Proceedings of the IEEE International Conference on Multimedia and Expo, New York, vol. II, pp. 1097-1100, 2000

# A CASCADE IMAGE TRANSFORM FOR SPEAKER INDEPENDENT AUTOMATIC SPEECHREADING

*G. Potamianos, A. Verma*[*]*, C. Neti, G. Iyengar,* and *S. Basu*

IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A.
Email: {gpotam,cneti,giyengar,sbasu}@us.ibm.com

## ABSTRACT

We propose a three-stage pixel based visual front end for automatic speechreading (lipreading) that results in improved recognition performance of spoken words or phonemes. The proposed algorithm is a cascade of three transforms applied to a three-dimensional video region of interest that contains the speaker's mouth area. The first stage is a typical image compression transform that achieves a high "energy", reduced-dimensionality representation of the video data. The second stage is a linear discriminant analysis based data projection, which is applied to a concatenation of a small number of consecutive image transformed video data. The third stage is a data rotation by means of a maximum likelihood linear transform. Such transform optimizes the likelihood of the observed data under the assumption of their class conditional Gaussian distribution with diagonal covariance. We apply the algorithm to visual-only 52-class phonetic and 27-class visemic classification on a 162-subject, 7-hour long, large vocabulary, continuous speech audio-visual dataset. We demonstrate significant classification accuracy gains by each added stage of the proposed algorithm, which, when combined, can reach up to 27% improvement. Overall, we achieve a 49% (38%) visual-only frame level phonetic classification accuracy with (without) use of test set phone boundaries. In addition, we report improved audio-visual phonetic classification over the use of a single-stage image transform visual front end.

## 1. INTRODUCTION

*Automatic speech recognition* (ASR) by using the video sequence of the speaker's lips, namely *automatic lipreading*, or *speechreading*, has attracted significant interest as a means of improving traditional audio-only ASR [1]-[12]. Such an improvement depends on both the audio and visual information fusion technique, as well as on the *visual front end*, namely, on extracting appropriate visual features which contain relevant speech information. Various such features have been proposed in the literature that, in general, can be grouped into *lip contour* based and *pixel* based ones [1]. In the first approach, the speaker's lip contours are extracted from the image sequence, a parametric or statistical lip contour model is obtained, and the model parameters are used as visual features. In the second approach, the entire image containing the speaker's mouth is considered as informative for lipreading (*region of interest* - ROI), and appropriate transformations of its pixel values are used.

In this paper, we concentrate on the visual front end for automatic speechreading, and we investigate the pixel based approach to it. Specifically, we propose a three-stage algorithm that consists of a cascade of three transforms applied to the ROI data vector. The first stage is a typical image transform, such as the *discrete cosine* (DCT) [2]-[4], the *discrete wavelet* (DWT) [4], [12], [13], and the *Karhunen-Loève* transform (KLT) (or, *principal component analysis*-PCA) [2]-[9], that seeks data dimensionality reduction, through data compression. The second stage is a *linear discriminant analysis* (LDA) data projection [2], [11], that seeks optimal classification performance and further data dimensionality reduction. The third stage is a *maximum likelihood linear transformation* (MLLT) aimed at optimizing the observed data likelihood under the assumption of their class conditional Gaussian distribution with diagonal covariance [14]. The proposed algorithm is novel in two aspects: First, MLLT has never before been used for speechreading, and, second, both DCT and DWT have up to date been considered as a one-step visual front end [2]-[4], [12].

The paper is structured as follows: The three algorithm stages are discussed in Sections 2, 3, and 4, respectively. Specifics of all components of our speechreading system are discussed in Section 5, and experimental results are presented in Section 6.

