

CHAPTER 10

Audio-Visual Automatic Speech Recognition: An Overview

Gerasimos Potamianos, Chalapathy Neti

Human Language Technologies Department, IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: {gpotam, cneti}@us.ibm.com).

Juergen Luettn

Robert Bosch GmbH, Automotive Electronics, D-7152 Leonberg, Germany (e-mail: Juergen.Luettn@de.bosch.com).

Iain Matthews

Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA (e-mail: iainm@cs.cmu.edu).

INTRODUCTION

We have made significant progress in *automatic speech recognition* (ASR) for well-defined applications like dictation and medium vocabulary transaction processing tasks in relatively controlled environments. However, ASR performance has yet to reach the level required for speech to become a truly *pervasive user interface*. Indeed, even in “clean” acoustic environments, and for a variety of tasks, state of the art ASR system performance lags human speech perception by up to an order of magnitude (Lippmann, 1997). In addition, current systems are quite sensitive to channel, environment, and style of speech variations. A number of techniques for improving ASR *robustness* have met limited success in severely degraded environments, mismatched to system training (Ghitza, 1986; Nadas et al., 1989; Juang, 1991; Liu et al., 1993; Hermansky and Morgan, 1994; Neti, 1994; Gales, 1997; Jiang et al., 2001). Clearly, novel, non-traditional approaches, that use orthogonal sources of information to the acoustic input, are needed to achieve ASR performance closer to the human speech perception level, and robust enough to be deployable in field applications. *Visual speech* is the most promising source of additional speech information, and it is obviously not affected by the acoustic environment and noise.

Human speech perception is *bimodal* in nature: Humans combine audio and visual information in deciding what has been spoken, especially in noisy environments. The visual modality benefit to speech intelligibility in noise has been quantified as far back as in Sumbly and Pollack (1954). Furthermore, bimodal fusion of audio and visual stimuli in perceiving speech has been demonstrated by the *McGurk effect* (McGurk and MacDonald, 1976). For example, when the spoken sound /ga/ is superimposed on the video of a person uttering /ba/, most people perceive the speaker as uttering the sound /da/. In addition, visual speech is of particular importance to the *hearing impaired*: Mouth movement is known to play an important role in both sign language and simultaneous communication between the deaf (Marschark et al., 1998). The hearing impaired speechread well, and possibly better than the general population (Bernstein et al., 1998).

There are three key reasons why vision benefits human speech perception (Summerfield, 1987): It helps speaker (audio source) localization, it contains speech segmental information that supplements the audio, and it provides complimentary information about the *place of articulation*. The latter is due to the partial or full

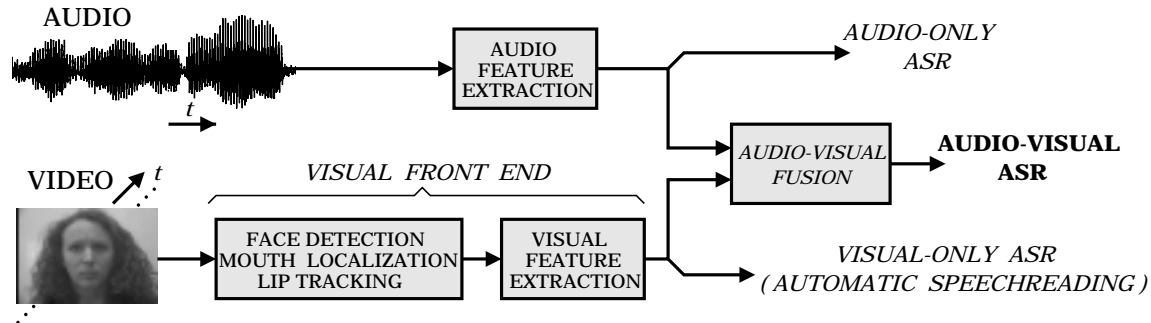


Figure 1: The main processing blocks of an audio-visual automatic speech recognizer. The visual front end design and the audio-visual fusion modules introduce additional challenging tasks to automatic recognition of speech, as compared to traditional audio-only ASR. They are discussed in detail in this chapter.

visibility of articulators, such as the tongue, teeth, and lips. Place of articulation information can help disambiguate, for example, the unvoiced consonants /p/ (a bilabial) and /k/ (a velar), the voiced consonant pair /b/ and /d/ (a bilabial and alveolar, respectively), and the nasal /m/ (a bilabial) from the nasal alveolar /n/ (Masaro and Stork, 1998). All three pairs are highly confusable on basis of acoustics alone. In addition, jaw and lower face muscle movement is correlated to the produced acoustics (Yehia et al., 1998; Barker and Berthommier, 1999), and its visibility has been demonstrated to enhance human speech perception (Summerfield et al., 1989; Smeele, 1996).

The above facts have motivated significant interest in automatic recognition of visual speech, formally known as *automatic lipreading*, or *speechreading* (Stork and Hennecke, 1996). Work in this field aims at improving ASR by exploiting the visual modality of the speaker's mouth region in addition to the traditional audio modality, leading to *audio-visual automatic speech recognition* systems. Not surprisingly, including the visual modality has been shown to outperform audio-only ASR over a wide range of conditions. Such performance gains are particularly impressive in noisy environments, where traditional acoustic-only ASR performs poorly. Improvements have also been demonstrated when speech is degraded due to speech impairment (Potamianos and Neti, 2001a) and *Lombard* effects (Huang and Chen, 2001). Coupled with the diminishing cost of quality video capturing systems, these facts make automatic speechreading tractable for achieving robust ASR in certain scenarios (Hennecke et al., 1996).

Automatic recognition of audio-visual speech introduces new and challenging tasks compared to traditional, audio-only ASR. The block-diagram of Figure 1 highlights these: In addition to the usual audio front end (feature extraction stage), visual features that are informative about speech must be extracted from video of the speaker's face. This requires robust face detection, as well as location estimation and tracking of the speaker's mouth or lips, followed by extraction of suitable visual features. In contrast to audio-only recognizers, there are now *two* streams of features available for recognition, one for each modality. The combination of the audio and visual streams should ensure that the resulting system performance is better than the best of the two single modality recognizers, and hopefully, significantly outperform it. Both issues, namely the *visual front end design* and *audio-visual fusion*, constitute difficult problems, and they have generated significant research work by the scientific community.

Indeed, since the mid-eighties, over a hundred articles have concentrated on audio-visual ASR, with the vast majority appearing during the last decade. The first automatic speechreading system was reported by Petajan (1984). Given the video of the speaker's face, and by using simple image thresholding, he was able to extract binary (black and white) mouth images, and subsequently, mouth height, width, perimeter, and area, as visual speech features. He then developed a visual-only recognizer based on dynamic time warping (Rabiner and Juang, 1993) to rescore the best two choices of the output of the baseline audio-only system. His method improved ASR for a single-speaker, isolated word recognition task on a 100-word vocabulary that included digits and letters. Petajan's work generated significant excitement, and soon various sites established research in audio-visual ASR. Among the pioneer sites was the group headed by Christian Benoît at the Institute

de la Communication Parlée (ICP), in Grenoble, France. For example, Adjoudani and Benoît (1996) have investigated the problem of audio-visual fusion for ASR, and compared early vs. late integration strategies. In the latter case, they considered modality reliability estimation based on the dispersion of the likelihoods of the top four recognized words using the audio-only and visual-only inputs. They reported significant ASR gains on a single-speaker corpus of 54 French non-sense words. Later, they developed a multimedia platform for audio-visual speech processing, containing a head mounted camera to robustly capture the speaker's mouth region (Adjoudani et al., 1997). Recently, work at ICP has continued in this arena, with additional audio-visual corpora collected (French connected letters and English connected digits) and a new audio-visual ASR system reported by Heckmann et al. (2001).

As shown in Figure 1, audio-visual ASR systems differ in three main aspects (Hennecke et al., 1996): The visual front end design, the audio-visual integration strategy, and the speech recognition method used. Unfortunately, the diverse algorithms suggested in the literature for automatic speechreading are very difficult to compare, as they are rarely tested on a common audio-visual database. In addition, until very recently (Neti et al., 2000), audio-visual ASR studies have been conducted on databases of small duration, and, in most cases, limited to a very small number of speakers (mostly less than ten, and often single-subject) and to small vocabulary tasks (Hennecke et al., 1996; Chibelushi et al., 1996, 2002). Such tasks are typically non-sense words (Adjoudani and Benoît, 1996; Su and Silsbee, 1996), isolated words (Petajan, 1984; Movellan and Chadderdon, 1996; Matthews et al., 1996; Chan et al., 1998; Dupont and Luetin, 2000; Gurbuz et al., 2001; Huang and Chen, 2001; Nefian et al., 2002), connected letters (Potamianos et al., 1998), connected digits (Potamianos et al., 1998; Zhang et al., 2000), closed-set sentences (Goldschen et al., 1996), or small-vocabulary continuous speech (Chu and Huang, 2000). Databases are commonly recorded in English, but other examples are French (Adjoudani and Benoît, 1996; Alissali et al., 1996; André-Obrecht et al., 1997; Teissier et al., 1999; Dupont and Luetin, 2000), German (Bregler et al., 1993; Krone et al., 1997), Japanese (Nakamura et al., 2000), and Hungarian (Czap, 2000). However, if the visual modality is to become a viable component in real-word ASR systems, research work is required on larger vocabulary tasks, developing speechreading systems on data of sizable duration and of large subject populations. A first attempt towards this goal was the authors' work during the Johns Hopkins summer 2000 workshop (Neti et al., 2000), where a *speaker-independent* audio-visual ASR system for *large vocabulary continuous speech recognition* (LVCSR) was developed for the first time. Significant performance gains in both clean and noisy audio conditions were reported.

In this chapter, we present the main techniques for audio-visual speech recognition that have been developed over the past two decades. We first discuss the visual feature extraction problem, followed by audio-visual fusion. In both cases, we provide details of some of the techniques employed during the Johns Hopkins summer 2000 workshop (Neti et al., 2000). We also consider the problem of audio-visual speaker adaptation, an issue of significant importance when building speaker specific models, or developing systems across databases. We then discuss the main audio-visual corpora, used in the literature for ASR experiments, including the IBM audio-visual LVCSR database. Subsequently, we present experimental results on automatic speechreading and audio-visual ASR. As an application of speaker adaptation, we consider the problem of automatic recognition of impaired speech. Finally, we conclude the chapter with a discussion on the current state of audio-visual ASR, and on what we view as open problems in this area.

