International Journal of Speech Technology, Special Issue on Multimedia, 4, 193-208, 2001

A cascade visual front end for speaker independent automatic speechreading

GERASIMOS POTAMIANOS, CHALAPATHY NETI, GIRIDHARAN IYENGAR,
ANDREW W. SENIOR AND ASHISH VERMA*

Human Language Technologies, IBM Thomas J. Watson Research Center,
Yorktown Heights, NY 10598, U.S.A.

gpotam@us.ibm.com
cneti@us.ibm.com
giyengar@us.ibm.com
aws@watson.ibm.com
vashish@in.ibm.com

Received May, 2000; Revised March 2001

Abstract. We propose a three-stage pixel based visual front end for automatic speechreading (lipreading) that results in significantly improved recognition performance of spoken words or phonemes. The proposed algorithm is a cascade of three transforms applied on 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 on 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 that optimizes the likelihood of the observed data under the assumption of their class-conditional multi-variate normal distribution with diagonal covariance. We apply the algorithm to visual-only 52-class phonetic and 27-class visemic classification on a 162-subject, 8-hour long, large-vocabulary, continuous speech audio-visual database. 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 60% (49%) visual-only frame-level visemic classification accuracy with (without) use of test set viseme boundaries. In addition, we report improved audio-visual phonetic classification over the use of a single-stage image transform visual front end. Finally, we discuss preliminary speech recognition results.

Keywords: automatic speechreading, lipreading, phonetic classification, discrete cosine transform, linear discriminant analysis, maximum likelihood linear transform, audio-visual speech recognition.

*IBM India Research Lab, New Delhi, India 110016

1. Introduction

Automatic speech recognition (ASR) by using the image sequence (video) of the speaker's lips, referred to as *automatic lipreading*, or *speechreading*, has recently attracted significant interest (Stork and Hennecke, 1996; Teissier, et al., 1999; Dupont and Luettin, 2000; Neti, et al., 2000; Chen, 2001). Much of this interest is motivated by the fact that the *visual* modality contains some complementary information to the *audio* modality (Massaro and Stork, 1998), as well as by the way that humans "fuse" *audio-visual* stimuli to recognize speech (McGurk and MacDonald, 1976; Summerfield, 1987). Not surprisingly, automatic speechreading has been shown to improve traditional audio-only ASR performance over a wide range of conditions (Adjoudani and Benoît, 1996; Rogozan, et al., 1997; Potamianos and Graf, 1998b; Teissier, et al., 1999; Dupont and Luettin, 2000; Neti, et al., 2000). Such performance gains are particularly impressive in noisy environments, where traditional ASR performs poorly. Coupled with the diminishing cost of quality video capturing systems, this fact makes automatic speechreading tractable for achieving robust ASR in certain scenarios (Hennecke, et al., 1996).

Two issues are key in the design and the resulting performance of audio-visual ASR systems. The first is the *visual front end* algorithm, namely the extraction of appropriate visual features that contain relevant speech information, given the video of the speaker's face. Various sets of visual features have been proposed in the literature that, in general, can be grouped into *lip-contour* (*shape*) based and *pixel* (*appearance*) based features (Hennecke, et al., 1996). In the first approach, the speaker's inner and (or) outer lip contours are extracted from the image sequence. A parametric (Hennecke, et al., 1996; Chiou and Hwang, 1997), or statistical (Dupont and Luettin, 2000), lip-contour model is obtained, and the model parameters are used as visual features. Alternatively, lip-contour geometric features are used, such as mouth height and width, as in (Petajan, 1984; Adjoudani and Benoît, 1996; Chandramohan and Silsbee, 1996; Rogozan, et al., 1997). In contrast, in the pixel based approach, the entire image containing the speaker's mouth is considered as the *region-of-interest* (ROI) for lipreading, and appropriate transformations of its pixel values are used as visual features. For example, Gray, et al. (1997) use video frame ROI differences, while Matthews, et al. (1996) suggest a nonlinear ROI image decomposition for feature extraction. Such approach is motivated by the fact that, in addition to the lips, visible parts of the mouth cavity, such as the teeth and tongue, as well as certain facial muscle movements, are informative about visual speech (Summerfield, et al., 1989). Often, the two approaches are combined into joint shape and appearance features, such as the *active appearance models* in (Matthews, 1998) and the visual front ends of (Chiou and Hwang, 1997; Dupont and Luettin, 2000). To-date, there exists little comparative work on the relative performance of shape versus appearance based visual features (Potamianos, et al., 1998; Neti, et al., 2000).

The second factor that affects the performance of automatic speechreading systems is the audio-visual *"integration"* strategy, which is used to combine the extracted visual representation with the

traditional audio features into a *bimodal* (audio-visual) speech recognizer (Hennecke, et al., 1996). This is also referred to as audio-visual *fusion*, and it constitutes an instance of the general *classifier combination* problem (Jain, et al., 2000). A number of techniques have appeared in the literature for audio-visual integration, which can be broadly grouped into *feature fusion* and *decision fusion* methods. The first ones are based on training a *single* classifier (i.e., of the same form as the audio- and visual-only classifiers) on the concatenated vector of audio and visual features, or on any appropriate transformation of it. Such methods include feature concatenation (Adjoudani and Benoît, 1996), dominant and motor recording (Teissier, et al., 1999), hierarchical linear discriminant feature extraction (Potamianos, et al., 2001), and feature weighting (Teissier, et al., 1999; Chen, 2001). In contrast, decision fusion algorithms utilize the two single-modality (audio- and visual-only) classifier outputs to recognize audio-visual speech. Typically, this is achieved by linearly combining the class-conditional observation log-likelihoods of the two classifiers into a joint audio-visual classification score, using appropriate weights that capture the reliability of each single-modality classifier, or data stream (Hennecke, et al., 1996; Rogozan, et al., 1997; Potamianos and Graf, 1998b; Dupont and Luettin, 2000; Neti, et al., 2000). This combination can be performed at various possible levels, the one extreme being the feature frame level, assuming time-synchronous audio and visual features (*"early"* integration), whereas the other extreme being the *"late"* integration at the utterance level (Adjoudani and Benoît, 1996).

In this paper, we concentrate on the first aspect of the audio-visual speech recognition problem, namely the issue of visual feature extraction. In addition to the obvious importance of the visual front end design to automatic speechreading, the problem is of interest by itself: How accurately can one hope to recognize speech using the visual information alone? Furthermore, the visual front end is not only limited to automatic speechreading: Lip-region visual features can readily be used in multimodal biometric systems (Wark and Sridharan, 1998; Fröba, et al., 1999), as well as to detect speech activity and intent to speak (De Cuetos, et al., 2000), among others.

In particular, we investigate the pixel (appearance) based approach to the visual front end for automatic speechreading, proposing a three-stage algorithm that consists of a cascade of three transforms applied on the ROI data vector. The *first* algorithm stage is a traditional image transform, such as the *discrete cosine* (DCT), suggested in the context of speechreading in (Duchnowski, et al., 1995), the *discrete wavelet* (DWT), as in (Potamianos, et al., 1998), and the *Karhunen-Loève* transform (KLT), or *principal component analysis* (PCA), used, among others, in (Bregler and Konig, 1994; Brooke, 1996; Basu, et al., 1999; Chiou and Hwang, 1997; Dupont and Luettin, 2000). This first algorithm stage seeks data dimensionality reduction through data *compression*. The *second* stage is a *linear discriminant analysis* (LDA) data projection (Rao, 1965), that seeks optimal classification performance and further data dimensionality reduction. In the literature, LDA has been used as a stand-alone visual front end in (Duchnowski, et al., 1995; Potamianos and Graf, 1998a), and as the second and final visual front end stage, following the application of PCA, in (Wark and Sridharan, 1998; Basu, et al., 1999). In our proposed algorithm, and in order to capture

dynamic visual speech information, LDA is applied on the concatenation of a small number of consecutive DCT feature vectors. The final, *third* stage of the proposed algorithm is a *maximum likelihood linear transformation* (MLLT) aimed at optimizing the observed data likelihood under the assumption of their class-conditional multi-variate normal distribution with diagonal covariance (Gopinath, 1998).

This proposed three-stage 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. Furthermore, the cascade algorithm is tested on a large-vocabulary, continuous speech audio-visual corpus suitable for ASR, namely on a 162-subject, 8-hour long subset of the IBM *ViaVoice*$^{\rm TM}$ *Audio-Visual* (VVAV) database (Neti, et al., 2000), thus allowing statistically meaningful comparisons and conclusions.

The paper is structured as follows: The three algorithm stages are discussed in Section 2, each in a separate subsection. Specifics of all components of our speechreading system are presented in Section 3. Such include a brief description of the face detection algorithm used, the mouth ROI extraction method, the cascade algorithm implementation, as well as the statistical classifier used in the phonetic, or visemic, automatic recognition of speech. Our audio-visual database and experimental results are reported in Section 4. Finally, conclusions and a short discussion follow in Section 5.