## 2. IMAGE TRANSFORMS FOR DATA COMPRESSION

Let us consider, for every video frame $t$, a three dimensional region of interest (ROI), centered around the speaker's mouth center, $(m_t, n_t)$, obtained as described in Section 5.1 (see also Fig. 1). The (monochrome) ROI pixel values are placed into vector

$$\underline{x}_t^{(\mathrm{I})} \leftarrow \{\, V(m,n,k):\ m_t - \lfloor M/2 \rfloor \le m < m_t + \lceil M/2 \rceil\,, \quad n_t - \lfloor N/2 \rfloor \le n < n_t + \lceil N/2 \rceil,\ t - \lfloor K/2 \rfloor \le k < t + \lceil K/2 \rceil \,\}\,, \tag{1}$$

of length $d^{(\mathrm{I})} = MNK$. We seek a $D^{(\mathrm{I})} \times d^{(\mathrm{I})}$-dimensional *linear transform* matrix $\mathbf{P}^{(\mathrm{I})} = [\,\underline{P}_1, ..., \underline{P}_{D^{(\mathrm{I})}}\,]^\top$, such that the transformed data vector $\underline{y}_t^{(\mathrm{I})} = \mathbf{P}^{(\mathrm{I})} \underline{x}_t^{(\mathrm{I})}$ contains most speechreading information in its $D^{(\mathrm{I})} \ll d^{(\mathrm{I})}$ elements. To obtain matrix $\mathbf{P}^{(\mathrm{I})}$, $L$ training examples are given, denoted by $\underline{x}_l^{(\mathrm{I})}$, $l = 1,...,L$.

### 2.1. Discrete Wavelet and Cosine Transforms

A number of linear, *separable* image transforms can be used in place of $\mathbf{P}^{(\mathrm{I})}$. In this work, we consider both the discrete cosine transform (DCT) [2]-[4], and the discrete wavelet transform (DWT) implemented by means of the Daubechies class wavelet filter of approximating order 2 [4], [13], [15]. Matrix $\mathbf{P}^{(\mathrm{I})}$ has as rows the image transform matrix $\mathcal{T}$ rows that maximize the *energy*

$$\sum_{d=1}^{D^{(\mathrm{I})}} \sum_{l=1}^{L} < \underline{x}_l^{(\mathrm{I})}, \underline{\mathcal{T}}_{j_d}^\top >^2, \quad \text{where } j_d \in \{1,...,d^{(\mathrm{I})}\}\,,$$

are disjoint, $<\bullet\,,\bullet>$ denotes vector *inner product*, and $\bullet^\top$ denotes vector or matrix *transpose*.

[*] IBM India Research Lab, New Delhi, India; vashish@in.ibm.com

### 2.2. Principal Component Analysis

Principal component analysis (PCA) [2]-[9] achieves optimal data compression, in the sense of minimum mean square error between $\underline{x}_t^{(\mathrm{I})}$ and its reconstruction based on $\underline{y}_t^{(\mathrm{I})}$. In our PCA implementation, we scale the data according to their inverse variance. Namely, we compute the data *mean* and *variance* as

$$\mu_d = \frac{1}{L}\sum_{l=1}^{L} x_{l,d}^{(\mathrm{I})}\,, \text{ and } \sigma_d^2 = \frac{1}{L}\sum_{l=1}^{L}(x_{l,d}^{(\mathrm{I})}-\mu_d)^2\,, \quad d=1,...,d^{(\mathrm{I})},$$

respectively, and the *correlation* $d^{(\mathrm{I})}\times d^{(\mathrm{I})}$ matrix $\mathbf{R}$ with elements

$$r_{d,d'} = \frac{1}{L}\sum_{l=1}^{L}\frac{(x_{l,d}^{(\mathrm{I})}-\mu_d)}{\sigma_d}\,\frac{(x_{l,d'}^{(\mathrm{I})}-\mu_{d'})}{\sigma_{d'}}\,, \text{ for } d,d'=1,...,d^{(\mathrm{I})}.$$