VISUAL FRONT ENDS FOR AUTOMATIC SPEECHREADING

As it was briefly mentioned in the Introduction (see also Figure 1), the first main difficulty in the area of audio-visual ASR is the visual front end design. The problem is two-fold: Face, lips, or mouth tracking is first required, followed by visual speech representation in terms of a small number of informative features. Clearly, the two issues are closely related: Employing a lip tracking algorithm allows one to use visual features such as mouth height or width (Adjoudani and Benoît, 1996; Chan et al., 1998; Potamianos et al., 1998), or parameters of a suitable lip model (Chandramohan and Silsbee, 1996; Dalton et al., 1996; Luetin et al., 1996). On the other hand, only a crude detection of the mouth region is sufficient to obtain visual features, using transformations of this region's pixel values, that achieve sufficient dimensionality reduction (Bregler

et al., 1993; Duchnowski et al., 1994; Matthews et al., 1996; Potamianos et al., 2001b). Needless to say, robust tracking of the lips or mouth region is of paramount importance for good performance of automatic speechreading systems (Iyengar et al., 2001).

Face Detection, Mouth and Lip Tracking

The problem of face and facial part detection has attracted significant interest in the literature (Graf et al., 1997; Rowley et al., 1998; Sung and Poggio, 1998; Senior, 1999). In addition to automatic speechreading, it has applications to other areas, such as visual text-to-speech (Cohen and Massaro, 1994; Chen et al., 1995; Cosatto et al., 2000), person identification and verification (Jourlin et al., 1997; Wark and Sridharan, 1998; Fröba et al., 1999; Jain et al., 1999; Maison et al., 1999; Chibelushi et al., 2002; Zhang et al., 2002), speaker localization (Bub et al., 1995; Wang and Brandstein, 1999; Zotkin et al., 2002), detection of intent to speak (De Cuetos et al., 2000), and image retrieval (Swets and Weng, 1996), among others. In general, robust face and mouth detection is quite difficult, especially in cases where the background, face pose, and lighting are varying (Iyengar and Neti, 2001).

In the audio-visual ASR literature, where issues such as visual feature design, or audio-visual fusion algorithms are typically of more interest, face and mouth detection are often ignored, or at least, care is taken for the simplification of the problem: In some databases for example, the speaker's lips are suitably colored, so that their automatic extraction becomes trivial by chroma-key methods (Adjoudani and Benoît, 1996; Heckmann et al., 2001). In other works, where audio-visual corpora are shared (for example, the Tulips1, (X)M2VTS, and AMP/CMU databases, discussed later), the mouth regions are extracted once and re-used in subsequent work by other researchers, or sites. In addition, in practically all databases, the faces are frontal with minor face pose and lighting variation.

In general, all audio-visual ASR systems require determining a *region-of-interest* (ROI) for the visual feature extraction algorithm to proceed. For example, such a ROI can be the entire face, in which case, a subsequent active appearance model can be used to match to the exact face location (Cootes et al., 1998). Alternatively, such a ROI can be the mouth only region, in which case, an active shape model of the lips can be used to fit a lip contour (Luetin et al., 1996). If appearance based visual features are to be extracted (see below), the latter is all that is required. Many techniques of varying complexity can be used to locate such ROIs. Some use traditional image processing techniques, such as color segmentation, edge detection, image thresholding, template matching, or motion information (Graf et al., 1997), whereas other methods use statistical modeling techniques, employing neural networks for example (Rowley et al., 1998). In the following, we describe an algorithm within the second category.

Face Detection and Mouth Region-of-Interest Extraction

A typical algorithm for face detection and facial feature localization is described in Senior (1999). This technique is used in the visual front end design of Neti et al. (2000) and Potamianos et al. (2001b), when processing the video of the IBM ViaVoiceTM audio-visual database, described later. 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 Potamianos et al. (2001b), 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.

In case the video signal is in color, *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

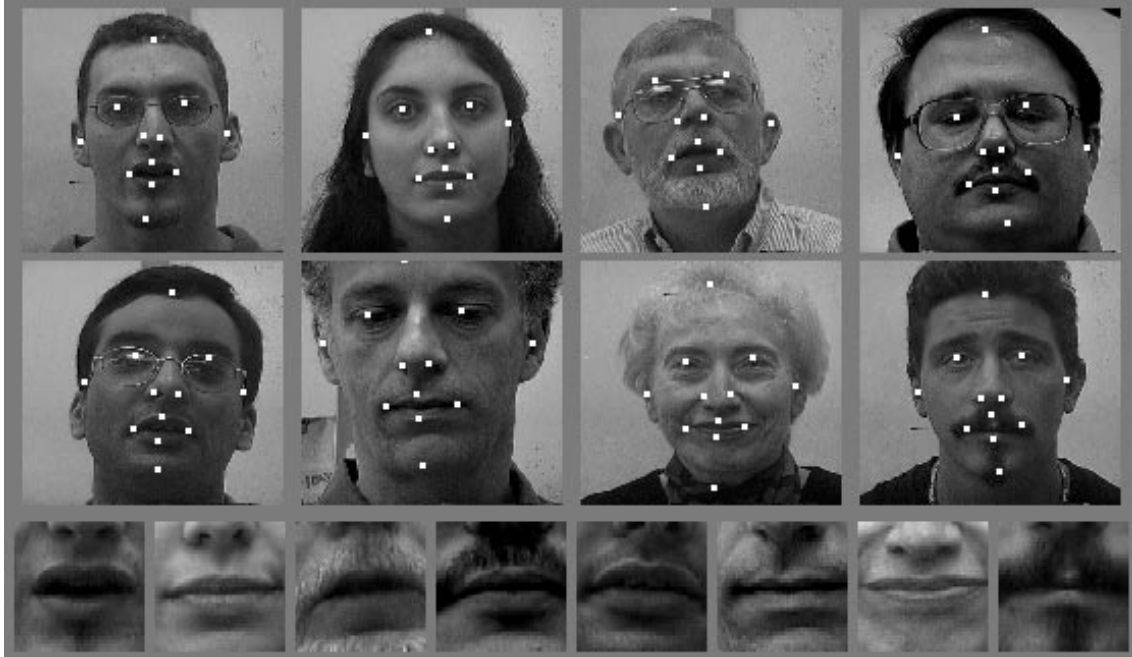


Figure 2: Region-of-interest extraction examples. *Upper rows:* Example video frames of eight subjects from the IBM ViaVoice™ audio-visual database (described in a later section), with superimposed facial features, detected by the algorithm of Senior (1999). *Lower row:* Corresponding mouth regions-of-interest, extracted as in Potamianos et al. (2001b).

to occupy a largely invariant to most humans and lighting conditions range of values (Graf et al., 1997; Senior, 1999). In the particular implementation, 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 a 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 *principal components analysis* (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 (twelve such facial features are marked on the frames of Figure 2). 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).

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. 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).

Given the output of the face detection and facial feature finding algorithm described above, five located lip

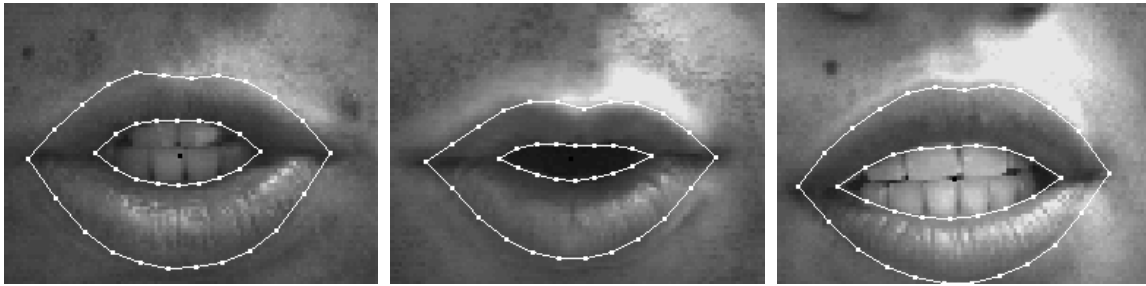


Figure 3: Examples of lip contour estimation by means of active shape models (Luettin et al., 1996). Depicted mouth regions are from the Tulips1 audio-visual database (Movellan and Chadderdon, 1996), and they have been extracted preceding lip contour estimation.

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, whereas the mouth size estimates are averaged over each utterance. A size-normalized square ROI is then extracted (see (1), below), with sides $M = N = 64$ (see also Figure 2). This can contain just the mouth region, or also parts of the lower face (Potamianos and Neti, 2001b).

Lip Contour Tracking

Once the mouth region is located, a number of algorithms can be used to obtain lip contour estimates. Some popular methods are *snakes* (Kass et al., 1988), *templates* (Yuille et al., 1992; Silsbee, 1994), and *active shape* and *appearance models* (Cootes et al., 1995, 1998).

A snake is an elastic curve represented by a set of control points. The control point coordinates are iteratively updated, by converging towards the local minimum of an energy function, defined on basis of curve smoothness constraints and a matching criterion to desired features of the image (Kass et al., 1988). Such an algorithm is used for lip contour estimation in the speechreading system of Chiou and Hwang (1997). Another widely used technique for lip tracking is by means of lip *templates*, employed in the system of Chandramohan and Silsbee (1996) for example. Templates constitute parametrized curves that are fitted to the desired shape by minimizing an energy function, defined similarly to snakes. *B-splines*, used by Dalton et al. (1996), work similarly to the above techniques as well.

Active shape and appearance models construct a lip shape or ROI appearance statistical model, as discussed in following subsections. These models can be used for tracking lips by means of the algorithm proposed by Cootes et al. (1998). This assumes that, given small perturbations from the actual fit of the model to a target image, a linear relationship exists between the difference in the model projection and image and the required updates to the model parameters. An iterative algorithm is used to fit the model to the image data (Matthews et al., 1998). Alternatively, the fitting can be performed by the downhill simplex method (Nelder and Mead, 1965), as in Luettin et al. (1996). Examples of lip contour estimation by means of active shape models using the latter fitting technique are depicted in Figure 3.

Visual Features

Various sets of visual features for automatic speechreading have been proposed in the literature over the last 20 years. In general, they can be grouped into three categories: (a): *Video pixel* (or, *appearance*) based ones; (b): *Lip contour* (or, *shape*) based features; and (c): Features that are a combination of *both* appearance and

shape (Hennecke et al., 1996). In the following, we present each category in more detail. Possible feature post-extraction processing is discussed at the end of this section.