2. A Three-Stage Feature Extraction Algorithm

Let us assume, that, for every video frame V_t at instant t, a two-dimensional region-of-interest (ROI) centered around the speaker's mouth center $(m_t\,,n_t)$ is extracted by means of an appropriate *face detection* and facial part location estimation algorithm (Graf, et al., 1997; Senior, 1999). Such an algorithm is described in more detail in Section 3.1. The ROI pixel values are placed into the vector[1]

$$\mathbf{x}_t^{(\rm I)} \leftarrow \{\, V_t(m\,,n\,):\; m_t - \lfloor M/2 \rfloor \le m < m_t + \lceil M/2 \rceil\,,\; n_t - \lfloor N/2 \rfloor \le n < n_t + \lceil N/2 \rceil \,\}\,, \qquad (1)$$

of length $d^{(\rm I)} = MN$. The proposed algorithm seeks three matrices, $\mathbf{P}^{(\rm I)}$, $\mathbf{P}^{(\rm II)}$, and $\mathbf{P}^{(\rm III)}$, that, when applied on the data vector $\mathbf{x}_t^{(\rm I)}$, in a cascade fashion, they result in a "compact" visual feature vector $\mathbf{y}_t^{(\rm III)}$ of dimension $D^{(\rm III)} \ll d^{(\rm I)}$ (see also Figure 1). Such vector should contain most discriminant and relevant to visual speech information, according to criteria defined in Sections 2.1, 2.2, and 2.3. Each such matrix $\mathbf{P}^{(\bullet)}$ is of dimension $D^{(\bullet)} \times d^{(\bullet)}$, where $\bullet = {\rm I, II, III}$. To obtain matrices $\mathbf{P}^{(\bullet)}$, L *training* examples are given, denoted by $\mathbf{x}_l^{(\rm I)}$, for $l = 1,...,L$.

FIG. 1 HERE

[1] Throughout this work, boldface lowercase symbols denote column vectors, and boldface capital symbols denote matrices.

2.1. Stage I: Image Transform Based Data Compression

At the first algorithm stage, we seek a $D^{(\mathrm{I})} \times d^{(\mathrm{I})}$-dimensional *linear transform* matrix $\mathbf{P}^{(\mathrm{I})} = [\mathbf{p}_1,...,\mathbf{p}_{D^{(\mathrm{I})}}]^\top$, such that the transformed data vector $\mathbf{y}_t^{(\mathrm{I})} = \mathbf{P}^{(\mathrm{I})}\mathbf{x}_t^{(\mathrm{I})}$ contains most speechreading information in its $D^{(\mathrm{I})} \ll d^{(\mathrm{I})}$ elements, thus achieving significant data compression. This can be quantified by seeking such elements to maximize the total *energy* of the transformed training feature vectors $\mathbf{y}_l^{(\mathrm{I})} = \mathbf{P}^{(\mathrm{I})}\mathbf{x}_l^{(\mathrm{I})}$, for $l = 1,...,L$, given the desired output vector length $D^{(\mathrm{I})}$ (see (2), below). Alternatively, one can seek to minimize the *mean square error* between the training data vectors $\mathbf{x}_l^{(\mathrm{I})}$ and their reconstruction based on $\mathbf{y}_l^{(\mathrm{I})}$, for $l = 1,...,L$, as in Section 2.1.2.

2.1.1. 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) and the discrete wavelet transform (DWT) implemented by means of the Daubechies class wavelet filter of approximating order 2 (Daubechies, 1992; Press, et al., 1988). Let square matrix $\mathbf{B} = [\mathbf{b}_1,...,\mathbf{b}_{d^{(\mathrm{I})}}]^\top$ denote the image transform matrix, where $\bullet^\top$ denotes vector or matrix *transpose*. Then, matrix $\mathbf{P}^{(\mathrm{I})}$ contains as its rows the rows of $\mathbf{B}$ that maximize the transformed data energy

$$\sum_{d=1}^{D^{(\mathrm{I})}} \sum_{l=1}^{L} < \mathbf{x}_l^{(\mathrm{I})}, \mathbf{b}_{j_d} >^2, \tag{2}$$

where $j_d \in \{1,...,d^{(\mathrm{I})}\}$ are disjoint, and $<\bullet\,,\bullet>$ denotes vector *inner product*. Obtaining the optimal values of j_d, for $d = 1,...,D^{(\mathrm{I})}$, that maximize (2) is straightforward. It is important to note that both DCT and DWT allow fast implementations, when M and N are powers of 2 (Press, et al., 1988). It is therefore advantageous to choose such values in (1).

2.1.2. Principal Component Analysis. Principal component analysis (PCA) achieves optimal data compression in the minimum mean square error sense, by projecting the data vectors onto the directions of their greatest variance. However, the problem of appropriate data *scaling* arises when applying PCA to classification (Chatfield and Collins, 1980). In our experiments, we found it beneficial to scale the data according to their inverse variance. Namely, we first compute the data *mean* and *variance* as

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

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

$$r_{d,d'} = \frac{1}{L}\sum_{l=1}^{L} \frac{(x_{l,d}^{(\mathrm{I})} - m_d)}{\sigma_d}\,\frac{(x_{l,d'}^{(\mathrm{I})} - m_{d'})}{\sigma_{d'}}\,, \quad \text{for } d, d' = 1,...,d^{(\mathrm{I})}\,. \tag{3}$$

Next, we *diagonalize* the correlation matrix as $\mathbf{R} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^\top$ (Chatfield and Collins, 1980; Golub and Van Loan, 1983), where $\mathbf{A} = [\mathbf{a}_1,...,\mathbf{a}_{d^{(\mathrm{I})}}]$ 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 of $\mathbf{\Lambda}$. Then, given data vector $\mathbf{x}_t^{(\mathrm{I})}$, we normalize it element-wise as $x_{t,d}^{(\mathrm{I})} \leftarrow (x_{t,d}^{(\mathrm{I})} - m_d)/\sigma_d$, and subsequently we extract its feature vector $\mathbf{y}_t^{(\mathrm{I})} = \mathbf{P}^{(\mathrm{I})}\mathbf{x}_t^{(\mathrm{I})}$, where $\mathbf{P}^{(\mathrm{I})} = [\,\mathbf{a}_{j_1},...,\mathbf{a}_{j_{D^{(\mathrm{I})}}}\,]^\top$.

2.2. Stage II: 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 on the concatenation of J consecutive image transformed feature vectors

$$\mathbf{x}_t^{(\mathrm{II})} = [\,\mathbf{y}_{t-\lfloor J/2\rfloor}^{(\mathrm{I})\top},...,\mathbf{y}_t^{(\mathrm{I})\top},...,\mathbf{y}_{t+\lceil J/2\rceil-1}^{(\mathrm{I})\top}\,]^\top, \tag{4}$$

of length $d^{(\mathrm{II})} = D^{(\mathrm{I})}J$ (see also Figure 1).

In general, LDA (Rao, 1965) assumes that a set of *classes* $\mathcal{C}$ is a-priori given, as well as that the training set data vectors $\mathbf{x}_l^{(\mathrm{II})}$, $l = 1,...,L$, are *labeled* as $c(l) \in \mathcal{C}$. LDA seeks a projection $\mathbf{P}^{(\mathrm{II})}$, such that the projected training sample $\{\,\mathbf{P}^{(\mathrm{II})}\,\mathbf{x}_l^{(\mathrm{II})},\ l = 1,...,L\,\}$ is "well separated" into the set of classes $\mathcal{C}$. Formally, $\mathbf{P}^{(\mathrm{II})}$ maximizes

$$Q(\mathbf{P}^{(\mathrm{II})}) = \frac{\mathsf{det}\,(\,\mathbf{P}^{(\mathrm{II})\top}\,\mathbf{S}_B\,\mathbf{P}^{(\mathrm{II})}\,)}{\mathsf{det}\,(\,\mathbf{P}^{(\mathrm{II})\top}\,\mathbf{S}_W\,\mathbf{P}^{(\mathrm{II})}\,)}\,, \tag{5}$$

where $\mathsf{det}(\bullet)$ denotes matrix *determinant*. In (5), $\mathbf{S}_W$, $\mathbf{S}_B$ denote the *within-class scatter* and *between-class scatter* matrices of the training sample. These matrices are given by

$$\mathbf{S}_W = \sum_{\mathsf{c}\in\mathcal{C}} Pr(\mathsf{c})\,\mathbf{\Sigma}^{(\mathsf{c})}, \quad \text{and} \quad \mathbf{S}_B = \sum_{\mathsf{c}\in\mathcal{C}} Pr(\mathsf{c})\,(\,\mathbf{m}^{(\mathsf{c})} - \mathbf{m}\,)\,(\,\mathbf{m}^{(\mathsf{c})} - \mathbf{m}\,)^\top, \tag{6}$$

respectively. In (6), $Pr(\mathsf{c}) = L_\mathsf{c}/L$, $\mathsf{c} \in \mathcal{C}$, 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

$$\mathbf{m}^{(\mathsf{c})} = [\,m_1^{(\mathsf{c})},...,m_{d^{(\mathrm{II})}}^{(\mathsf{c})}\,]^\top, \quad \text{where } m_d^{(\mathsf{c})} = \frac{1}{L_\mathsf{c}}\sum_{l=1}^{L}\delta_{\,c(l)}^{\,\mathsf{c}}\,x_{l,d}^{(\mathrm{II})}\,, \quad \text{for } d = 1,...,d^{(\mathrm{II})}\,,$$

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

$$\sigma_{d,d'}^{(\mathsf{c})} = \frac{1}{L_\mathsf{c}}\sum_{l=1}^{L}\delta_{\,c(l)}^{\,\mathsf{c}}(\,x_{l,d}^{(\mathrm{II})} - m_d^{(\mathsf{c})})\,(\,x_{l,d'}^{(\mathrm{II})} - m_{d'}^{(\mathsf{c})})\,, \quad \text{for } d, d' = 1,...,d^{(\mathrm{II})}\,.$$

Finally, $\mathbf{m} = \Sigma_{\mathsf{c}\in\mathcal{C}}\,Pr(\mathsf{c})\,\mathbf{m}^{(\mathsf{c})}$, denotes the total sample mean.