We then *diagonalize* the correlation matrix as $\mathbf{R}=\mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top$ [15], [16], where $\mathbf{A}=[\,\underline{A}_1,...,\underline{A}_{d^{(\mathrm{I})}}\,]$ has as columns the *eigenvectors* of $\mathbf{R}$, and $\mathbf{\Lambda}$ is a diagonal matrix containing the *eigenvalues* of $\mathbf{R}$. Let the $D^{(\mathrm{I})}$ largest such eigenvalues be located at the $j_1,...,j_{D^{(\mathrm{I})}}$ diagonal positions. Given data vector $\underline{x}_t^{(\mathrm{I})}$, we normalize it elementwise as $x_{t,d}^{(\mathrm{I})} \leftarrow (x_{t,d}^{(\mathrm{I})}-\mu_d)/\sigma_d$, and subsequently we extract its feature vector $\underline{y}_t^{(\mathrm{I})} = \mathbf{P}^{(\mathrm{I})}\underline{x}_t^{(\mathrm{I})}$, where $\mathbf{P}^{(\mathrm{I})}=[\,\underline{A}_{j_1},...,\underline{A}_{j_{D^{(\mathrm{I})}}}\,]^\top$.

## 3. LINEAR DISCRIMINANT DATA PROJECTION

In the proposed cascade algorithm, and in order to capture important *dynamic* visual speech information, linear discriminant analysis (LDA) is applied to the concatenation of $J$ consecutive image transformed feature vectors

$$\underline{x}_t^{(\mathrm{II})} = [\,\underline{y}_{t-\lfloor J/2\rfloor}^{(\mathrm{I})\top},...,\underline{y}_t^{(\mathrm{I})\top},...,\underline{y}_{t+\lceil J/2\rceil-1}^{(\mathrm{I})\top}\,]^\top,$$

of length $d^{(\mathrm{II})}=D^{(\mathrm{I})}J$.

In general, LDA [2], [11], [14] assumes that a set of *classes* $\mathcal{C}$ is a-priori given, as well as that the training set data vectors $\underline{x}_l^{(\mathrm{II})}$, $l=1,...,L$, are *labeled* as $c(l)\in\mathcal{C}$. LDA seeks a projection $\mathbf{P}^{(\mathrm{II})}$, such that the projected training sample $\{\,\mathbf{P}^{(\mathrm{II})}\underline{x}_l^{(\mathrm{II})},\ l=1,...,L\,\}$ is "well separated" into $\mathcal{C}$. Let $\mathbf{S_W}$, $\mathbf{S_B}$ be the *within-class scatter* and *between-class scatter* matrices of the training sample, given by

$$\mathbf{S_W} = \sum_{\mathsf{c}\in|\mathcal{C}|} Pr(\mathsf{c})\,\mathbf{\Sigma}^{(\mathsf{c})}, \quad \mathbf{S_B} = \sum_{\mathsf{c}\in|\mathcal{C}|} Pr(\mathsf{c})\,(\underline{\mu}^{(\mathsf{c})}-\underline{\mu})(\underline{\mu}^{(\mathsf{c})}-\underline{\mu})^\top,$$

respectively. Above, $Pr(\mathsf{c})=L_\mathsf{c}/L$, $\mathsf{c}\in\mathcal{C}$, is the class empirical probability mass function, where $L_\mathsf{c}=\Sigma_{l=1}^{L}\delta_{c(l)}^{\mathsf{c}}$, and $\delta_i^j=1$, if $i=j$; 0, otherwise. In addition, each class sample mean is

$$\underline{\mu}^{(\mathsf{c})} = [\,\mu_1^{(\mathsf{c})},...,\mu_{d^{(\mathrm{II})}}^{(\mathsf{c})}\,]^\top, \text{ where } \mu_d^{(\mathsf{c})} = \frac{1}{L_\mathsf{c}}\sum_{l=1}^{L}\delta_{c(l)}^{\mathsf{c}}\,x_{l,d}^{(\mathrm{II})}\,,$$

and each class sample covariance is $\mathbf{\Sigma}^{(\mathsf{c})}$, with elements given by

$$\sigma_{d,d'}^{(\mathsf{c})} = \frac{1}{L_\mathsf{c}}\sum_{l=1}^{L}\delta_{c(l)}^{\mathsf{c}}(\,x_{l,d}^{(\mathrm{II})}-\mu_d^{(\mathsf{c})})\,(\,x_{l,d'}^{(\mathrm{II})}-\mu_{d'}^{(\mathsf{c})})\,,$$