Appearance Based Features

In this approach to visual feature extraction, the image part typically containing the speaker's mouth region is considered as informative for lipreading, i.e., the region-of-interest (ROI). Such region can be a rectangle containing the mouth, and possibly include larger parts of the lower face, such as the jaw and cheeks (Potamianos and Neti, 2001b), or the entire face (Matthews et al., 2001). Often, it can be a three-dimensional rectangle, containing adjacent frame rectangular ROIs, in an effort to capture dynamic speech information at this early stage of processing (Li et al., 1995; Potamianos et al., 1998). Alternatively, the ROI can correspond to a number of image profiles vertical to the lip contour (Dupont and Luetttin, 2000), or be just a disc around the mouth center (Duchnowski et al., 1994). By concatenating the ROI pixel *greyscale* (Bregler et al., 1993; Duchnowski et al., 1994; Potamianos et al., 1998; Dupont and Luetttin, 2000), or *color* values (Chiou and Hwang, 1997), a feature vector is obtained. For example, in the case of an $M \times N$ -pixel rectangular ROI, which is centered at location (m_t, n_t) of video frame $V_t(m, n)$ at time t , the resulting feature vector of length $d = MN$ will be (after a lexicographic ordering)¹

$$\mathbf{x}_t \leftarrow \{ V_t(m, n) : m_t - \lfloor M/2 \rfloor \leq m < m_t + \lceil M/2 \rceil, n_t - \lfloor N/2 \rfloor \leq n < n_t + \lceil N/2 \rceil \}. \quad (1)$$

This vector is expected to contain most visual speech information. Notice that approaches that use *optical flow* as visual features (Mase and Pentland, 1991; Gray et al., 1997) can fit within this framework, by replacing in (1) the video frame ROI pixels with optical flow estimates.

Typically, the dimensionality d of vector (1) is too large to allow successful statistical modeling (Chatfield and Collins, 1991) of speech classes, by means of a hidden Markov model (HMM) for example (Rabiner and Juang, 1993). Therefore, appropriate transformations of the ROI pixel values are used as visual features. Movellan and Chadderdon (1996) for example, use low-pass filtering followed by image subsampling and video frame ROI differencing, whereas Matthews et al. (1996) propose a nonlinear image decomposition using "image sieves" for dimensionality reduction and feature extraction. By far however, the most popular appearance feature representations achieve such reduction by using traditional *image transforms* (Gonzalez and Wintz, 1977). These transforms are typically borrowed from the image compression literature, and the hope is that they will preserve most relevant to speechreading information. In general, a $D \times d$ -dimensional *linear transform* matrix \mathbf{P} is sought, such that the transformed data vector $\mathbf{y}_t = \mathbf{P} \mathbf{x}_t$ contains most speechreading information in its $D \ll d$ elements. To obtain matrix \mathbf{P} , L training examples are given, denoted by \mathbf{x}_l , $l = 1, \dots, L$. A number of possible such matrices are described in the following.

Principal components analysis (PCA) - This constitutes the most popular pixel based feature representation for automatic speechreading (Bregler et al., 1993; Bregler and Konig, 1994; Duchnowski et al., 1994; Li et al., 1995; Brooke, 1996; Tomlinson et al., 1996; Chiou and Hwang, 1997; Gray et al., 1997; Luetttin and Thacker, 1997; Potamianos et al., 1998; Dupont and Luetttin, 2000). The PCA data projection achieves optimal information compression, in the sense of minimum square error between the original vector \mathbf{x}_t and its reconstruction based on its projection \mathbf{y}_t , however appropriate data scaling constitutes a problem in the classification of the resulting vectors (Chatfield and Collins, 1991). In the PCA implementation of Potamianos et al. (1998), the data are scaled according to their inverse variance, and their correlation matrix \mathbf{R} is computed. Subsequently, \mathbf{R} is diagonalized as $\mathbf{R} = \mathbf{A} \mathbf{\Lambda} \mathbf{A}^T$ (Chatfield and Collins, 1991; Press et al., 1995), where $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_d]$ has as columns the *eigenvectors* of \mathbf{R} , and $\mathbf{\Lambda}$ is a diagonal matrix containing the *eigenvalues* of \mathbf{R} . Assuming that the D largest such eigenvalues are located at the j_1, \dots, j_D diagonal positions, the data projection matrix is $\mathbf{P}_{\text{PCA}} = [\mathbf{a}_{j_1}, \dots, \mathbf{a}_{j_D}]^T$. Given a data vector \mathbf{x}_t , this is first element-wise mean and variance normalized, and subsequently, its feature vector is extracted as $\mathbf{y}_t = \mathbf{P}_{\text{PCA}} \mathbf{x}_t$.

Discrete cosine, wavelet, and other image transforms - As an alternative to PCA, a number of popular linear image transforms (Gonzalez and Wintz, 1977) have been used in place of \mathbf{P} for obtaining speechreading

¹Throughout this work, boldface lowercase symbols denote column vectors, and boldface capital symbols denote matrices. In addition, \bullet^T denotes vector or matrix *transpose*, and $\text{diag}(\bullet)$, $\det(\bullet)$, denote matrix *diagonal* and *determinant*, respectively.

features. For example, the *discrete cosine transform* (DCT) has been adopted in several systems (Duchnowski et al., 1994; Potamianos et al., 1998; Nakamura et al., 2000; Neti et al., 2000; Scanlon and Reilly, 2001; Nefian et al., 2002), the *discrete wavelet transform* (DWT - Daubechies, 1992) in others (Potamianos et al., 1998), and the *Hadamard* and *Haar* transforms by Scanlon and Reilly (2001). Most researchers use separable transforms (Gonzalez and Wintz, 1977), which allow fast implementations (Press et al., 1995) when M and N are powers of 2 (typically, values $M, N = 16, 32$, or 64 are considered). Notice that, in each case, matrix \mathbf{P} can have as rows the image transform matrix rows that maximize the transformed data energy over the training set (Potamianos et al., 1998), or alternatively, that correspond to a-priori chosen locations (Nefian et al., 2002).

Linear discriminant analysis (LDA) - The data vector transforms presented above are more suitable for ROI compression than ROI classification into the set of speech classes of interest. For the latter task, LDA (Rao, 1965) is more appropriate, as it maps features to a new space for improved classification. LDA was first proposed for automatic speechreading by Duchnowski et al. (1994). There, it was applied directly to vector (1). LDA has also been considered in a cascade, following the PCA projection of a single frame ROI vector, or on the concatenation of a number of adjacent PCA projected vectors (Matthews et al., 2001).

LDA assumes that a set of *classes* \mathcal{C} (such as HMM states) is a-priori chosen, and, in addition, that the training set data vectors \mathbf{x}_l , $l = 1, \dots, L$, are *labeled* as $c(l) \in \mathcal{C}$. Then, it seeks matrix \mathbf{P}_{LDA} , such that the projected training sample $\{\mathbf{P}_{\text{LDA}} \mathbf{x}_l, l = 1, \dots, L\}$ is “well separated” into the set of classes \mathcal{C} , according to a function of the training sample *within-class scatter* matrix \mathbf{S}_W and its *between-class scatter* matrix \mathbf{S}_B (Rao, 1965). These matrices are given by

$$\mathbf{S}_W = \sum_{c \in \mathcal{C}} Pr(c) \Sigma^{(c)}, \quad \text{and} \quad \mathbf{S}_B = \sum_{c \in \mathcal{C}} Pr(c) (\mathbf{m}^{(c)} - \mathbf{m})(\mathbf{m}^{(c)} - \mathbf{m})^\top, \quad (2)$$

respectively. In (2), $Pr(c) = L_c/L$, $c \in \mathcal{C}$, is the class empirical probability mass function, where $L_c = \sum_{l=1}^L \delta_{c(l),c}$, and $\delta_{i,j} = 1$, if $i = j$; 0, otherwise; in addition, $\mathbf{m}^{(c)}$ and $\Sigma^{(c)}$ denote the class sample mean and covariance, respectively; and finally, $\mathbf{m} = \sum_{c \in \mathcal{C}} Pr(c) \mathbf{m}^{(c)}$ is the total sample mean. To estimate \mathbf{P}_{LDA} , the *generalized* eigenvalues and *right* eigenvectors of the matrix pair $(\mathbf{S}_B, \mathbf{S}_W)$, that satisfy $\mathbf{S}_B \mathbf{F} = \mathbf{S}_W \mathbf{F} \mathbf{\Lambda}$, are first computed (Rao, 1965; Golub and Van Loan, 1983). Matrix $\mathbf{F} = [\mathbf{f}_1, \dots, \mathbf{f}_d]$ has as columns the generalized eigenvectors. Assuming that the D largest eigenvalues are located at the j_1, \dots, j_D diagonal positions of $\mathbf{\Lambda}$, then, $\mathbf{P}_{\text{LDA}} = [\mathbf{f}_{j_1}, \dots, \mathbf{f}_{j_D}]^\top$. It should be noted that, due to (2), 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 $D \leq |\mathcal{C}| - 1$ should hold. In addition, the rank of the $d \times d$ -dimensional matrix \mathbf{S}_W cannot exceed $L - |\mathcal{C}|$, therefore having insufficient training data, with respect to the input feature vector dimension d , is a potential problem.

Maximum Likelihood Data Rotation (MLLT) - In our speechreading system (Potamianos et al., 2001b), LDA is followed by the application of a data *maximum likelihood linear transform* (MLLT). This transform seeks a square, non-singular, data rotation matrix \mathbf{P}_{MLLT} that maximizes the observation data likelihood in the original feature space, under the assumption of diagonal data covariance in the transformed space (Gopinath, 1998). Such a rotation is beneficial, since in most ASR systems, diagonal covariances are typically assumed, when modeling the observation class conditional probability distribution with Gaussian mixture models. The desired rotation matrix is obtained as

$$\mathbf{P}_{\text{MLLT}} = \arg \max_{\mathbf{P}} \{ \det(\mathbf{P})^L \prod_{c \in \mathcal{C}} (\det(\text{diag}(\mathbf{P} \Sigma^{(c)} \mathbf{P}^\top))^{-\frac{L_c}{2}}) \}$$

(Gopinath, 1998). This can be solved numerically (Press et al., 1995).

Notice that LDA and MLLT are data transforms aiming in improved classification performance and maximum likelihood data modeling. Therefore, their application can be viewed as a feature post-processing stage, and clearly, should not be limited to appearance only visual data.