To maximize (5), we compute the *generalized* eigenvalues and *right* eigenvectors of the matrix pair $(\mathbf{S}_B,\mathbf{S}_W)$ that satisfy $\mathbf{S}_B\mathbf{F} = \mathbf{S}_W\mathbf{F}\mathbf{D}$ (Rao, 1965; Golub and Van Loan, 1983). Matrix $\mathbf{F} = [\,\mathbf{f}_1,...,\mathbf{f}_{d^{(\mathrm{II})}}]$ 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 $\mathbf{x}_t^{(\mathrm{II})}$, we extract its feature vector

of length $D^{(\mathrm{II})}$ as $\mathbf{y}_t^{(\mathrm{II})} = \mathbf{P}^{(\mathrm{II})}\mathbf{x}_t^{(\mathrm{II})}$, where $\mathbf{P}^{(\mathrm{II})} = [\,\mathbf{f}_{j_1},...,\mathbf{f}_{j_{D^{(\mathrm{II})}}}\,]^\top$. Vectors $\mathbf{f}_{j_d}$, for $d = 1,...,D^{(\mathrm{II})}$, are the linear discriminant "eigensequences" that correspond to the directions where the data vector projection yields high discrimination among the classes of interest.

We should note that, due to (6), the rank of $\mathbf{S}_B$ is at most $|\mathcal{C}|-1$, where $|\mathcal{C}|$ denotes the number of classes (the *cardinality* of set $\mathcal{C}$); hence we consider $D^{(\mathrm{II})} \leq |\mathcal{C}|-1$. In addition, the rank of the $d^{(\mathrm{II})} \times d^{(\mathrm{II})}$-dimensional matrix $\mathbf{S}_W$ cannot exceed $L-|\mathcal{C}|$ (since the rank of each $\mathbf{\Sigma}^{(\mathsf{c})}$ cannot exceed L_{c}–1), therefore having insufficient training data, with respect to the input feature vector dimension $d^{(\mathrm{II})}$, is a potential problem (matrix $\mathbf{S}_W$ might not be of full rank). In our case, however, first, the input data dimensionality is significantly reduced by using Stage I of the proposed algorithm, and, second, the available training data are of the order $L = O(10^6)$. Therefore, in our experiments, $L-|\mathcal{C}| \gg d^{(\mathrm{II})}$ (see also Section 3.3).

2.3. Stage III: Maximum Likelihood Data Rotation

In difficult classification problems, such as *large-vocabulary, continuous speech recognition* (LVCSR), many high-dimensional multi-variate normal densities are used to model the observation class-conditional probability distribution. Due to lack of sufficient data, diagonal covariances are typically assumed, although the observed data vector class 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 $\mathbf{x}_t^{(\mathrm{III})} = \mathbf{y}_t^{(\mathrm{II})}$, of dimension $d^{(\mathrm{III})} = D^{(\mathrm{II})}$, obtained by the two first stages of the proposed cascade algorithm as discussed in Sections 2.1 and 2.2. The final feature vector is of length $D^{(\mathrm{III})} = d^{(\mathrm{III})}$, and it is derived as $\mathbf{y}_t^{(\mathrm{III})} = \mathbf{P}^{(\mathrm{III})}\mathbf{x}_t^{(\mathrm{III})}$.

MLLT considers the observed data likelihood in the original feature space, under the assumption of diagonal data covariance in the transformed space. The desired matrix $\mathbf{P}^{(\mathrm{III})}$ is obtained by maximizing the *original* data likelihood, namely (Gopinath, 1998)

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

where $\mathsf{diag}(\bullet)$ denotes matrix *diagonal*. Differentiating the logarithm of the objective function with respect to $\mathbf{P}$ and setting it to zero, we obtain (Gopinath, 1998)

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

The latter can be solved numerically (Press, et al., 1988).

3. The Automatic Speechreading System

Until now, we have presented an algorithm for obtaining visual features given the mouth ROI, deferring the issue of extracting the data vector $\mathbf{x}_t^{(\mathrm{I})}$ (see (1)). Obtaining a sequence of such vectors, given the video of the face region of a subject, requires two components: Face detection, and mouth localization, discussed in Section 3.1, and subsequent extraction of normalized pixel values to be placed in (1), discussed in Section 3.2. To complete the presentation of our automatic speechreading system, Section 3.3 is devoted to discussing implementation issues related to the proposed cascade algorithm, whereas Section 3.4 presents the specifics of the statistical classifier used to recognize speech in this work.

3.1. Face Detection and Mouth Location Estimation

We use the face detection and facial feature localization method described in (Senior, 1999). Given a video frame, face detection is first performed by employing a combination of methods, some of which are also used for subsequent face feature finding. A face template size is first chosen (an 11×11-pixel square, here), and an image pyramid over all permissible face locations and scales (given the video frame and face template sizes) is used to search for possible face candidates. This search is constrained by the minimum and maximum allowed face candidate size with respect to the frame size, the face size increment from one pyramid level to the next, the spatial shift in searching for faces within each pyramid level, and the fact that no candidate face can be of smaller size than the face template. In this work, the face square side is restricted to lie within 10% and 75% of the frame width, with a face size increase of 15% across consecutive pyramid levels. Within each pyramid level, a local horizontal and vertical shift of one pixel is used to search for candidate faces.

Since the video signal is in *color* (see Section 4, below), *skin-tone* segmentation can be used to quickly narrow the search to face candidates that contain a relatively high proportion of skin-tone pixels. The normalized (red, green, blue) values of each frame pixel are first transformed to the (*hue*, *saturation*) color space, where skin tone is known to occupy a largely invariant to most humans and lighting conditions range of values (Graf, et al., 1997; Senior, 1999). In our case, all face candidates that contain less than 25% of pixels with hue and saturation values that fall within the skin-tone range, are eliminated. This substantially reduces the number of face candidates (depending on the frame background), speeding up computation and reducing spurious face detections.

Every remaining face candidate is subsequently size-normalized to the 11×11 face template size, and its *greyscale* pixel values are placed into an 121-dimensional face candidate vector. Each such vector is given a score based on both a two-class (face versus non-face) Fisher linear discriminant and the candidate's *"distance from face space"* (DFFS), i.e., the face vector projection error onto a lower, 40-dimensional space, obtained by means of PCA (see below). All candidate regions exceeding a threshold score are considered as faces. Among such faces at neighboring scales and locations, the one achieving the maximum score is returned by the algorithm as a detected face

(Senior, 1999).

Once a face has been detected, an ensemble of facial feature detectors are used to estimate the locations of 26 facial features, including the lip corners and centers (ten such facial features are marked on the frames of Figure 2). The search for these features occurs *hierarchically*. First, a few "high"-level features are located, and, subsequently, the 26 "low"-level features are located relative to the high-level feature locations. Each feature location is determined by using a score combination of prior feature location statistics, linear discriminant and "distance from feature space" (similar to the DFFS discussed above), based on the chosen feature template size (such as 11×11 pixels). FIG. 2 HERE

Before incorporating the described algorithm into our speechreading system, a training step is required to estimate the Fisher discriminant and eigenvectors (PCA) for face detection and facial feature estimation, as well as the facial feature location statistics. Such training requires a number of frames manually annotated with the faces and their visible features (see Section 4). When training the Fisher discriminant, both face and non-face (or facial feature and non-feature) vectors are used, whereas in the case of PCA, face and facial feature only vectors are considered (Senior, 1999).

Finally, it is worth mentioning that, when applying the algorithm on image sequences, the required computations can be substantially reduced by constraining the face candidates in the current frame (as well as the corresponding facial features) to be within a small scale and location variation of the previous-frame detected faces. Full image pyramid searches can be reduced to only one in 15 frames, for example. This approach however introduces temporal dependencies in the face detection errors. To improve robustness of the subsequent ROI extraction to such errors, we instead carry the full image pyramid search at each video frame.