for $d,d'=1,...,d^{(\mathrm{II})}$. Finally, $\underline{\mu}=\Sigma_{\mathsf{c}\in\mathcal{C}}Pr(\mathsf{c})\,\underline{\mu}^{(\mathsf{c})}$ is the total sample mean. We subsequently compute the *generalized* eigenvalues and *right* eigenvectors of the matrix pair $(\mathbf{S_B},\mathbf{S_W})$ that satisfy $\mathbf{S_B}\mathbf{F}=\mathbf{S_W}\mathbf{F}\mathbf{D}$ [11], [16]. Matrix $\mathbf{F}=[\,\underline{F}_1,...,\underline{F}_{d^{(\mathrm{II})}}]$ has as columns the generalized eigenvectors. Let the $D^{(\mathrm{II})}$ largest eigenvalues be located at the $j_1,...,j_{D^{(\mathrm{II})}}$ diagonal positions of $\mathbf{D}$. Then, given data vector $\underline{x}_t^{(\mathrm{II})}$, we extract its feature vector of length $D^{(\mathrm{II})}$ as $\underline{y}_t^{(\mathrm{II})}=\mathbf{P}^{(\mathrm{II})}\underline{x}_t^{(\mathrm{II})}$, where $\mathbf{P}^{(\mathrm{II})}=[\,\underline{F}_{j_1},...,\underline{F}_{j_{D^{(\mathrm{II})}}}\,]^\top$.

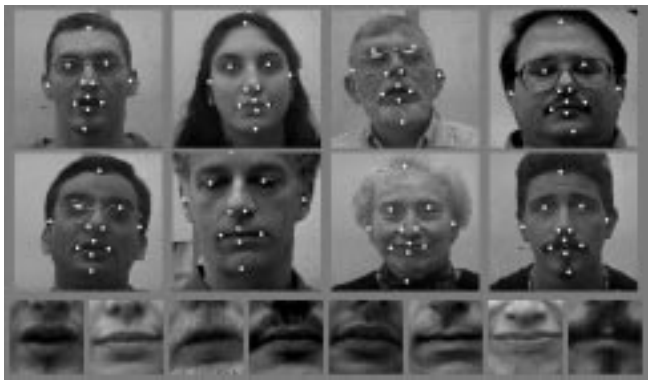

Figure 1: ROI extraction examples. *Upper rows*: Example video frames from 8 database subjects, with detected facial features superimposed. *Lower row*: Corresponding extracted mouth regions of interest.

## 4. MAXIMUM LIKELIHOOD DATA ROTATION

In difficult classification problems, such as large vocabulary continuous speech recognition, many high dimensional Gaussian densities are used to model the observation class conditional probability distribution. Due to lack of sufficient data, diagonal covariances are typically assumed, although the data class observation vector covariance matrices $\mathbf{\Sigma}^{(\mathsf{c})}$, $\mathsf{c}\in\mathcal{C}$, are not diagonal. To alleviate this, we employ the maximum likelihood linear transform (MLLT) algorithm. MLLT provides a *non-singular* matrix $\mathbf{P}^{(\mathrm{III})}$ that "rotates" feature vector $\underline{x}_t^{(\mathrm{III})}=\underline{y}_t^{(\mathrm{II})}$, of dimension $d^{(\mathrm{III})}=D^{(\mathrm{II})}$, obtained by the two first stages of the proposed cascade algorithm as discussed in Sections 2 and 3. The final feature vector is of length $D^{(\mathrm{III})}=d^{(\mathrm{III})}$, and it is derived as $\underline{y}_t^{(\mathrm{III})}=\mathbf{P}^{(\mathrm{III})}\underline{x}_t^{(\mathrm{III})}$.