Shape Based Features

In contrast to appearance based features, shape based feature extraction assumes that most speechreading information is contained in the *shape* (contours) of the speaker’s lips, or more generally (Matthews et al.,

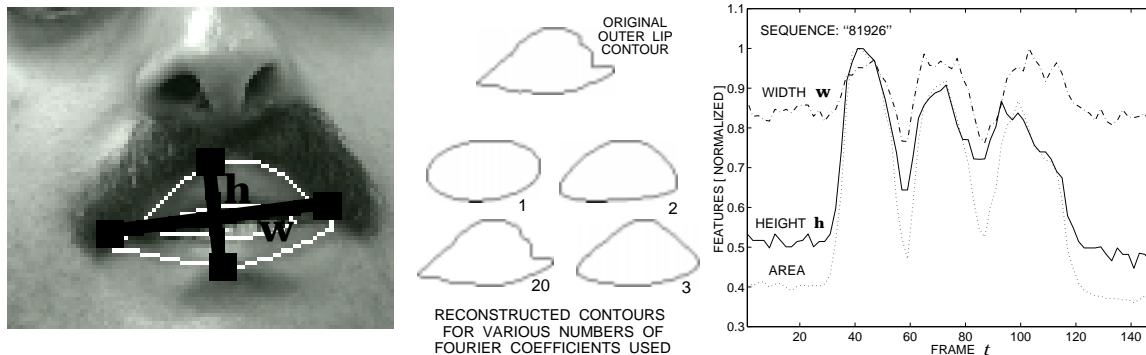


Figure 4: Geometric feature approach. *Left:* Outer lip width (w) and height (h). *Middle:* Reconstruction of an estimated outer lip contour (upper part) from 1, 2, 3, and 20 sets of its Fourier coefficients (lower part, clockwise). *Right:* Three geometric visual features, displayed on a normalized scale, tracked over the spoken utterance “81926” of the connected digits database of Potamianos et al. (1998). Lip contours are estimated as in Graf et al. (1997).

2001), in the face contours (e.g., jaw and cheek shape, in addition to the lips). Two types of features fall within this category: Geometric type ones, and shape model based features. In both cases, an algorithm that extracts the inner and/or outer lip contours, or in general, the face shape, is required. A variety of such algorithms were discussed above.

Lip geometric features - Given the lip contour, a number of high level features, meaningful to humans, can be readily extracted, such as the contour *height*, *width*, *perimeter*, as well as the *area* contained within the contour. As demonstrated in Figure 4, such features do contain significant speech information. Not surprisingly, a large number of speechreading systems makes use of all or a subset of them (Petajan, 1984; Adjoudani and Benoît, 1996; Alissali et al., 1996; Goldschen et al., 1996; André-Obrecht et al., 1997; Jourlin, 1997; Chan et al., 1998; Rogozan and Deléglise, 1998; Teissier et al., 1999; Zhang et al., 2000; Gurbuz et al., 2001; Heckmann et al., 2001; Huang and Chen, 2001).

Additional visual features can be derived from the lip contours, such as lip *image moments* and lip contour *Fourier descriptors* (see Figure 4), that are invariant to affine image transformations. Indeed, a number of central moments of the contour interior binary image, or its *normalized moments*, as defined in Dougherty and Giardina (1987), have been considered as visual features (Czap, 2000). Normalized Fourier series coefficients of a contour parametrization (Dougherty and Giardina, 1987) have also been used to augment previously discussed geometric features in some speechreading systems, resulting to improved automatic speechreading (Potamianos et al., 1998; Gurbuz et al., 2001).

Lip model features - A number of *parametric* models (Basu et al., 1998) have been used for lip- or face-shape tracking in the literature, and briefly reviewed in a previous subsection. The parameters of these models can be readily used as visual features. For example, Chiou and Hwang (1997) employ a snake based algorithm to estimate the lip contour, and subsequently use a number of snake radial vectors as visual features. Su and Silsbee (1996), as well as Chandramohan and Silsbee (1996), use lip template parameters instead.

Another popular lip model is the *active shape model* (ASM). ASMs are flexible *statistical* models that represent an object by a set of labeled points (Cootes et al., 1995; Luetttin et al., 1996). Such object can be the inner and/or outer lip contour (Luetttin and Thacker, 1997), or the union of various face shape contours as in Matthews et al. (2001). To derive an ASM, a number of K contour points are first labeled on available training set images, and their co-ordinates are placed on $2K$ -dimensional “shape” vectors

$$\mathbf{x}^{(s)} = [x_1, y_1, x_2, y_2, \dots, x_K, y_K]^T. \quad (3)$$

Given a set of vectors (3), PCA can be used to identify the optimal orthogonal linear transform $\mathbf{P}_{PCA}^{(s)}$ in terms of the variance described along each dimension, resulting in a statistical model of the lip or facial shape (see

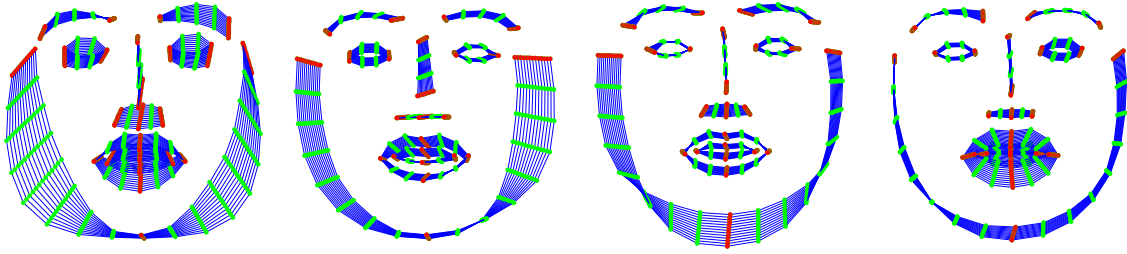


Figure 5: Statistical shape model. The top four modes are plotted (*left-to-right*) at ± 3 standard deviations around the mean. These four modes describe 65% of the variance of the training set, which consists of 4072 labeled images from the IBM ViaVoiceTM audio-visual database (Neti et al., 2000; Matthews et al., 2001).

Figure 5). To identify axes of genuine shape variation, each shape in the training set must be aligned. This is achieved using a similarity transform (translation, rotation, and scaling), by means of an iterative procrustes analysis (Cootes et al., 1995; Dryden and Mardia, 1998). Given a tracked lip contour, the extracted visual features will be $\mathbf{y}^{(S)} = \mathbf{P}_{PCA}^{(S)} \mathbf{x}^{(S)}$. Note that vectors (3) can be the output of a tracking algorithm based on B-splines for example, as in Dalton et al. (1996).

Joint Appearance and Shape Features

Appearance and shape based visual features are quite different in nature. In a sense they code low- and high-level information about the speaker's face and lip movements. Not surprisingly, combinations of features from both categories have been employed in a number of automatic speechreading systems.

In most cases, features from each category are just concatenated. For example, Chan (2001) combines geometric lip features with the PCA projection of a subset of pixels contained within the mouth. Luetin et al. (1996), as well as Dupont and Luetin (2000), combine ASM features with PCA based ones, extracted from a ROI that consists of short image profiles around the lip contour. Chiou and Hwang (1997) on the other hand, combine a number of snake lip contour radial vectors with PCA features of the color pixel values of a rectangle mouth ROI.

A different approach to combining the two classes of features is to create a *single* model of face shape and appearance. An *active appearance model* (AAM - Cootes et al., 1998) provides a framework to statistically combine them. Building an AAM requires three applications of PCA:

- (a) Shape eigenspace calculation that models shape deformations, resulting in PCA matrix $\mathbf{P}_{PCA}^{(S)}$, computed as above (see (3)).
- (b) Appearance eigenspace calculation to model appearance changes, resulting in a PCA matrix $\mathbf{P}_{PCA}^{(A)}$, of the ROI appearance vectors. If the *color* values of the $M \times N$ -pixel ROI are considered, such vectors are

$$\mathbf{x}^{(A)} = [r_1, g_1, b_1, r_2, g_2, b_2, \dots, r_{MN}, g_{MN}, b_{MN}]^T, \quad (4)$$

similar to vectors (1).

- (c) Using these, calculation of a combined shape and appearance eigenspace. The latter is a PCA matrix $\mathbf{P}_{PCA}^{(A,S)}$ on training vectors

$$\mathbf{x}^{(A,S)} = [\mathbf{x}^{(A)\top} \mathbf{W} \mathbf{P}_{PCA}^{(A)\top}, \mathbf{x}^{(S)\top} \mathbf{P}_{PCA}^{(S)\top}]^T = [\mathbf{y}^{(A)\top} \mathbf{W}, \mathbf{y}^{(S)\top}]^T,$$

where \mathbf{W} is a suitable diagonal scaling matrix (Matthews et al., 2001). The aim of this final PCA is to remove the redundancy due to the shape and appearance correlation, and to create a single model that compactly describes shape and the corresponding appearance deformation.



Figure 6: Combined shape and appearance statistical model. *Center row:* Mean shape and appearance. *Top row:* Mean shape and appearance +3 standard deviations. *Bottom row:* Mean shape and appearance -3 standard deviations. The top four modes, depicted *left-to-right*, describe 46% of the combined shape and appearance variance of 4072 labeled images from the IBM ViaVoice™ audio-visual database (Neti et al., 2000; Matthews et al., 2001).

Such a model has been used for speechreading in Neti et al. (2000) and Matthews et al. (2001). An example of the resulting learned joined model is depicted in Figure 6. A block diagram of the method, including the dimensionalities of the input shape and appearance vectors ((3) and (4), respectively), their PCA projections $\mathbf{y}^{(S)}$, $\mathbf{y}^{(A)}$, and the final feature vector $\mathbf{y}^{(A,S)} = \mathbf{P}_{\text{PCA}}^{(A,S)} \mathbf{x}^{(A,S)}$, is depicted in Figure 7.