3.2. Region of Interest Extraction

Given the output of the face detection and facial feature finding algorithm described above, five located lip-contour points are used to estimate the mouth center and its size at every video frame (four such points are marked on the frames of Figure 2). To improve ROI extraction robustness to face and mouth detection errors, the mouth center estimates are smoothed over twenty neighboring frames using *median* filtering to obtain the ROI center $(m_t\,, n_t)$, whereas the mouth size estimates are averaged over each utterance. A size-normalized ROI is then extracted as in (1), with $M = N = 64$, in order to allow for fast DCT and DWT implementation (see also Figure 2). The ROI greyscale only pixel values are placed in $\mathbf{x}_t^{(\mathrm{I})}$, as we have found no visual speech classification benefit by including color information in (1). Furthermore, in our current implementation, no rotation normalization, general three-dimensional pose compensation, or lighting normalization is directly applied on the ROI.

3.3. Cascade Algorithm Implementation

Stage I (image transform) of the visual feature extraction algorithm is applied on each ROI vector $\mathbf{x}_t^{(\mathrm{I})}$ of length $d^{(\mathrm{I})} = 4096$ at the video rate of 60 Hz. To simplify subsequent LDA and MLLT training, as well as bimodal (audio-visual) fusion, we interpolate the resulting features $\mathbf{y}_t^{(\mathrm{I})}$ to the audio feature rate, 100 Hz. Furthermore, and in order to account for lighting and other variations, we apply *feature mean normalization* (FMN) by simply subtracting the feature mean computed over the entire utterance length T, i.e., $\mathbf{y}_t^{(\mathrm{I})} \leftarrow \mathbf{y}_t^{(\mathrm{I})} - \Sigma_{t'=1}^{T} \mathbf{y}_{t'}^{(\mathrm{I})}/T$. This is akin to the audio front end processing (Rabiner and Juang, 1993), and it is known to help visual speech recognition (Potamianos, et al., 1998; Vanegas, et al., 1998); see also Section 4, below. When using Stage I as the sole visual front end, and in order to capture visual speech dynamics, we augment $\mathbf{y}_t^{(\mathrm{I})}$ by its first- and second-order derivatives, each computed over a 9-frame window, similarly to a widely used audio front end (Rabiner and Juang, 1993). In such case, we consider $D^{(\mathrm{I})} = 54 = 3 \times 18$.

At Stage II (LDA) and Stage III (MLLT) in the current visual front end implementation we use values $D^{(\mathrm{I})} = 24$, $D^{(\mathrm{II})} = D^{(\mathrm{III})} = 41$, and $J = 15$. In order to train the LDA projection matrix $\mathbf{P}^{(\mathrm{II})}$ and the MLLT rotation matrix $\mathbf{P}^{(\mathrm{III})}$, we consider $|\mathcal{C}| \approx 3400$ *context-dependent* sub-phonetic classes that coincide with the context-dependent states of an available audio-only *hidden Markov Model* (HMM), developed, in-house, for LVCSR. Such an HMM has been trained on a collection of audio corpora as described in (Polymenakos, et al., 1998), using the traditional speech recognition maximum likelihood estimation approach (Rabiner and Juang, 1993). Its class-conditional observation probability contains a total of approximately 90000 Gaussian mixtures. We use this HMM to label vectors $\mathbf{x}_t^{(\mathrm{II})}$, $\mathbf{x}_t^{(\mathrm{III})}$, as $\mathsf{c}_t \in \mathcal{C}$, by means of *Viterbi forced segmentation* (Rabiner and Juang, 1993), based on the audio channel of our audio-visual data. In addition to estimating matrices $\mathbf{P}^{(\mathrm{II})}$ and $\mathbf{P}^{(\mathrm{III})}$ (see Sections 2.2 and 2.3), such labels are used for training the phonetic classifiers described in Section 3.4, as well as for providing the ground truth, when testing them.

3.4. Phonetic and Visemic Classification

In order to test the effectiveness of the three stages of the proposed algorithm to the recognition of visual speech, we have decided to mostly report phonetic classification experiments, as opposed to large-vocabulary, continuous speech recognition results. Relative performance of visual feature extraction algorithms in the latter case is often masked by the language model effects (Rabiner and Juang, 1993). Furthermore, visual-only ASR performance is low, even for small-vocabulary tasks: For example, Potamianos and Graf (1998a) report a 36.5% visual-only word accuracy on a multi-subject connected-letter task (26-word problem). One clearly expects visual-only ASR performance to degrade when, for example, a 60000-word vocabulary is considered.

In this work, we consider 52 phoneme classes, and, for visual-only classification, also 27 *viseme* classes, both listed in Table 1. The training set utterance alignments are used to bootstrap visual-only *Gaussian mixture models* (GMMs), using the *expectation-maximization* (EM) algorithm

TABLE 1 HERE

(Dempster, et al., 1977). The GMM class-conditional probability is

$$Pr(\mathbf{y}_t|\mathsf{c}) = \sum_{m=1}^{M_\mathsf{c}} w_{\mathsf{c}m}\, \mathcal{N}_D(\mathbf{y}_t;\, \mathbf{m}_{\mathsf{c}m},\, \mathbf{\Sigma}_{\mathsf{c}m}), \quad \text{for all } \mathsf{c} \in \mathcal{C}. \tag{7}$$

In (7), *mixture weights* $w_{\mathsf{c}m}$ are positive adding up to one, M_c denotes the number of class mixtures, and $\mathcal{N}_D(\mathbf{y};\mathbf{m},\mathbf{\Sigma})$ denotes the D-variate normal distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$, assumed to be diagonal. In this work, we mostly consider $M_\mathsf{c} = 5$, or 64.

Frame-level *classification accuracy* is calculated by comparing, at each instance of t, the audio forced alignment class label c_t, obtained as described in Section 3.3, to its *maximum-a-posteriori* (MAP) class estimate $\hat{\mathsf{c}}_t$, obtained as (see also (7))

$$\hat{\mathsf{c}}_t = \arg\max_{\mathsf{c}\in\mathcal{C}} \{Pr(\mathbf{y}_t|\mathsf{c})\, Pr(\mathsf{c})\}. \tag{8}$$

In (8), the *smoothed* class *prior* $Pr(\mathsf{c}) = (L_\mathsf{c} + 1)/(L + |\mathcal{C}|)$, $\mathsf{c} \in \mathcal{C}$, is used (see also Section 2.2).

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 (7). Such HMMs are trained using the EM algorithm. MAP estimation becomes Viterbi decoding over each utterance phone segment (Rabiner and Juang, 1993).

It is of course also of interest to consider audio-visual phonetic classification, as a means of judging the possible effects of improved visual front end processing to audio-visual automatic speech recognition. As mentioned in the introduction, the problem of audio-visual sensory fusion constitutes a very active research area (Hennecke, et al., 1996). A number of traditional classifier fusion techniques can be used in our phonetic classification scenario (Jain, et al., 2000). In this work, we consider a simple but effective decision fusion method, assuming the following class-conditional audio-visual observation scoring function,

$$\text{Score}(\mathbf{y}_t^{(\text{AV})}|\,\mathsf{c}) = Pr(\mathbf{y}_t^{(\text{A})}|\,\mathsf{c})^{\gamma_\text{A}}\, Pr(\mathbf{y}_t^{(\text{V})}|\,\mathsf{c})^{\gamma_\text{V}}, \tag{9}$$

where $\gamma_\text{A}, \gamma_\text{V} \geq 0$, $\gamma_\text{A}+\gamma_\text{V} = 1$, and $\mathbf{y}_t^{(\text{AV})} = [\mathbf{y}_t^{(\text{A})\top}, \mathbf{y}_t^{(\text{V})\top}]^\top$ denotes the concatenation of time-synchronous audio[2] and visual features. Notice that, in general, (9) does not represend a probability density function. Nevertheless, (8) can still be used to estimate the most likely class $\hat{\mathsf{c}}_t \in \mathcal{C}$, with $Pr(\mathbf{y}_t|\mathsf{c})$ being replaced by $\text{Score}(\mathbf{y}_t^{(\text{AV})}|\,\mathsf{c})$. In (9), exponents $\gamma_\text{A}, \gamma_\text{V}$ are used to capture the relative reliability ("confidence") of the audio and visual feature streams, as information sources about the spoken utterance. As it is demonstrated in Section 4, their values greatly influence the performance of the joint audio-visual system. Optimal exponent values can be obtained by various methods (Adjoudani and Benoît, 1996; Rogozan, et al., 1997; Potamianos and Graf, 1998b; Neti, et al., 2000); here, they are estimated by simply maximizing the bimodal phonetic classification accuracy on a held-out data set (see Section 4). Notice that $\gamma_\text{A}, \gamma_\text{V}$ are assumed constant over the

[2]Throughout this work, the audio front end reported in (Basu, et al., 1999) is used to obtain $\mathbf{y}_t^{(\text{A})}$.

entire set of utterances. Generalizations, where the exponents are estimated locally on a per-frame or per-utterance basis, have been considered in (Neti, et al., 2000; Potamianos and Neti, 2000).