MLLT considers the observation data likelihood at the original feature space. The desired matrix $\mathbf{P}^{(\mathrm{III})}$ is obtained as [14]

$$\mathbf{P}^{(\mathrm{III})} = \arg\max_{\mathbf{P}}\{\mathsf{det}(\mathbf{P})^L\prod_{\mathsf{c}\in\mathcal{C}}(\mathsf{det}(\mathsf{diag}(\mathbf{P}\mathbf{\Sigma}^{(\mathsf{c})}\mathbf{P}^\top)))^{-L_\mathsf{c}/2}\}\,,$$

where $\mathsf{det}(\bullet)$ and $\mathsf{diag}(\bullet)$ denote matrix *determinant* and *diagonal*, respectively. Equivalently,

$$\sum_{\mathsf{c}\in\mathcal{C}} L_\mathsf{c}(\mathsf{diag}(\mathbf{P}^{(\mathrm{III})}\mathbf{\Sigma}^{(\mathsf{c})}\mathbf{P}^{(\mathrm{III})\top}))^{-1}\,\mathbf{P}^{(\mathrm{III})}\mathbf{\Sigma}^{(\mathsf{c})} = L(\mathbf{P}^{(\mathrm{III})\top})^{-1}\,.$$

The latter can be solved numerically [14].

## 5. THE AUTOMATIC SPEECHREADING SYSTEM

### 5.1. Region of interest extraction

We use the statistical face tracking algorithm reported in [17] to first estimate the face location, size, and orientation at each video frame, and to subsequently locate a number of facial features. Five located lip contour points are used to estimate the mouth center and its size at every video frame. The mouth center estimate is smoothed over neighboring frames using median filtering to obtain the ROI center $(m_t,n_t)$, whereas the mouth size estimate is averaged over each utterance. A size normalized ROI is then extracted as in (1), with $M=N=64$, and $K=1$, in order to allow for fast DCT and DWT implementation [15] (see also Fig. 1).

| $\mathbf{P}^{(\mathrm{I})} \rightarrow$ | DCT | | DWT | | PCA | |
|---|---|---|---|---|---|---|
| STAGE $\downarrow$ | GMM | HMM | GMM | HMM | GMM | HMM |
| I ($\mathbf{P}^{(\mathrm{I})}$) | 27.31 | 37.94 | 28.01 | 37.37 | 26.88 | 37.28 |
| II (LDA) | 32.94 | 38.81 | 31.33 | 38.15 | 31.72 | 39.26 |
| III (MLLT) | 34.64 | 41.48 | 33.67 | 41.80 | 32.65 | 41.28 |

Table 1: Test set visual-only phonetic classification accuracy (%) using each stage of the proposed algorithm and DCT, DWT, or PCA features at the first stage. Both GMM and segmental based HMM classification are reported (5 mixtures per GMM class or HMM state are used).

### 5.2. Cascade algorithm implementation

*Stage I* (image transform) is applied to each ROI vector $\underline{x}_t^{(\mathrm{I})}$ at a rate of 60 Hz. To simplify subsequent LDA and MLLT training, as well as bimodal (audio-visual) fusion, we interpolate the resulting features $\underline{y}_t^{(\mathrm{I})}$ to the audio feature rate, 100 Hz. Furthermore, in order to account for lighting and other variations, we apply *feature mean normalization* (FMN) by simply subtracting the feature mean computed over the utterance length $T$, i.e., $\underline{y}_t^{(\mathrm{I})} \leftarrow \underline{y}_t^{(\mathrm{I})} - \Sigma_{t'=1}^{T} \underline{y}_{t'}^{(\mathrm{I})}/T$. When using Stage I as the sole visual front end, and in order to capture visual speech dynamics, we augment $\underline{y}_t^{(\mathrm{I})}$ by its first and second-order derivatives, each computed over a 9-frame window [4], [18]. In such case, we consider $D^{(\mathrm{I})} = 54 = 3 \times 18$ .

At *Stage II* (LDA) and *Stage III* (MLLT), and in order to train matrices $\mathbf{P}^{(\mathrm{II})}$, $\mathbf{P}^{(\mathrm{III})}$, we consider approximately 3,400 context dependent sub-phonetic classes. We label vectors $\underline{x}_t^{(\mathrm{II})}$, $\underline{x}_t^{(\mathrm{III})}$, by means of *Viterbi forced segmentation* [18], based on the audio channel and an available audio-only *hidden Markov Model* (HMM). In the current front end implementation, we use $D^{(\mathrm{I})} = 24$, $D^{(\mathrm{II})} = D^{(\mathrm{III})} = 41$, and $J = 15$.