Visual Feature Post-Extraction Processing

In an audio-visual speech recognition system, in addition to the visual features, audio features are also extracted from the acoustic waveform. For example, such features could be *mel-frequency cepstral coefficients* (MFCCs), or *linear prediction coefficients* (LPCs), typically extracted at a 100 Hz rate (Deller et al., 1993; Rabiner and Juang, 1993; Young et al., 1999). In contrast, visual features are generated at the video frame rate, commonly 25 or 30 Hz, or twice that, in case of interlaced video. Since feature stream synchrony is required in a number of algorithms for audio-visual fusion, as discussed in the next section, the two feature streams must attain the same rate. Typically, this is accomplished (whenever required), either after feature extraction, by simple element-wise *linear interpolation* of the visual features to the audio frame rate (as in Figure 7), or before feature extraction, by frame duplication to achieve a 100 Hz video input rate to the visual

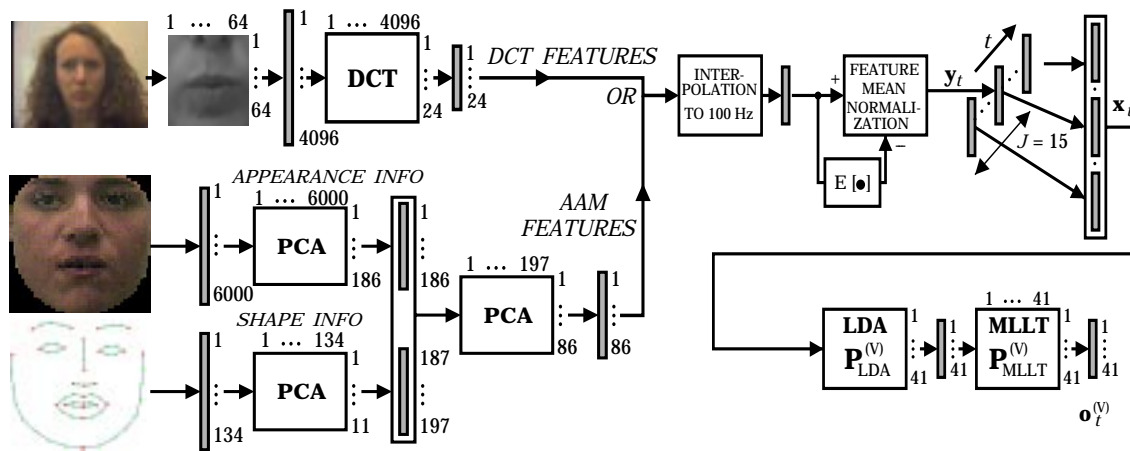


Figure 7: DCT versus AAM based visual feature extraction for automatic speechreading, followed by visual feature post-extraction processing using linear interpolation, feature mean normalization, adjacent frame feature concatenation, and the application of LDA and MLLT. Vector dimensions as implemented in the system of Neti et al. (2000) are depicted.

front end. Occasionally, the audio front end processing is performed at the lower video rate.

Another interesting issue in visual feature extraction has to do with feature normalization. In a traditional audio front end, *cepstral mean subtraction* is often employed to enhance robustness to speaker and environment variations (Liu et al., 1993; Young et al., 1999). A simple visual *feature mean normalization* (FMN) by element-wise subtraction of the vector mean over each sentence has been demonstrated to improve appearance feature based, visual-only recognition (Potamianos et al., 1998, 2001b). Alternatively, linear intensity compensation has been investigated preceding the appearance feature extraction by Vanegas et al. (1998).

A very important issue in the visual feature design is capturing the dynamics of visual speech. Temporal information, often spanning multiple phone segments, is known to help human perception of visual speech (Rosenblum and Saldaña, 1998). Borrowing again from the ASR literature, dynamic speech information can be captured by augmenting the visual feature vector by its *first* and *second* order temporal *derivatives* (Rabiner and Juang, 1993; Young et al., 1999). Alternatively, LDA can be used, as a means of “learning” a transform that optimally captures the speech dynamics. Such a transform is applied on the concatenation of consecutive feature vectors adjacent and including the current frame (see also Figure 7), i.e., on

$$\mathbf{x}_t = [\mathbf{y}_{t-\lceil J/2 \rceil}^\top, \dots, \mathbf{y}_t^\top, \dots, \mathbf{y}_{t+\lceil J/2 \rceil - 1}^\top]^\top, \tag{5}$$

with $J = 15$ for example, as in Neti et al. (2000) and Potamianos et al. (2001b).

Clearly, and as we already mentioned, LDA could be applied to any category of features discussed. The same holds for MLLT, a method that aims in improving maximum likelihood data modeling, and in practice, ASR performance. For example, a number of feature post-processing steps discussed above, including LDA and MLLT, have been interchangeably applied to DCT appearance features, as well as to AAM ones, in our visual front end experiments during the Johns Hopkins workshop, as depicted in Figure 7 (Neti et al., 2000; Matthews et al., 2001). Alternate ways of combining feature post-extraction processing steps can easily be envisioned. For example, LDA and MLLT can be applied to obtain within-frame discriminant features (Potamianos and Neti, 2001b), which can then be augmented by their first and second order derivatives, or followed by LDA and MLLT across frames (see also Figure 11).

Finally, an important problem in data classification is the issue of *feature selection* within a larger pool of candidate features (Jain et al., 2000). In the context of speechreading, this matter has been directly addressed in the selection of geometric, lip contour based features by Goldschen et al. (1996).

Summary of Visual Front End Algorithms

We have presented a summary of the most common visual feature extraction algorithms proposed in the literature for automatic speechreading. Such techniques differ both in their assumptions about where the speechreading information lies, as well as in the requirements that they place on face detection, facial part localization, and tracking. On the one extreme, appearance based visual features consider a broadly defined ROI, and then rely on traditional pattern recognition and image compression techniques to extract relevant speechreading information. On the opposite side, shape based visual features require adequate lip or facial shape tracking, and assume that the visual speech information is captured by this shape's form and movement alone. Bridging the two extremes, various combinations of the two types of features have also been used, ranging from simple concatenation to their joint modeling.

Comparisons between features within the same class are often reported in the literature (Duchnowski et al., 1994; Goldschen et al., 1996; Gray et al., 1997; Potamianos et al., 1998; Matthews et al., 2001; Scanlon and Reilly, 2001). Unfortunately however, comparisons across the various types of features are rather limited, as they require widely different sets of algorithms for their implementation. Nevertheless, Matthews et al. (1998) demonstrate AAMs to outperform ASMs, and to result in similar visual-only recognition to alternative appearance based features. Chiou and Hwang (1997) report that their joint features outperform their shape and appearance feature components, whereas Potamianos et al. (1998), as well as Scanlon and Reilly (2001), report that DCT transform based visual features are superior to a set of lip contour geometric features. However, the above results are reported on single-subject data and/or small vocabulary tasks. In a larger experiment, Matthews et al. (2001) compare a number of appearance based features with AAMs on a speaker-independent LVCSR task. All appearance features considered outperformed AAMs, however it is suspected that the AAM used there was not sufficiently trained.

Although much progress has been made in visual feature extraction, it seems that the question of what are the best visual features for automatic speechreading, that are robust in a variety of visual environments, remains to a large extent unresolved. Of particular importance is that such features should exhibit sufficient speaker, pose, camera, and environment independence. However, it is worth mentioning two arguments in favor of appearance based features: First, their use is well motivated by human perception studies of visual speech. Indeed, significant information about the place of articulation, such as tongue and teeth visibility, cannot be captured by the lip contours alone. Human speech perception based on the mouth region is superior than perception on basis of the lips alone, and it further improves when the entire lower face is visible (Summerfield et al., 1989). Second, the extraction of certain highly performing appearance based features such as the DCT is computationally efficient. Indeed, it requires a crude mouth region detection algorithm, which can be applied at a low frame rate, whereas the subsequent pixel vector transform is amenable to fast implementation for suitable ROI sizes (Press et al., 1995). These facts enable the implementation of real-time automatic speechreading systems.

AUDIO-VISUAL INTEGRATION FOR SPEECH RECOGNITION

Audio-visual fusion is an instance of the general classifier combination problem (Jain et al., 2000). In our case, two observation streams are available (audio and visual modalities) and provide information about speech classes, such as context-dependent sub-phonetic units, or at a higher level, word sequences. Each observation stream can be used alone to train single-modality statistical classifiers to recognize such classes. However, one hopes that combining the two streams will give rise to a bimodal classifier with superior performance to both single-modality ones.

Various information fusion algorithms have been considered in the literature for audio-visual ASR (for example, Bregler et al., 1993; Adjoudani and Benoît, 1996; Hennecke et al., 1996; Potamianos and Graf, 1998; Rogozan, 1999; Teissier et al., 1999; Dupont and Luetin, 2000; Neti et al., 2000; Chen, 2001; Chu and Huang, 2002). The proposed techniques differ both in their basic design, as well as in the adopted terminology. The architecture of some of these methods (Robert-Ribes et al., 1996; Teissier et al., 1999) is motivated by models of human speech perception (Massaro, 1996; Massaro and Stork, 1998; Berthommier, 2001). In

FUSION TYPE	AUDIO-VISUAL FEATURES	CLASSIFICATION LEVEL
Feature fusion: <i>One classifier is used</i>	1. Concatenated features 2. Hierarchical discriminant features 3. Enhanced audio features	Sub-phonetic (early)
Decision fusion: <i>Two classifiers are used</i>	Concatenated features	1. Sub-phonetic (early) 2. Phone or word (intermediate) 3. Utterance (late)

Table 1: Taxonomy of the audio-visual integration methods considered in this section. Three feature fusion techniques, that differ in the features used for recognition, and three decision fusion methods, that differ in the combination stage of the audio and visual classifiers, are described in more detail in this chapter.

most cases however, research in audio-visual ASR has followed a separate track from work on modeling the human perception of audio-visual speech.

Audio-visual integration techniques 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 (Adjoudani and Benoit, 1996; Teissier et al., 1999; Potamianos et al., 2001a). 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, 1998; Dupont and Luettin, 2000; Neti et al., 2000).

In this section, we provide a detailed description of some popular fusion techniques from each category (see also Table 1). In addition, we briefly address two issues relevant to automatic recognition of audio-visual speech: One is the problem of speech modeling for ASR, which poses particular interest in automatic speechreading, and helps establish some background and notation for the remainder of the section. We also consider the subject of speaker adaptation, an important element in practical ASR systems.

Audio-Visual Speech Modeling for ASR

Two central aspects in the design of ASR systems are the choice of speech classes, that are assumed to generate the observed features, and the statistical modeling of this generation process. In the following, we briefly discuss both issues, since they are often embedded into the design of audio-visual fusion algorithms.

Speech Classes for Audio-Visual ASR

The basic unit that describes how speech conveys linguistic information is the *phoneme*. For American English, there exist approximately 42 such units (Deller et al., 1993), generated by specific positions or movements of the vocal tract articulators. Only some of the articulators are visible however, therefore among these phonemes, the number of visually distinguishable units is much smaller. Such units are called *visemes* in the audio-visual ASR and human perception literatures (Stork and Hennecke, 1996; Campbell et al., 1998; Massaro and Stork, 1998). In general, phoneme to viseme mappings are derived by human speechreading studies. Alternatively, such mappings can be generated using statistical clustering techniques, as proposed by Goldschen et al. (1996) and Rogozan (1999). There is no universal agreement about the exact partitioning of phonemes into visemes, but some visemes are well-defined, such as the bilabial viseme consisting of phoneme set $\{ /p/, /b/, /m/ \}$. A typical clustering into 13 visemes is used by Neti et al. (2000) to conduct visual speech modeling experiments, and is depicted in Table 2.