4. Database and Experiments

We have collected a 290-subject, large-vocabulary, continuous speech audio-visual database, using IBM ViaVoice^{TM} training utterance scripts (Neti, et al., 2000). The database contains full frontal face color video of the subjects with minor face-camera distance and lighting variations (see also Figure 2). The video is captured at a resolution of 704×480 pixels (interlaced), a frame rate of 30 Hz (i.e., 60 fields per second are available at a resolution of 240 lines), and it is MPEG2 encoded to about 0.5 MBytes/sec. The audio is captured at a relatively "clean" office environment, at a sampling rate of 16 KHz, and it is time-synchronous to the video stream. For faster experimentation, a subset of this database, consisting of 162 subjects and close to 8 hours of speech (3,888 utterances), has been exclusively used for experiments in this paper. For each of the 162 subjects, we have randomly selected 24 database videos and randomly split them into 16 training, 4 test, and 4 held-out utterances, thus creating a *multi-subject* 2,592 utterance *training set* (5.5 hours) and two 648 utterance sets, namely a *test set* and a *held-out set* of about 1.3 hours each. The latter is used for optimizing exponents γ_{A} and γ_{V} in (9).

We first process the video data to extract the mouth ROI, as discussed in Sections 3.1 and 3.2. The statistical face detection and feature localization templates (Fisher discriminant and eigenvectors) are trained using 10 video frames for each of the 162 database subjects, each manually annotated with the 26 facial feature locations. The performance of the trained face detector is subsequently evaluated on a test set containing 3 marked frames per subject. Following some fine tuning of the image pyramid parameters and of the minimum percentage of skin-tone pixels within a face candidate, face detection becomes 100% accurate on this test set, and each mouth feature is estimated "close" to its true location (within a radius of 0.10 times the eye separation) in more than 90% of the test video frames. Subsequently, all 2106 annotated frames are pulled together to train new statistical templates. The face detection performance tested on all 3,888 database videos (containing approximately 0.9 million frames) is 99.5% correct, assuming that one face is present per video frame. Given the face detection and mouth feature localization results, ROI extraction and visual feature computation follow, as explained in Sections 3.2 and 3.3.

We next compare the phonetic classification performance of the various algorithm stages discussed in Section 2, using, at first, $M_{\mathrm{c}} = 5$ in (7). As shown in Table 2, and regardless of the visual feature extraction method employed at Stage I (DCT, DWT, or PCA), the use of LDA (Stage II) results in significant accuracy improvement (20% relative, in the DCT, GMM based classification case, for example). Using the additional MLLT data rotation (Stage III) further improves performance (10% in the DWT, HMM classification case). Both stages combined can account for up to 27% accuracy relative improvement over the image transform only (Stage I) visual front end

TABLE 2 HERE

(DCT, GMM based classification case, for example). Notice that, within each column of Table 2, all performance differences are *statistically significant*, as computed using *McNemar's test* (Gillick and Cox, 1989) for independent algorithm errors, a valid assumption in the case of phonetic classification. Indeed, for all same-column comparisons in Table 2, the probability that the observed difference between any two algorithm stages would arise by chance is computed to be less than 10^{-6}. Interestingly, GMM based phonetic classification benefits, in general, relatively more than HMM based classification, by the added stages of the algorithm. This is possibly due to the fact that the latter uses the entire phone segment (containing a number of feature frames), to obtain estimate (8), thus being more robust than single-frame (GMM) based classification.

Overall, the performance within each algorithm stage does not vary much 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 the PCA (34.64%, 33.67%, and 32.65% GMM based, 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 (see (3), in addition to the required diagonalization of correlation matrix $\mathbf{R}$). Clearly therefore, DCT and DWT are preferable to the use of PCA. Notice that, within each row of Table 2 (and for the same type of classification method), all performance differences are statistical significant, with the exception of the HMM based, Stage I classification accuracy difference between DWT and PCA (a by chance occurrence of such a difference is computed to be 0.10).

It is worth reporting that feature mean normalization (FMN) improves classification performance. Indeed, GMM, DCT feature based Stage I classification accuracy without FMN drops to 25.99%, as compared to 27.31% when FMN is applied (see also Table 2). 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.

Classification using various size GMM and HMM systems is addressed in Figure 3(a). Clearly, larger systems (with larger values of M_{c}) perform better, but the relative performance of the three algorithm stages remains mostly unchanged. Figure 3(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. For clarity, we also depict GMM classification using a uniform prior in (8). Wider temporal windows improve performance, however at an increased computational cost. Such increase occurs both, when computing matrices $\mathbf{S}_W$ and $\mathbf{S}_B$ in (6) (particularly, when computing the sample covariance matrices $\mathbf{\Sigma}^{(\mathsf{c})}$, a task of $O(J^2)$ computational complexity), as well as when solving the generalized eigenvalue and eigenvector problem to obtain the LDA projection matrix $\mathbf{P}^{(\mathrm{II})}$ (see Section 2.2), a task of $O(J^3)$ complexity (Golub and Van Loan, 1983).

FIG. 3 HERE

In Table 3, we concentrate on the DCT based visual front end, and we first report improved (compared to Table 2) visual-only classification accuracy, using a classifier with 64 mixtures per GMM class, or HMM state. Such a system achieves a 48.85% segmental (HMM) based visual-only

phonetic classification accuracy at Stage III. This corresponds to a 59.77% visemic classification accuracy (see also Table 1). For completeness, audio-only as well as audio-visual phonetic classification accuracies obtained by means of (9), are also reported. Notice that both Stages II and III improve audio-visual phonetic classification over Stage I. Indeed, the reported HMM based 80.52% clean audio-only accuracy improves in a statistically significant manner to 83.20%, 83.81%, and 84.14%, when Stage I, II, and III visual features are respectively used to augment the audio modality by means of (9). The audio front end remains unchanged in all three cases. Notice that the best improvement corresponds to a 18% reduction of the classification error in the clean speech case, when the Stage III, HMM based, audio-visual system is used in place of its audio-only counterpart. This is mostly due to the visual modality resolving phoneme confusions across visemes. For example, confusions between /T/ and /P/, which belong to different visemes (see Table 1), drop from 249 frames, when the audio-only classifier is used, to only 59 frames, when the audio-visual HMM is employed (a 76% reduction). Similarly, /N/ and /M/ confusions drop from 1998 to 879 frames (a 66% reduction). However, the visual modality does not benefit discrimination among phonemes that cluster into the same viseme. For example, /P/ and /B/ confusions actually slightly increase from 505 to 566, when the visual modality is added.

TABLE 3 HERE

Next, in Figure 4, we further consider the use of (9) in audio-visual phonetic classification. Figure 4(a) demonstrates the dependence of the bimodal GMM classifier accuracy on the choice of exponent $\gamma_{\rm A} \in [\,0\,,1]$ (recall that $\gamma_{\rm V} = 1 - \gamma_{\rm A}$). A near-optimal audio stream exponent value can be simply estimated by considering a fine grid of $\gamma_{\rm A} \in [\,0\,,1]$, subsequently computing the corresponding audio-visual phonetic classification accuracies on the held-out data set, and retaining the exponent value associated with the best performance. This is easily accomplished in the case of GMM frame-level classification, whereas, for HMM segmental classification, more elaborate schemes such as discriminative training could be used instead (Potamianos and Graf, 1998b). Figure 4(a) shows the bimodal phonetic classification accuracy on both the test and held-out sets, using 201 equally spaced exponents $\gamma_{\rm A} \in [\,0\,,1]$. Notice that the best audio stream exponent values are almost identical for the two sets ($\gamma_{\rm A} = 0.605$ and 0.610 for the test and held-out sets, respectively).

In Figure 4(b), we concentrate on audio-visual phonetic classification in the case of noisy speech. The audio-only channel is artificially corrupted by additive, non-stationary, speech ("babble") noise, at a number of *signal-to-noise ratios* (SNRs). At every SNR considered, an audio-only GMM phonetic classifier (with $M_{\rm c} = 64$) is first trained on the *matched*-condition training set, and its accuracy is compared to that of its corresponding audio-visual GMM classifier (9), with optimal exponent values estimated on the basis of the matched-condition held-out data set. Notice that the audio-visual classifier exhibits superior robustness to noise. For example, the audio-visual phonetic classification accuracy at 2 dB SNR is almost identical to the audio-only accuracy at 10 dB, thus amounting to an "effective SNR gain" of 8 dB.