### 5.3. Phonetic classification

We consider 52 phoneme classes, and, for visual-only classification, also 27 *viseme* classes, both listed in [9]. The training set utterance alignments are used to bootstrap visual-only *Gaussian mixture models* (GMMs), using the *expectation-maximization* (EM) algorithm [18]. The GMM class conditional probability is

$$Pr(\underline{y}_t|\mathsf{c}) = \sum_{m=1}^{M_\mathsf{c}} w_{\mathsf{c}m}\, \mathcal{N}_D(\underline{y}_t;\, \underline{\mu}_{\mathsf{c}m},\, \underline{\sigma}_{\mathsf{c}m}),\ \text{for all } \mathsf{c} \in \mathcal{C}, \quad (2)$$

where *mixture weights* $w_{\mathsf{c}m}$ are positive adding up to one, $M_\mathsf{c}$ denotes the number of class mixtures, and $\mathcal{N}_D(\underline{y}; \underline{\mu}, \underline{\sigma})$ is the $D$–variate normal distribution with mean $\underline{\mu}$ and diagonal covariance $\underline{\sigma}$.

*Frame* level classification accuracy is calculated by comparing, at each instance of $t$ , the audio forced alignment class label to its *maximum-a-posteriori* (MAP) class estimate, obtained as

$$\mathsf{c}_t = \arg\max_{\mathsf{c}\in\mathcal{C}} \{Pr(\underline{y}_t|\mathsf{c})\, Pr(\mathsf{c})\}. \quad (3)$$

In (3), the smoothed class prior $Pr(\mathsf{c}) = (L_\mathsf{c}+1)/(L+|\mathcal{C}|)$ , $\mathsf{c} \in \mathcal{C}$, is used.

Significantly superior frame classification accuracy is obtained, if the class boundaries of the test utterances are assumed known (*segmental* approach). In this case, we consider 52 phoneme (or, 27 viseme) class HMMs, each consisting of three *states* per class and state conditional probabilities as in (2). Such HMMs are trained using the EM algorithm [18]. MAP estimation becomes Viterbi decoding over each utterance phone segment [18].

| | VI-27 (G,H) | VI-52 (G,H) | AU (G,H) | AV (G,H) |
|---|---|---|---|---|
| I | 44.47 , 57.64 | 31.77 , 46.07 | 62.78 , 80.52 | 64.73 , 83.51 |
| II | 47.66 , 58.56 | 35.74 , 46.52 | | 66.03 , 83.57 |
| III | 49.29 , 59.77 | 37.71 , 48.85 | | 66.20 , 84.04 |

Table 2: Test set visual-only visemic (VI-27) and phonetic (VI-52) classification accuracy (%) using each stage of the visual front end and DCT features. Audio-only (AU) and audio-visual (AV) phonetic classification accuracy are also depicted ($\gamma_\mathrm{A} = 0.675$, $\gamma_\mathrm{V} = 0.325$ are used in (4)). Both GMM (G) and segmental based HMM (H) classification are reported (64 mixtures per GMM class or HMM state are used).

It is finally of interest to consider audio-visual phonetic classification. A number of classifier fusion techniques can be used [19]. In this work, we employ a simple algorithm that considers [12]

$$\mathrm{Score}(\underline{y}_t^{(\mathrm{AV})}|\mathsf{c}) = Pr(\underline{y}_t^{(\mathrm{A})}|\mathsf{c})^{\gamma_\mathrm{A}}\, Pr(\underline{y}_t^{(\mathrm{V})}|\mathsf{c})^{\gamma_\mathrm{V}}, \quad (4)$$

where $\gamma_\mathrm{A}, \gamma_\mathrm{V} \geq 0$ , and $\underline{y}_t = [\, \underline{y}_t^{(\mathrm{A})\,\top},\ \underline{y}_t^{(\mathrm{V})\,\top}\,]^\top$ denotes the concatenation of time synchronous audio and visual features. The audio front end reported in [9] is used.