In traditional audio-only ASR, the set of classes $c \in \mathcal{C}$ that need to be estimated on basis of the observed feature

Silence	/sil/, /sp/	Alveolar-fricatives	/s/, /z/
Lip-rounding based vowels	/ao/, /ah/, /aa/, /er/, /oy/, /aw/, /hh/	Alveolar	/t/, /d/, /n/, /en/
	/uw/, /uh/, /ow/	Palato-alveolar	/sh/, /zh/, /ch/, /jh/
	/ae/, /eh/, /ey/, /ay/	Bilabial	/p/, /b/, /m/
	/ih/, /iy/, /ax/	Dental	/th/, /dh/
Alveolar-semivowels	/l/, /el/, /r/, /y/	Labio-dental	/f/, /v/
		Velar	/ng/, /k/, /g/, /w/

Table 2: The 44 phoneme to 13 viseme mapping considered by Neti et al. (2000), using the HTK phone set (Young et al., 1999).

sequence most often consist of sub-phonetic units, and occasionally of sub-word units in small vocabulary recognition tasks. For LVCSR, a large number of context-dependent sub-phonetic units are used, obtained by clustering the possible phonetic contexts (tri-phone ones, for example) by means of a decision tree (Deller et al., 1993; Rabiner and Juang, 1993; Young et al., 1999). In this chapter, such units are exclusively used, defined over tri- or eleven-phone contexts, as described in the Experiments section.

For automatic speechreading, it seems appropriate, from the human visual speech perception point of view, to use visemic sub-phonetic classes, and their decision tree clustering based on visemic context. Such clustering experiments are reported by Neti et al. (2000). In addition, visual-only recognition of visemes is occasionally considered in the literature (Potamianos et al., 2001b). Visemic speech classes are also used for audio-visual ASR at the second stage of a cascade decision fusion architecture proposed by Rogozan (1999). However, the use of different classes for its audio- and visual-only components complicates audio-visual fusion, with unclear performance gains. Therefore, in the remainder of this section, identical classes and decision trees are being used for both modalities.

HMM Based Speech Recognition

The most widely used classifier for audio-visual ASR is the *hidden Markov model* (HMM), a very popular method for traditional audio-only speech recognition (Deller et al., 1993; Rabiner and Juang, 1993; Young et al., 1999). Additional methods also exist for automatic recognition of speech, and have been employed in audio-visual ASR systems, such as *dynamic time warping* (DTW), used for example by Petajan (1984), *artificial neural networks* (ANN), as in Krone et al. (1997), hybrid ANN-DTW systems (Bregler et al., 1993; Duchnowski et al., 1994), or hybrid ANN-HMM ones (Heckmann et al., 2001). Various types of HMMs have also been used for audio-visual ASR, such as HMMs with discrete observations after vector quantization of the feature space (Silsbee and Bovik, 1996), or HMMs with non-Gaussian continuous observation probabilities (Su and Silsbee, 1996). However, the vast majority of audio-visual ASR systems, and to which we restrict the presentation in this chapter, employ HMMs with a continuous observation probability density, modeled as a mixture of Gaussian densities.

Typically in the literature, *single-stream* HMMs are used to model the generation of a sequence of audio-only or visual-only speech informative features, $\{\mathbf{o}_t^{(s)}\}$, of dimensionality D_s , where $s = A, V$ denotes the audio or visual modality (stream). The HMM *emission* (class conditional observation) probabilities are modeled by *Gaussian mixture* densities, given by

$$Pr[\mathbf{o}_t^{(s)} | c] = \sum_{k=1}^{K_{sc}} w_{sck} \mathcal{N}_{D_s}(\mathbf{o}_t^{(s)}; \mathbf{m}_{sck}, \mathbf{s}_{sck}), \quad (6)$$

for all classes $c \in \mathcal{C}$, whereas the HMM *transition* probabilities between the various classes are given by $\mathbf{r}_s = [\{Pr[c' | c'']\}, c', c'' \in \mathcal{C}]^T$. The HMM parameter vector is therefore

$$\mathbf{a}_s = [\mathbf{r}_s^T, \mathbf{b}_s^T]^T, \quad \text{where } \mathbf{b}_s = [\{[w_{sck}, \mathbf{m}_{sck}^T, \mathbf{s}_{sck}^T]^T, k = 1, \dots, K_{sc}, c \in \mathcal{C}\}]^T. \quad (7)$$

In (6) and (7), $c \in \mathcal{C}$ denote the HMM context dependent states, whereas mixture weights w_{sck} are positive

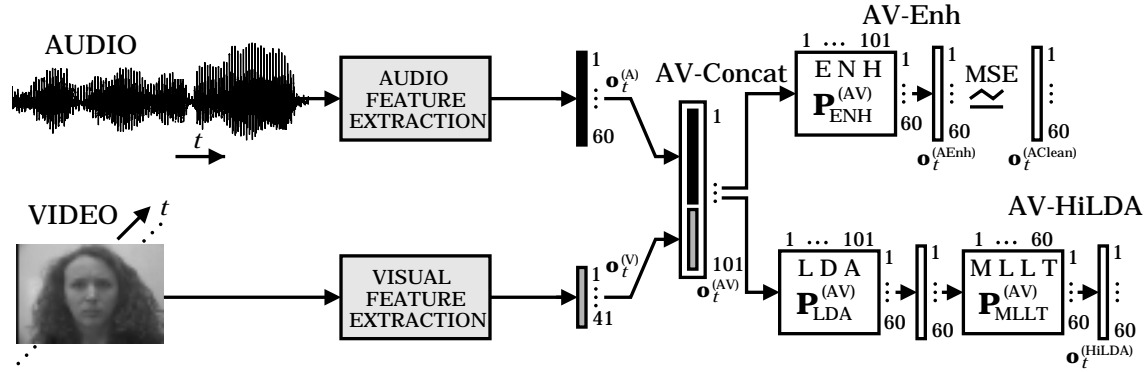


Figure 8: Three types of feature fusion considered in this section: Plain audio-visual feature concatenation (AV-Concat), hierarchical discriminant feature extraction (AV-HiLDA), and audio-visual speech enhancement (AV-Enh).

adding to one, K_{sc} denotes the number of mixtures, and $\mathcal{N}_D(\mathbf{o}; \mathbf{m}, \mathbf{s})$ is the D -variate normal distribution with mean \mathbf{m} and a diagonal covariance matrix, its diagonal being denoted by \mathbf{s} .

The *expectation-maximization* (EM) algorithm (Dempster et al., 1977) is typically used to obtain *maximum likelihood* estimates of (7). Given a current HMM parameter vector at EM algorithm iteration j , $\mathbf{a}_s^{(j)}$, a re-estimated parameter vector is obtained as

$$\mathbf{a}_s^{(j+1)} = \arg \max_{\mathbf{a}} Q(\mathbf{a}_s^{(j)}, \mathbf{a} | \mathbf{O}^{(s)}). \quad (8)$$

In (8), $\mathbf{O}^{(s)}$ denotes training data observations from L utterances $\mathbf{O}_l^{(s)}$, $l = 1, \dots, L$, and $Q(\bullet, \bullet | \bullet)$ represents the EM algorithm *auxiliary function*, defined as (Rabiner and Juang, 1993)

$$Q(\mathbf{a}', \mathbf{a}'' | \mathbf{O}^{(s)}) = \sum_{l=1}^L \sum_{\mathbf{c}(l)} Pr[\mathbf{O}_l^{(s)}, \mathbf{c}(l) | \mathbf{a}'] \log Pr[\mathbf{O}_l^{(s)}, \mathbf{c}(l) | \mathbf{a}'']. \quad (9)$$

In (9), $\mathbf{c}(l)$ denotes any HMM state sequence for utterance l . Replacing it with the best HMM path, reduces EM to *Viterbi* training (Deller et al., 1993). As an alternative to maximum likelihood, *discriminative training* methods can instead be used for HMM parameter estimation (Bahl et al., 1986; Chou et al., 1994).

Feature Fusion Techniques for Audio-Visual ASR

As already mentioned, feature fusion uses a single classifier to model the concatenated vector of time-synchronous audio and visual features, or appropriate transformations of it. Such methods include plain feature concatenation (Adjoudani and Benoît, 1996), feature weighting (Teissier et al., 1999; Chen, 2001), both also known as *direct identification* fusion (Teissier et al., 1999), and hierarchical linear discriminant feature extraction (Potamianos et al., 2001a). The *dominant* and *motor* recording fusion models discussed by Teissier et al. (1999) also belong to this category, as they seek a data-to-data mapping of either the visual features into the audio space, or of both modality features to a new common space, followed by linear combination of the resulting features. Audio feature enhancement on basis of either visual input (Girin et al., 1995; Barker and Berthommier, 1999), or concatenated audio-visual features (Girin et al., 2001b; Goecke et al., 2002) falls also within this category of fusion, under its general definition adopted above. In this section, we expand on three feature fusion techniques, schematically depicted in Figure 8.

Concatenative Feature Fusion

Given time-synchronous audio and visual feature vectors $\mathbf{o}_t^{(A)}$ and $\mathbf{o}_t^{(V)}$, with dimensionalities D_A and D_V , respectively, the joint, concatenated audio-visual feature vector at time t becomes

$$\mathbf{o}_t^{(AV)} = [\mathbf{o}_t^{(A)\top}, \mathbf{o}_t^{(V)\top}]^\top \in \mathbb{R}^D, \quad (10)$$

where $D = D_A + D_V$. As with all feature fusion methods (i.e., also for vectors (11) and (12), below), the generation process of a sequence of features (10) is modeled by a single-stream HMM, with emission probabilities (see also (6))

$$Pr[\mathbf{o}_t^{(AV)} | c] = \sum_{k=1}^{K_c} w_{ck} \mathcal{N}_D(\mathbf{o}_t^{(AV)}; \mathbf{m}_{ck}, \mathbf{s}_{ck}),$$

for all classes $c \in \mathcal{C}$ (Adjoudani and Benoît, 1996). Concatenative feature fusion constitutes a simple approach for audio-visual ASR, implementable in most existing ASR systems with minor changes. However, the dimensionality of (10) can be rather high, causing inadequate modeling in (6) due to the curse of dimensionality (Chatfield and Collins, 1991). The following fusion technique aims to avoid this, by seeking lower dimensional representations of (10).