FIG. 4 HERE

Finally, we briefly report the *word error rate* (WER) for some large-vocabulary, continuous speech recognition (LVCSR) preliminary experiments on this database. We consider the HMM

based LVCSR system reported in (Polymenakos, et al., 1998) with a 60000-word vocabulary and a tri-gram language model. After 3 iterations of the EM algorithm, and starting with an initial segmentation based on the original audio-only HMM and the audio-only front end (see Section 3.3), we obtain an audio-only WER of 13.94%, a visual-only WER of 87.60%, and an audio-visual WER of 13.78%. The latter is obtained by training HMMs with state-conditional probability densities (7), where $\mathbf{y}_t \leftarrow [\mathbf{y}_t^{(A)\top}, \mathbf{y}_t^{(V)\top}]^\top$ (concatenative feature fusion). Such an audio-visual integration approach is known to often degrade ASR performance for both small- and large-vocabulary tasks in the clean audio case (Potamianos and Graf, 1998b; Neti, et al., 2000), therefore the above LVCSR results are not surprising. Nevertheless, significant ASR improvement can be achieved when the audio stream is degraded. For example, and for audio corrupted by "babble" noise at 8.5 dB, the *matched*-trained noisy audio WER improves from 41.57% to 31.30% by incorporating the visual information. Neti, et al. (2000) investigate audio-visual LVCSR decision fusion strategies by means of the *multi-stream* HMM (Dupont and Luettin, 2000), as well as various stream "confidence" estimation techniques. Significant LVCSR WER reduction is achieved by such methods, even in the clean audio case. Additional research work in this area is currently in progress.

5. Conclusions and Discussion

In this paper, we have described 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. In a visual-only classification of 52 phonemes, we have demonstrated that all three stages allow to improve 5-mixture GMM based classification (with no prior phonetic segmentation of the test data) from 27.31% accuracy (DCT based Stage I alone) to 34.64% (three-stage DCT based visual front end), corresponding to a 27% accuracy relative gain. In a visual-only classification of 27 visemes, a 64-mixture GMM classifier reaches 49.29% recognition when all three stages are applied, amounting to an 11% relative improvement over a single-stage DCT based front end. When using a 64-mixture HMM based system (with knowledge of the test set viseme boundaries), a 59.77% classification accuracy is achieved, amounting to a 4% relative improvement over the corresponding single-stage system. In addition, the proposed algorithm has resulted in improved audio-visual phonetic classification over the use of a single-stage image transform visual front end. Noisy audio-visual phonetic classification results and preliminary large-vocabulary, continuous speech recognition experiments have also been presented.

The experiments in this paper have been reported on a large audio-visual database, suitable for large-vocabulary, continuous ASR. This fact allows our conclusions on comparing visual front end algorithms to be statistically significant, as discussed in our experiments. In light of the rich phonetic context of our data, it is also very encouraging to record an 18% phone classification error reduction in the clean speech case, using a crude 64-mixture HMM based phonetic classifier and a

simple audio-visual decision fusion model with constant audio and visual stream exponents.

It is worth stressing that, given the mouth ROI, the proposed visual front end is computationally efficient: It consists of a fast image transform (DCT, or DWT), followed by a data projection (LDA) and a subsequent data rotation (MLLT) applied on vectors of low dimensionality. Such efficiency allows real-time automatic speechreading system implementation, assuming that it is adequate to perform face detection and facial feature localization at a lower frame rate. Practical automatic speechreading systems are therefore feasible.

Acknowledgments

The authors would like to acknowledge contributions to this work by Ramesh Gopinath of the IBM Thomas J. Watson Research Center, specifically for insights in the maximum likelihood linear transform algorithm, and to thank the anonymous reviewers for helpful suggestions and comments.

References

Adjoudani, A. and Benoît, C. (1996) . On the integration of auditory and visual parameters in an HMM-based ASR. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 461–471.

Basu, S., Neti, C., Rajput, N., Senior, A., Subramaniam, L., and Verma, A. (1999) . Audio-visual large vocabulary continuous speech recognition in the broadcast domain. *Proceedings IEEE Workshop on Multimedia Signal Processing (MMSP)'99*, Copenhagen, Denmark, pp. 475–481.

Bregler, C. and Konig, Y. (1994) . 'Eigenlips' for robust speech recognition. *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)'94*, Adelaide, Australia, pp. 669–672.

Brooke, N.M. (1996) . Talking heads and speech recognizers that can see: The computer processing of visual speech signals. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 351–371.

Chandramohan, D. and Silsbee, P.L. (1996) . A multiple deformable template approach for visual speech recognition. *Proceedings International Conference on Spoken Language Processing (ICSLP)'96*, Philadelphia, PA, pp. 50–53.

Chatfield, C. and Collins, A.J. (1980) . *Introduction to Multivariate Analysis*. London, United Kingdom: Chapman and Hall.

Chen, T. (2001) . Audiovisual speech processing. Lip reading and lip synchronization. *IEEE Signal Processing Magazine, 18*(1):9–21.

Chiou, G.I. and Hwang, J.-N. (1997) . Lipreading from color video. *IEEE Transactions on Image Processing, 6*(8):1192–1195.

Daubechies, I. (1992) . *Wavelets*. Philadelphia, PA: S.I.A.M.

De Cuetos, P., Neti, C., and Senior, A. (2000) . Audio-visual intent-to-speak detection for human-computer interaction. *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)'00*, Istanbul, Turkey, pp. 1325–1328.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) . Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, 39*(1):1–38.

Duchnowski, P., Hunke, M., Büsching, D., Meier, U., and Waibel, A. (1995) . Toward movement-invariant automatic lip-reading and speech recognition. *Proceedings International Conference*

on Acoustics, Speech, and Signal Processing (ICASSP)'95, Detroit, MI, pp. 109–112.

Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia, 2*(3):141–151.

Fröba, B., Küblbeck, C., Rothe, C., and Plankensteiner, P. (1999). Multi-sensor biometric person recognition in an access control system. *Proceedings International Conference on Audio and Video-based Biometric Person Authentication (AVBPA)'99*, Washington, DC, pp. 55–59.

Gillick, L. and Cox, S.J. (1989). Some statistical issues in the comparison of speech recognition algorithms. *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)'89*, Glasgow, United Kingdom, pp. 532–535.

Golub, G.H. and Van Loan, C.F. (1983). *Matrix Computations*. Baltimore, MD: The Johns Hopkins University Press.

Gopinath, R.A. (1998). Maximum likelihood modeling with Gaussian distributions for classification. *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)'98*, Seattle, WA, pp. 661–664.

Graf, H.P., Cosatto, E., and Potamianos, G. (1997). Robust recognition of faces and facial features with a multi-modal system. *Proceedings International Conference on Systems, Man, and Cybernetics (ICSMC)'97*, Orlando, FL, pp. 2034–2039.

Gray, M.S., Movellan, J.R., and Sejnowski, T.J. (1997). Dynamic features for visual speech-reading: A systematic comparison. In Mozer, M.C., Jordan, M.I., and Petsche, T. (Eds.), *Advances in Neural Information Processing Systems 9*. Cambridge, MA: MIT Press, pp. 751–757.

Hennecke, M.E., Stork, D.G., and Prasad, K.V. (1996). Visionary speech: Looking ahead to practical speechreading systems. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 331–349.

Jain, A.K., Duin, R.P.W., and Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(1):4–37.

Massaro, D.W. and Stork, D.G. (1998). Speech recognition and sensory integration. *American Scientist, 86*(3):236–244.

Matthews, I. (1998). *Features for Audio-Visual Speech Recognition*. PhD Thesis, School of Information Systems, University of East Anglia, Norwich, United Kingdom.

Matthews, I., Bangham, J.A., and Cox, S. (1996). Audio-visual speech recognition using multiscale nonlinear image decomposition. *Proceedings International Conference on Spoken Lan-*

guage Processing (ICSLP)'96, Philadelphia, PA, pp. 38–41.

McGurk, H. and MacDonald, J.W. (1976) . Hearing lips and seeing voices. *Nature, 264*:746–748.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J. (2000). *Audio-Visual Speech Recognition.* Summer Workshop 2000 Final Technical Report, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD (`http://www.clsp.jhu.edu/ws2000/final_reports/avsr/`).

Petajan, E.D. (1984) . Automatic lipreading to enhance speech recognition. *Proceedings Global Telecommunications Conference (GLOBCOM)'84*, Atlanta, GA, pp. 265–272.

Polymenakos, L., Olsen, P., Kanevsky, D., Gopinath, R.A., Gopalakrishnan, P.S., and Chen, S. (1998). Transcription of broadcast news - some recent improvements to IBM's LVCSR system. *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)'98*, Seattle, WA, pp. 901–904.

Potamianos, G. and Graf, H.P. (1998a) . Linear discriminant analysis for speechreading. *Proceedings IEEE Workshop on Multimedia Signal Processing (MMSP)'98*, Los Angeles, CA, pp. 221–226.

Potamianos, G. and Graf, H.P. (1998b) . Discriminative training of HMM stream exponents for audio-visual speech recognition. *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)'98*, Seattle, WA, pp. 3733–3736.

Potamianos, G., Graf, H.P., and Cosatto, E. (1998) . An image transform approach for HMM based automatic lipreading. *Proceedings International Conference on Image Processing (ICIP)'98*, Chicago, IL, vol. III, pp. 173–177.