## 6. DATABASE AND EXPERIMENTS

We have been collecting a multi-subject, continuous, large vocabulary, audio-visual database, using ViaVoice$^{\mathrm{TM}}$ training utterance scripts. Currently, it consists of 162 subjects and close to 30 hours of speech (15,350 utterances). The database contains full frontal face color video of the subjects with minor face-camera distance and lighting variations (see also Fig. 1). The video is captured at a resolution of 704 $\times$ 480 pixels (interleaved), a frame rate of 60 Hz, and is MPEG2 encoded to about 0.5 MByte/sec. The audio is captured at 16 KHz, and it is time-synchronous to the video stream. For the sake of faster experimentation, we randomly select 20 database utterances per subject and randomly split them into 16 training and 4 test utterances per subject, thus creating a *multi-subject* 2,592 utterance *training set* (5.5 hours) and a 648 utterance *test set* (1.4 hour).

We first compare the phonetic classification performance of the various algorithm stages discussed in Sections 2-4. As shown in Table 1, and regardless of the visual feature extraction method employed at Stage I (DCT, DWT, or PCA), using LDA (Stage II) results in significant accuracy improvement (20% in the DCT GMM based classification case, for example). Using the additional MLLT data rotation (Stage III) further improves performance (7% in the DWT case). Both stages combined can account for up to 27% accuracy improvement over the image transform only (Stage I) visual front end (DCT GMM based classification case, for example).

Overall, the performance of each algorithm stage does not vary significantly when using any of the three image transforms (DCT, DWT, or PCA) considered in this paper. The DCT slightly outperforms the DWT and somewhat more PCA (34.64%, 33.67%, and 32.65% Stage III accuracy, respectively). Both DCT and DWT allow fast implementations, whereas PCA is computationally expensive, given the large dimensionality of the mouth ROI typically required. Clearly therefore, DCT and DWT are preferable to the use of PCA.

In Table 2, we report improved visual-only classification accuracy using a classifier with 64 mixtures per GMM class or HMM state. Such a system achieves a 48.85% segmental based (HMM)

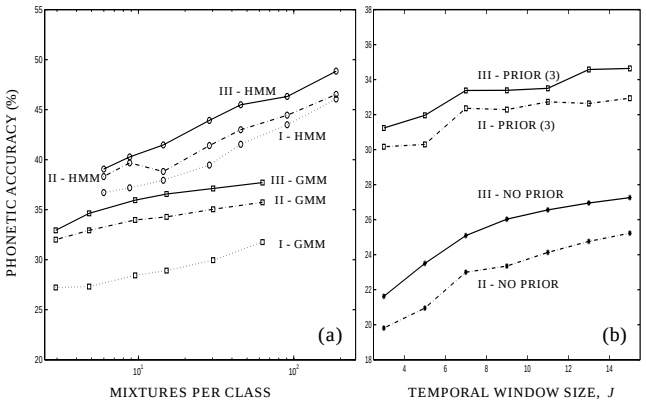


Figure 2: Visual-only phonetic classification accuracy using a DCT based visual front end, as a function of: (a) number of mixtures per GMM or HMM phone class; (b) temporal window size $J$ at Stage II (GMM only, with or without prior in (3)).

visual-only phonetic classification accuracy. This corresponds to 59.77% visemic classification accuracy. For completeness, audio-only and audio-visual phonetic classification accuracies are also reported. Notice that both Stages II and III improve audio-visual phonetic classification over Stage I. Indeed, the reported 80.52% clean audio-only accuracy improves to 83.51%, 83.57%, and 84.04%, when Stages I, II, and III of the visual modality front end are respectively used to augment the audio modality by means of (4).

Classification using various size GMM/HMM systems is addressed in Fig. 2(a). Clearly, larger systems perform better, but the relative performance of the three algorithm stages remains mostly unchanged. Fig. 2(b) depicts the dependence of phonetic classification accuracy on the size $J$ of the temporal window used to capture the visual speech dynamics at Stage II.[1] Wider temporal windows improve performance, however at an increased computational cost.