Hierarchical Discriminant Feature Fusion

The visual features contain less speech classification power than audio features, even in the case of extreme noise in the audio channel (see Table 4, in the Experiments section). One would therefore expect that an appropriate lower-dimensional representation of (10) could lead to equal and possibly better HMM performance, given the problem of accurate probabilistic modeling in high-dimensional spaces. Potamianos et al. (2001a) have considered LDA as a means of obtaining such a dimensionality reduction. Indeed, the goal being to obtain the best discrimination among the classes of interest, LDA achieves this on basis of the data (and their labels) alone, without a-priori bias in favor of any of the two feature streams. LDA is subsequently followed by an MLLT based data rotation (see also Figure 8), in order to improve maximum-likelihood data modeling using (6). In the audio-visual ASR system of Potamianos et al. (2001a), the proposed method amounts to a two-stage application of LDA and MLLT, first intra-modal on the original audio MFCC and visual DCT features, and then inter-modal on (10), as also depicted in Figure 11. It is therefore referred to as HiLDA (hierarchical LDA). The final audio-visual feature vector is (see also (10))

$$\mathbf{o}_t^{(\text{HiLDA})} = \mathbf{P}_{\text{MLLT}}^{(AV)} \mathbf{P}_{\text{LDA}}^{(AV)} \mathbf{o}_t^{(AV)}. \quad (11)$$

One can set the dimensionality of (11) to be equal to the audio feature vector size, as implemented by Neti et al. (2000).

Audio Feature Enhancement

Audio and visible speech are correlated, since they are produced by the same oral-facial cavity. Not surprisingly, a number of techniques have been proposed to obtain estimates of audio features utilizing the visual-only modality (Girin et al., 1995; Yehia et al., 1998; Barker and Berthommier, 1999), or joint audio-visual speech data, in the case where the audio signal is degraded (Girin et al., 2001b; Goecke et al., 2002). The latter scenario corresponds to the speech *enhancement* paradigm. Under this approach, the enhanced audio feature vector $\mathbf{o}_t^{(\text{AEnh})}$ can be simply obtained as a *linear* transformation of the concatenated audio-visual feature vector (10), namely as

$$\mathbf{o}_t^{(\text{AEnh})} = \mathbf{P}_{\text{ENH}}^{(AV)} \mathbf{o}_t^{(AV)}, \quad (12)$$

where matrix $\mathbf{P}_{\text{ENH}}^{(AV)} = [\mathbf{p}_1^{(AV)}, \mathbf{p}_2^{(AV)}, \dots, \mathbf{p}_{D_A}^{(AV)}]^\top$ consists of D -dimensional row vectors $\mathbf{p}_i^{(AV)\top}$, for $i = 1, \dots, D_A$, and has dimension $D_A \times D$ (see also Figure 8).

A simple way to estimate matrix $\mathbf{P}_{\text{ENH}}^{(\text{AV})}$ is by considering the approximation $\mathbf{o}_t^{(\text{AEnh})} \approx \mathbf{o}_t^{(\text{AClean})}$ in the *Euclidean* distance sense, where vector $\mathbf{o}_t^{(\text{AClean})}$ denotes clean audio features available in addition to visual and “noisy” audio vectors, for a number of time instants t in a training set, \mathcal{T} . Due to (12), this becomes equivalent to solving D_A *mean square error* (MSE) estimations

$$\mathbf{p}_i^{(\text{AV})} = \arg \min_{\mathbf{p}} \sum_{t \in \mathcal{T}} [\mathbf{o}_{t,i}^{(\text{AClean})} - \mathbf{p}^T \mathbf{o}_t^{(\text{AV})}]^2, \quad (13)$$

for $i = 1, \dots, D_A$, i.e., one per row of the matrix $\mathbf{P}_{\text{ENH}}^{(\text{AV})}$. Equations (13) result to D_A systems of Yule-Walker equations, that can be easily solved using Gauss-Jordan elimination (Press et al., 1995). A more sophisticated way of estimating $\mathbf{P}_{\text{ENH}}^{(\text{AV})}$ by using a Mahalanobis type distance instead of (13) is considered by Goecke et al. (2002), whereas non-linear estimation schemes are proposed by Girin et al. (2001b) and Deligne et al. (2002).

Decision Fusion Techniques for Audio-Visual ASR

Although feature fusion techniques (for example, HiLDA) have been documented to result in improved ASR over audio-only performance (Neti et al., 2000), they cannot explicitly model the reliability of each modality. Such modeling is extremely important, as speech information content and discrimination power of the audio and visual streams can vary widely, depending on the spoken utterance, acoustic noise in the environment, visual channel degradations, face tracker inaccuracies, and speaker characteristics. In contrast to feature fusion methods, the decision fusion framework provides a mechanism for capturing the reliability of each modality, by borrowing from classifier combination literature.

Classifier combination based on their individual decisions about the classes of interest is an active area of research with many applications (Xu et al., 1992; Kittler et al., 1998; Jain et al., 2000). Combination strategies differ in various aspects, such as the architecture used (parallel, cascade, or hierarchical combination), possible trainability (static, or adaptive), and information level considered at integration (abstract, rank-order, or measurement level), i.e., whether information is available about the best class only, the top n classes (or the ranking of all possible ones), or the scores (likelihoods) of them. In the audio-visual ASR literature, examples of most of these categories can be found. For example, Petajan (1984) rescores the two best outputs of the audio-only classifier by means of the visual-only classifier, a case of cascade, static, rank-order level decision fusion. Combinations of more than one categories, as well as cases where the one of the two classifiers of interest corresponds to a feature fusion technique are also possible. For example, Rogozan and Deléglise (1998) use a parallel, adaptive, measurement-level combination of an audio-visual classifier trained on concatenated features (10) with a visual-only classifier, whereas Rogozan (1999) considers a cascade, adaptive, rank-order level integration of the two. The lattice rescoring framework used during the Johns Hopkins University workshop (as described in the Experiments section that follows) is an example of a hybrid cascade/parallel fusion architecture (Neti et al., 2000; Glotin et al., 2001; Luettin et al., 2001).

By far however, the most commonly used decision fusion techniques for audio-visual ASR belong to the paradigm of audio- and visual-only classifier integration using a parallel architecture, adaptive combination weights, and class measurement level information. These methods derive the most likely speech class or word sequence by linearly combining the log-likelihoods of the two single-modality HMM classifier decisions, using appropriate weights (Adjoudani and Benoît, 1996; Jourlin, 1997; Potamianos and Graf, 1998; Teissier et al., 1999; Dupont and Luettin, 2000; Neti et al., 2000; Gurbuz et al., 2001; Heckmann et al., 2001). This corresponds to the adaptive product rule in the likelihood domain (Jain et al., 2000), and it is also known as the *separate identification* model to audio-visual fusion (Rogozan, 1999; Teissier et al., 1999).

Continuous speech recognition introduces an additional twist to the classifier fusion problem, due to the fact that sequences of classes (HMM states or words) need to be estimated. One can consider three possible *temporal* levels for combining stream (modality) likelihoods, as depicted in Table 1. (a): “Early” integration, i.e., likelihood combination at the HMM state level, which gives rise to the *multi-stream HMM* classifier (Bourlard and Dupont, 1996; Young et al., 1999), and forces synchrony between its two single-modality components; (b): “Late” integration, where typically a number of n -best audio and possibly visual-only recognizer hypotheses are rescored by the log-likelihood combination of the two streams, which allows complete

asynchrony between the two HMMs; and (c): “Intermediate” integration, typically implemented by means of the *product HMM* (Varga and Moore, 1990), or the *coupled HMM* (Brand et al., 1997), which force HMM synchrony at the phone, or word, boundaries. Notice that such terminology is not universally agreed upon, and our reference to early or late integration at the temporal level should not be confused with the feature vs. decision fusion meaning of these terms in other work (Adjoudani and Benoît, 1996).

Early Integration: The State-Synchronous Multi-Stream HMM

In its general form, the class conditional observation likelihood of the multi-stream HMM is the product of the observation likelihoods of its single-stream components, raised to appropriate *stream exponents* that capture the reliability of each modality, or equivalently, the confidence of each single-stream classifier. Such model has been considered in audio-only ASR, where for example, separate streams are used for the energy audio features, MFCC static features, as well as their first and possibly second order derivatives, as in Hernando et al. (1995) and Young et al. (1999), or for band-limited audio features in the multi-band ASR paradigm (Hermansky et al., 1996), as in Boulard and Dupont (1996), Okawa et al. (1999), and Glotin and Berthommier (2000), among others. In the audio-visual domain, the model becomes a two-stream HMM, with one stream devoted to the audio, and another to the visual modality. As such, it has been extensively used in small-vocabulary audio-visual ASR tasks (Jourlin, 1997; Potamianos and Graf, 1998; Dupont and Luettin, 2000; Miyajima et al., 2000; Nakamura et al., 2000). In the system reported by Neti et al. (2000) and Luettin et al. (2001), the method was applied for the first time to the LVCSR domain.

Given the bimodal (audio-visual) observation vector $\mathbf{o}_t^{(AV)}$, the state emission “score” (it no longer represents a probability distribution) of the multi-stream HMM is (see also (6) and (10))

$$Pr[\mathbf{o}_t^{(AV)} | c] = \prod_{s \in \{A, V\}} \left[\sum_{k=1}^{K_{sc}} w_{sck} \mathcal{N}_{D_s}(\mathbf{o}_t^{(s)}; \mathbf{m}_{sck}, \mathbf{s}_{sck}) \right]^{\lambda_{sct}}. \quad (14)$$

Notice that (14) corresponds to a linear combination in the log-likelihood domain. In (14), λ_{sct} denote the stream exponents (weights), that are *non-negative*, and in general, are a function of the modality s , the HMM state $c \in \mathcal{C}$, and locally, the utterance frame (time) t . Such state- and time-dependence can be used to model the speech class and “local” environment-based reliability of each stream. The exponents are often constrained to $\lambda_{Act} + \lambda_{Vct} = 1$, or 2. In most systems, they are set to global, modality-only dependent values, i.e., $\lambda_s \leftarrow \lambda_{sct}$, for all classes $c \in \mathcal{C}$ and time instants t , with the class dependence occasionally being preserved, i.e., $\lambda_{sc} \leftarrow \lambda_{sct}$, for all t . In the latter case, the parameters of the multi-stream HMM are (see also (6), (7), and (14))

$$\bar{\mathbf{a}}_{AV} = [\mathbf{a}_{AV}^\top, \{[\lambda_{Ac}, \lambda_{Vc}]^\top, c \in \mathcal{C}\}^\top]^\top, \quad \text{where } \mathbf{a}_{AV} = [\mathbf{r}^\top, \mathbf{b}_A^\top, \mathbf{b}_V^\top]^\top \quad (15)$$

consists of the HMM transition probabilities \mathbf{r} and the emission probability parameters \mathbf{b}_A and \mathbf{b}_V of its single-stream components.