Potamianos, G., Luettin, J., and Neti, C. (2001) . Hierarchical discriminant features for audio-visual LVCSR. *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)'01*, Salt Lake City, UT (In Press).

Potamianos, G. and Neti, C. (2000) . Stream confidence estimation for audio-visual speech recognition. *Proceedings International Conference on Spoken Language Processing (ICSLP)'00*, Beijing, China, vol. III, pp. 746–749.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1988) . *Numerical Recipes in C. The Art of Scientific Computing.* Cambridge, MA: Cambridge University Press.

Rabiner, L. and Juang, B.-H. (1993) . *Fundamentals of Speech Recognition.* Englewood Cliffs, NJ: Prentice Hall.

Rao, C.R. (1965) . *Linear Statistical Inference and Its Applications.* New York, NY: John Wiley and Sons.

Rogozan, A., Deléglise, P., and Alissali, M. (1997) . Adaptive determination of audio and visual weights for automatic speech recognition. *Proceedings European Tutorial Research Workshop on Audio-Visual Speech Processing (AVSP)'97*, Rhodes, Greece, pp. 61–64.

Senior, A.W. (1999) . Face and feature finding for a face recognition system. *Proceedings International Conference on Audio and Video-based Biometric Person Authentication (AVBPA)'99*, Washington, DC, pp. 154–159.

Stork, D.G. and Hennecke, M.E. (Eds.) (1996) . *Speechreading by Humans and Machines.* Berlin, Germany: Springer.

Summerfield, A.Q. (1987) . Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd, B. and Campbell, R. (Eds.), *Hearing by Eye: The Psychology of Lip-Reading.* Hillside, NJ: Lawrence Erlbaum Associates, pp. 97–113.

Summerfield, Q., MacLeod, A., McGrath, M., and Brooke, M. (1989) . Lips, teeth, and the benefits of lipreading. In Young, A.W. and Ellis, H.D. (Eds.), *Handbook of Research on Face Processing.* Amsterdam, The Netherlands: Elsevier Science Publishers, pp. 223–233.

Teissier, P., Robert-Ribes, J., Schwartz, J.-L., and Guérin-Dugué, A. (1999) . Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing, 7*(6):629–642.

Vanegas, O., Tanaka, A., Tokuda, K., and Kitamura, T. (1998) . HMM-based visual speech recognition using intensity and location normalization. *Proceedings International Conference on Spoken Language Processing (ICSLP)'98*, Sydney, Australia, pp. 289–292.

Wark, T. and Sridharan, S. (1998) . A syntactic approach to automatic lip feature extraction for speaker identification. *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)'98*, Seattle, WA, pp. 3693–3696.

Table 1. Phonetic clustering used as the set of 27 visemes in our experiments. Phones /SIL/ and /SP/ correspond to silence and short pause, respectively.

{/AA/,/AH/,/AX/}	{/AE/}	{/AO/}	{/AW/}	{/AXR/,/ER/}	{/AY/}
{/CH/}	{/EH/}	{/EY/}	{/HH/}	{/IH/,/IX/}	{/IY/}
{/JH/}	{/L/}	{/OW/}	{/OY/}	{/R/}	{/UH/,/UW/}
{/W/}	{/SIL/,/SP/}	{/TS/}	{/F/,/V/}	{/S/,/Z/}	{/SH/,/ZH/}
{/DH/,/TH/}	{/D/,/DD/,/DX/,/G/,/GD/,/K/,/KD/,/N/,/NG/,/T/,/TD/,/Y/}				{/B/,/BD/,/M/,/P/,/PD/}

Table 2. 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 accuracies are reported (5 mixtures per GMM class or HMM state are used).

$\mathbf{P}^{(\mathrm{I})} \rightarrow$	DCT		DWT		PCA	
STAGE	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 3. Test set visual-only visemic (VI-27) and phonetic (VI-52) classification accuracies (%) using each stage of the visual front end and DCT features. Audio-only (AU) and audio-visual (AV) phonetic classification accuracies are also depicted ($\gamma_A = 0.65$, $\gamma_V = 0.35$ are used in (9)). Both GMM and segmental based HMM classification accuracies are reported (64 mixtures per GMM class or HMM state are used).

TASK →	VI-27		VI-52		AU		AV	
STAGE	GMM	HMM	GMM	HMM	GMM	HMM	GMM	HMM
I	44.47	57.64	31.77	46.07	62.78	80.52	64.40	83.20
II	47.66	58.56	35.74	46.52	62.78	80.52	66.10	83.81
III	49.29	59.77	37.71	48.85	62.78	80.52	66.36	84.14

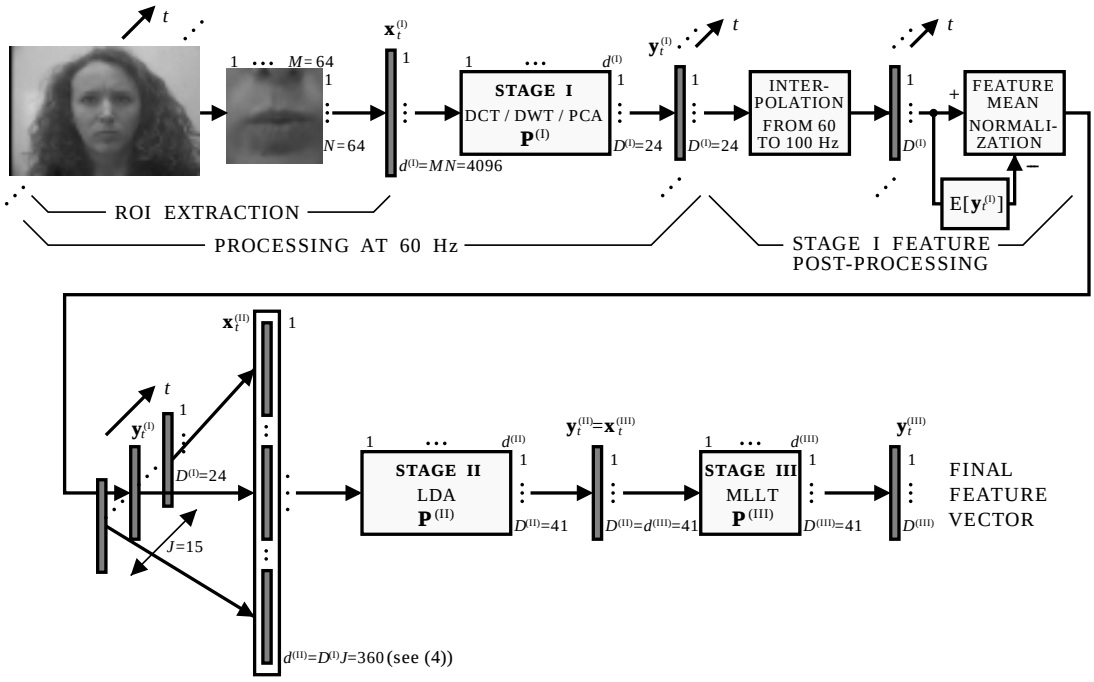

Figure 1. The proposed cascade algorithm block diagram.