Finally, it is worth reporting that feature mean normalization (FMN) improves classification performance. Indeed, DCT feature based Stage I classification accuracy without FMN is only 25.99%, compared to 27.31% when FMN is applied (see also Table 1). Furthermore, bypassing Stage II of the algorithm degrades performance: A DCT based Stage I, followed by MLLT, results to a 31.86% accuracy, as compared to 34.64%, obtained when all three stages are used. Clearly therefore, the proposed three-stage cascade approach is superior.

## 7. SUMMARY

We propose a new pixel based visual front end for automatic recognition of visual speech. It consists of a discrete cosine, or wavelet, transform of the video region of interest, followed by a linear discriminant data projection, and a maximum likelihood based data rotation. We have demonstrated that all three stages contribute to accuracy gains in phone classification that can reach up to 27% improvement, as compared to an image transform only based visual front end. Overall, we achieve a 49% (38%) visual-only frame level phonetic classification accuracy with (without) use of test set phone boundaries. In addition, the proposed algorithm results in improved audio-visual phonetic classification.

[1] For the sake of clarity, we also depict GMM classification using a uniform prior in (3).

## 8. ACKNOWLEDGMENTS

The authors would like to acknowledge contributions to this work by A.W. Senior for the face tracking algorithm and R. Gopinath for insights in the maximum likelihood linear transform algorithm, both with the IBM Thomas J. Watson Research Center.

## 9. REFERENCES

[1] M.E. Hennecke, D.G. Stork, and K.V. Prasad, "Visionary speech: Looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 331-349, 1996.

[2] C. Bregler and Y. Konig, "'Eigenlips' for robust speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, Adelaide, pp. 669-672, 1994.

[3] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel, "Toward movement-invariant automatic lip-reading and speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, Detroit, pp. 109-112, 1995.

[4] G. Potamianos, H.P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," *Proc. Int. Conf. Image Process.*, Chicago, pp. 173-177, 1998.

[5] N.M. Brooke, "Talking heads and speech recognizers that can see: The computer processing of visual speech signals," in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 351-371, 1996.

[6] M.S. Gray, J.R. Movellan, and T.J. Sejnowski, "Dynamic features for visual speech-reading: A systematic comparison," in *Advances in Neural Information Processing Systems 9*, M.C. Mozer, M.I. Jordan, and T. Petsche eds., MIT Press, Cambridge, pp. 751-757, 1997.

[7] G.I. Chiou and J.-N. Hwang, "Lipreading from color video," *IEEE Trans. Image Process.*, vol. 6, pp. 1192-1195, 1997.

[8] J. Luettin, "Towards speaker independent continuous speechreading," *Proc. Eurospeech*, Rhodes, pp. 1991-1994, 1997.

[9] S. Basu, C. Neti, N. Rajput, A. Senior, L. Subramaniam, and A. Verma, "Audio-visual large vocabulary continuous speech recognition in the broadcast domain," *Proc. Works. Multimedia Signal Process.*, Copenhagen, pp. 475-481, 1999.

[10] I. Matthews, J.A. Bangham, and S. Cox, "Audio-visual speech recognition using multiscale nonlinear image decomposition," *Proc. Int. Conf. Speech Lang. Process.*, Philadelphia, pp. 38-41, 1996.

[11] G. Potamianos and H.P. Graf, "Linear discriminant analysis for speechreading," *Proc. Works. Multimedia Signal Process.*, Los Angeles, pp. 221-226, 1998.

[12] G. Potamianos and H.P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," *Proc. Int. Conf. Acoust. Speech Signal Process.*, Seattle, pp. 3733-3736, 1998.

[13] I. Daubechies, *Wavelets*. S.I.A.M., Philadelphia, 1992.

[14] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," *Proc. Int. Conf. Acoust. Speech Signal Process.*, Seattle, pp. 661-664, 1998.

[15] W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge University Press, Cambridge, 1988.

[16] G.H. Golub and C.F. Van Loan, *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 1983.

[17] A.W. Senior, "Face and feature finding for a face recognition system," *Proc. Int. Conf. Audio and Video-based Biometr. Person Authent.*, Washington, pp. 154-159, 1999.

[18] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, 1993.

[19] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 4-37, 2000.