Advanced Research Projects Agency (US).

REFERENCES

[1] L.D. Braida. Crossmodal Integration in the Identification of Consonant Segments. *The Quart. J. of Exp. Psych.*, 43A(3), 1991, pp.647-677.

[2] C. Bregler, H. Hild, S. Manke, and A. Waibel. Improving Connected Letter Recognition by Lipreading. in *Proc. ICASSP '93.*

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* San Diego: Academic Press, 1990.

[4] A.J. Goldschen. Continuous Automatic Speech Recognition by Lipreading. Ph.D. Dissertation. George Washington University, 1993.

[5] H. Hild and A. Waibel. Connected Letter Recognition with a Multi-State Time Delay Neural Network. *Neural Information Processing Systems* (NIPS-5), 1993.

[6] H.M. Hunke. Lokalisieren von Gesichtern mit Hilfe von neuronalen Netzen. M.S. Thesis, University of Karlsruhe, 1994.

[7] I.T. Joliffe. *Principal Component Analysis.* New York: Springer-Verlag, 1986.

[8] J.S. Lim. *Two-Dimensional Signal and Image Processing.* Englewood Cliffs, N.J.: Prentice-Hall, 1990.

[9] K. Mase and A. Pentland. Automatic Lipreading by Optical-Flow Analysis. *Systems and Computers in Japan*, 22(6), 1991, pp. 67–76.

[10] E.D. Petajan. Automatic lipreading to enhance speech recognition. in *Proc. IEEE Communications Society Global Telecom. Conf.*, Atlanta GA, Nov. 1984.

[11] P. Silsbee and A. Bovik. Audio-visual speech recognition for a vowel discrimination task. *Proceedings of SPIE*, vol. 2094, 1993, pp.84–95.

[12] D.G. Stork, G. Wolff, and E. Levine. Neural network lipreading system for improved speech recognition. in *Proc. IJCNN'92.*

[13] Q. Summerfield. Audio-visual Speech Perception, Lipreading and Artificial Stimulation. in *Hearing Science and Hearing Disorders*, M.E. Lutman and M.P. Haggard eds., New York: Academic Press, 1983.

[14] A. Waibel, M.T. Vo, P. Duchnowski, and S. Manke. Multimodal Interfaces. to appear in *Artificial Intelligence Review Journal*, special issue, 1994.

[15] B.P. Yuhas, M.H. Goldstein, Jr., and T.J. Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, Nov. 1989.