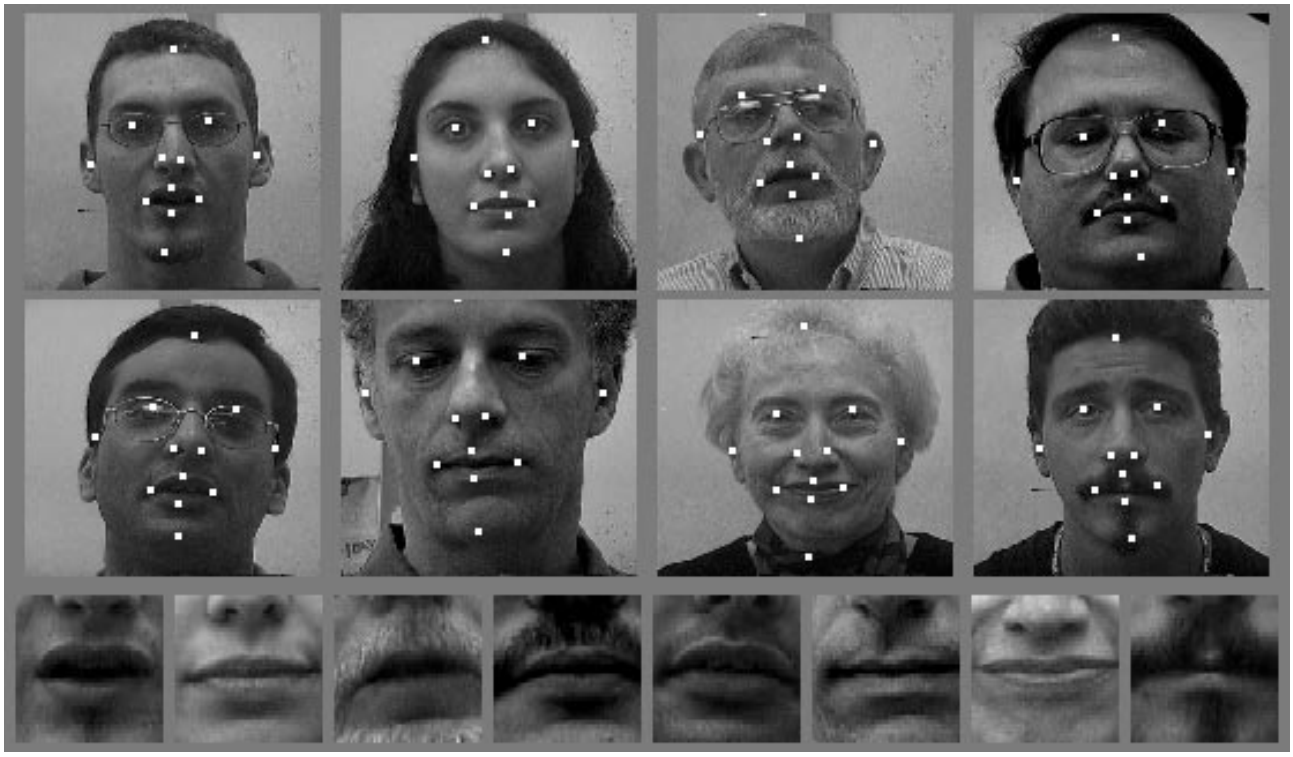

Figure 2. ROI extraction examples. *Upper rows*: Example video frames from 8 database subjects, with detected facial features superimposed. *Lower row*: Corresponding extracted mouth regions-of-interest, size-normalized.

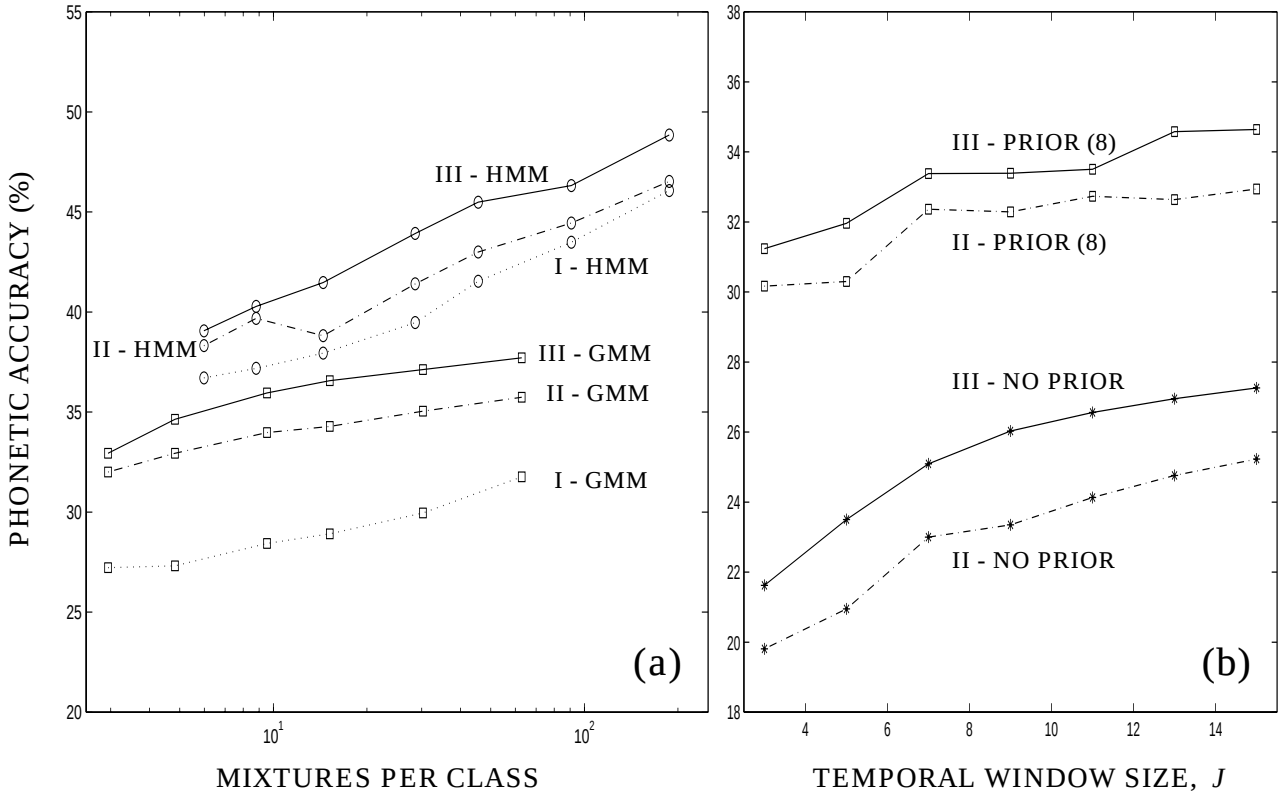

Figure 3. Visual-only phonetic classification accuracy using a DCT based visual front end, as a function of: (a) number of mixtures per GMM (M_c) or HMM phone class ($3\,M_c$); (b) temporal window size J at Stage II (GMM only, with or without prior in (8)).

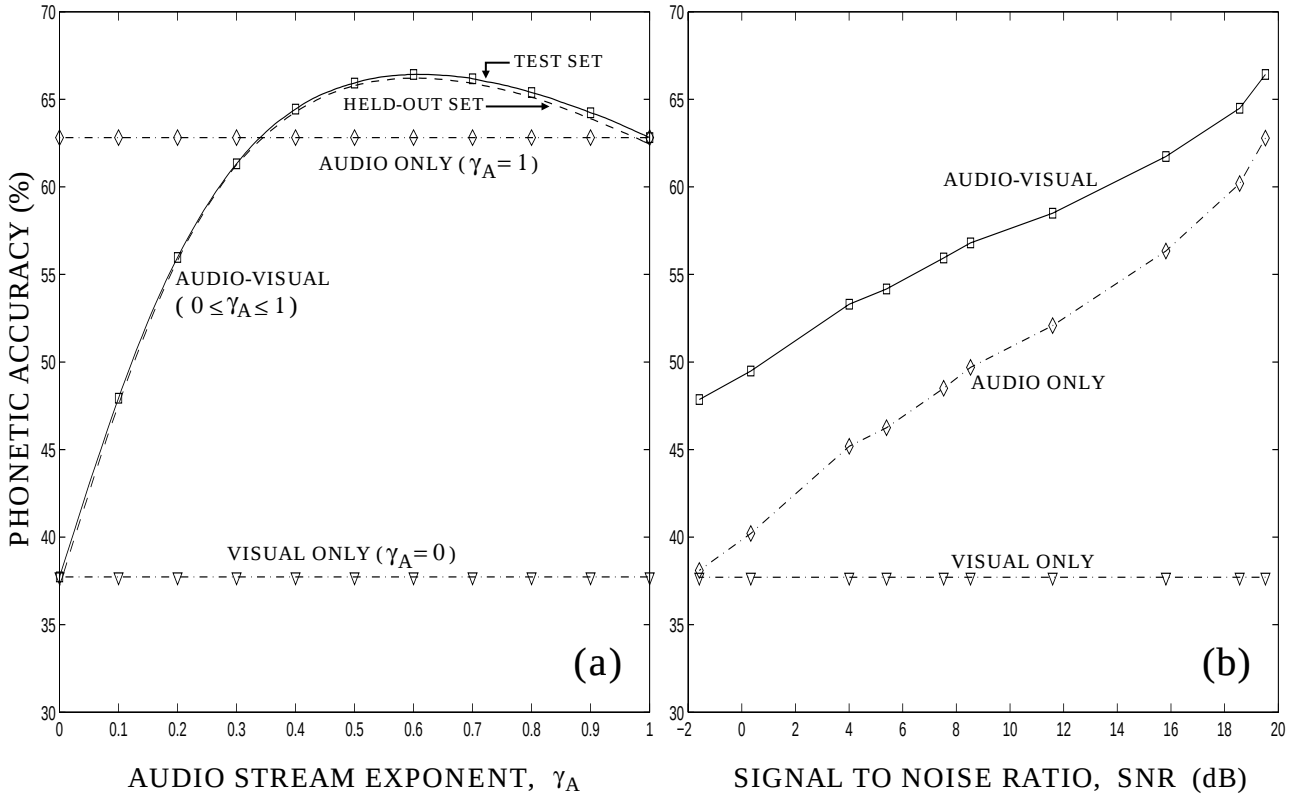

Figure 4. GMM based audio-visual phonetic classification accuracy using (9) and Stage III, DCT based visual features, as a function of: (a) audio exponent value γ_A, for clean audio (both test and held-out set performances are shown); (b) signal-to-noise ratio (SNR) for degraded audio (optimal exponents are estimated based on the held-out set). In both cases, test set audio- and visual-only phonetic classification accuracies are also depicted.