The parameters of \mathbf{a}_{AV} can be estimated *separately* for each stream component using the EM algorithm, namely (8) for $s \in \{A, V\}$, and subsequently, by setting the joint HMM transition probability vector equal to the audio-one, i.e., $\mathbf{r} = \mathbf{r}_A$, or alternatively, to the product of the transition probabilities of the two HMMs, i.e., $\mathbf{r} = \text{diag}(\mathbf{r}_A \mathbf{r}_V^\top)$ (see also (7)). The latter scheme is referred to in the Experiments section as *AV-MS-Sep*. An obvious drawback of this approach is that the two single-modality HMMs are trained asynchronously (i.e., using different forced alignments), whereas (14) assumes that the HMM stream components are state synchronous. The alternative is to *jointly* estimate parameters \mathbf{a}_{AV} , in order to enforce state synchrony. Due to the linear combination of stream log-likelihoods in (14), the EM algorithm carries on in the multi-stream HMM case with minor changes (Rabiner and Juang, 1993; Young et al., 1999). As a result,

$$\mathbf{a}_{AV}^{(j+1)} = \arg \max_{\mathbf{a}} Q(\bar{\mathbf{a}}_{AV}^{(j)}, \mathbf{a} | \mathbf{O}^{(AV)}), \quad (16)$$

can be used, a scheme referred to as *AV-MS-Joint*. Notice that the two approaches basically differ in the E-step of the EM algorithm.

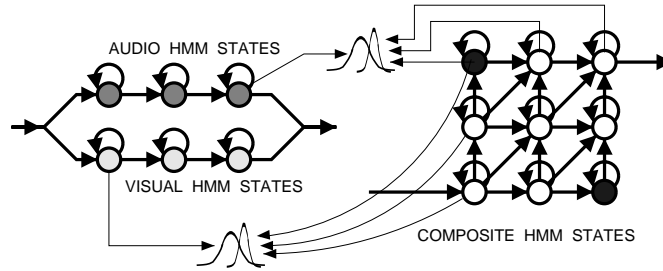


Figure 9: *Left:* Phone-synchronous (state-asynchronous) multi-stream HMM with three states per phone and modality. *Right:* Its equivalent product (composite) HMM; black circles denote states that are removed when limiting the degree of within-phone allowed asynchrony to one state. The single-stream emission probabilities are tied for states along the same row (column) to the corresponding audio (visual) state ones.

In both separate and joint HMM training, the remainder of parameter vector $\bar{\mathbf{a}}_{AV}$, consisting of the stream exponents, needs to be obtained. Maximum likelihood estimation cannot be used for such parameters, and discriminative training techniques have to be employed instead (Jourlin, 1997; Potamianos and Graf, 1998; Nakamura, 2001; Gravier et al., 2002a). The issue is discussed later. Notice that HMM stream parameter and stream exponent training iterations can be alternated in (16).

Intermediate Integration: The Product HMM

It is well known that visual speech activity precedes the audio signal by as much as 120 ms (Bregler and Konig, 1994; Grant and Greenberg, 2001), which is close to the average duration of a phoneme. A generalization of the state-synchronous multi-stream HMM can be used to model such audio and visual stream asynchrony to some extent, by allowing the single modality HMMs to be in asynchrony within a model, but forcing their synchrony at model boundaries instead. Single-stream log-likelihoods are linearly combined at such boundaries using weights, similarly to (14). For LVCSR, a reasonable choice for forcing synchrony constitute the phone boundaries. The resulting phone-synchronous audio-visual HMM is depicted in Figure 9, for the typical case of three states used per phone and modality.

Recognition based on this intermediate integration method requires the computation of the best state sequences for both audio and visual streams. To simplify decoding, the model can be formulated as a *product HMM* (Varga and Moore, 1990). Such model consists of *composite* states $\mathbf{c} \in \mathcal{C} \times \mathcal{C}$, that have audio-visual emission probabilities of a form similar to (14), namely

$$Pr[\mathbf{o}_t^{(AV)} | \mathbf{c}] = \prod_{s \in \{A, V\}} \left[\sum_{k=1}^{K_{s c_s}} w_{s c_s k} \mathcal{N}_{D_s}(\mathbf{o}_t^{(s)}; \mathbf{m}_{s c_s k}, \mathbf{s}_{s c_s k}) \right]^{\lambda_{s c_s t}}, \quad (17)$$

where $\mathbf{c} = [c_A, c_V]^T$. Notice that in (17), the audio and visual stream components correspond to the emission probabilities of certain audio and visual-only HMM states, as depicted in Figure 9. These single-stream emission probabilities are tied for states along the same row, or column (depending on the modality), therefore the original number of mixture weight, mean, and variance parameters is kept in the new model. However, this is usually not the case with the number of transition probability parameters $\{Pr[\mathbf{c}' | \mathbf{c}''], \mathbf{c}', \mathbf{c}'' \in \mathcal{C} \times \mathcal{C}\}$, as additional transitions between the composite states need to be modeled. Such probabilities are often factored as $Pr[\mathbf{c}' | \mathbf{c}''] = Pr[c'_A | c''_A] Pr[c'_V | c''_V]$, in which case the resulting product HMM is typically referred to in the literature as the *coupled HMM* (Brand et al., 1997; Chu and Huang, 2000, 2002; Nefian et al., 2002). A further simplification of this factorization is sometimes employed, namely $Pr[\mathbf{c}' | \mathbf{c}''] = Pr[c'_A | c''_A] Pr[c'_V | c''_V]$, as in Gravier et al. (2002b) for example, which results in a product HMM with the same number of parameters as the state synchronous multi-stream HMM.

Given audio-visual training data, product HMM training can be performed similarly to separate, or joint, multi-stream HMM parameter estimation, discussed in the previous subsection. In the first case, the composite model is constructed based on individual single-modality HMMs estimated by (8), and on transition

probabilities equal to the product of the audio- and visual-only ones. In the second case, referred to as *AV-MS-PROD* in the experiments reported later, all transition probabilities and HMM stream component parameters are estimated at a single stage using (16) with appropriate parameter tying. In both schemes, stream exponents need to be estimated separately. In the audio-visual ASR literature, product (or, coupled) HMMs have been considered in some small-vocabulary recognition tasks (Tomlinson et al., 1996; Dupont and Luettin, 2000; Huang and Chen, 2001; Nakamura, 2001; Chu and Huang, 2002; Nefian et al., 2002), where synchronization is sometimes enforced at the word level, and recently for LVCSR (Neti et al., 2000; Luettin et al., 2001; Gravier et al., 2002b).

It is worth mentioning, that the product HMM allows the restriction of the degree of asynchrony between the two streams, by excluding certain composite states in the model topology. In the extreme case, when only the states that lie in its “diagonal” are kept, the model becomes equivalent to the state-synchronous multi-stream HMM (see also Figure 9).

Late Integration: Discriminative Model Combination

A popular stage of combining audio- and visual-only recognition log-likelihoods is at the utterance end, giving rise to late intergration. In small-vocabulary, isolated word speech recognition, this can be easily implemented by calculating the combined likelihood for each word model in the vocabulary, given the acoustic and visual observations (Adjoudani and Benoît, 1996; Su and Silsbee, 1996; Cox et al., 1997; Gurbuz et al., 2001). However, for connected word recognition, and even more so for LVCSR, the number of possible hypotheses of word sequences becomes prohibitively large. Instead, one has to limit the log-likelihood combination to the top n -best only hypotheses. Such hypotheses can be generated by the audio-only HMM, an alternative audio-visual fusion technique, or can be the union of audio-only and visual-only n -best lists. In this approach, the list of n -best hypotheses for a particular utterance, $\{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$, are first forced-aligned to their corresponding phone sequences $\mathbf{h}_i = \{c_{i,1}, c_{i,2}, \dots, c_{i,N_i}\}$ by means of both audio- and visual-only HMMs. Let the resulting phone $c_{i,j}$ boundaries be denoted by $[t_{i,j,s}^{\text{start}}, t_{i,j,s}^{\text{end}}]$, for $s \in \{A, V\}$, $j = 1, \dots, N_i$, and $i = 1, \dots, n$. Then, the audio-visual likelihoods of the n -best hypotheses are computed as

$$Pr[\mathbf{h}_i] \sim Pr_{\text{LM}}(\mathbf{h}_i)^{\lambda_{\text{LM}}} \prod_{s \in \{A, V\}} \prod_{j=1}^{N_i} Pr(\mathbf{o}_t^{(s)}, t \in [t_{i,j,s}^{\text{start}}, t_{i,j,s}^{\text{end}}] | c_{i,j})^{\lambda_{sc_{i,j}}}, \quad (18)$$

where $Pr_{\text{LM}}(\mathbf{h}_i)$ denotes the *language model* (LM) probability of hypothesis \mathbf{h}_i . The exponents in (18) can be estimated using discriminative training criteria, as in the *discriminative model combination* method of Beyerlein (1998) and Vergyri (2000). The method is proposed for audio-visual LVCSR in Neti et al. (2000), and it is referred to as *AV-DMC* in the Experiments section.

Stream Exponent Estimation and Reliability Modeling

We now address the issue of estimating stream exponents (weights), when combining likelihoods in the audio-visual decision fusion techniques presented above (see (14), (17), and (18)). As already discussed, such exponents can be set to constant values, computed for a particular audio-visual environment and database. In this case, the audio-visual weights depend on the modality and possibly on the speech class, capturing the confidence of the individual classifiers for the particular database conditions, and are estimated by seeking optimal system performance on matched data. However, in a practical audio-visual ASR system, the quality of captured audio and visual data, and thus the speech information present in them, can change dramatically over time. To model this variability, utterance level or even frame level dependence of the stream exponents is required. This can be achieved by first obtaining an estimate of the local environment conditions and then using pre-computed exponents for this condition, or alternatively, by seeking a direct functional mapping between “environment” estimates and stream exponents. In the following, we expand on these methodologies.

In the first approach, constant exponents are estimated based on training data, or more often, on held-out data. Such stream exponents cannot be obtained by maximum likelihood estimation (Potamianos and Graf, 1998; Nakamura, 2001). Instead, discriminative training techniques have to be used. Some of these methods

seek to minimize a smooth function of the *minimum classification error* (MCE) of the resulting audio-visual model on the data, and employ the *generalized probabilistic descent* (GPD) algorithm (Chou et al., 1994) for stream exponent estimation (Potamianos and Graf, 1998; Miyajima et al., 2000; Nakamura et al. 2000; Gravier et al., 2002a). Other techniques use *maximum mutual information* (MMI) training (Bahl et al., 1986), such as the system reported by Jourlin (1997). Alternatively, one can seek to directly minimize the word error rate of the resulting audio-visual ASR system on a held-out data set. In the case of global exponents across all speech classes, constrained to add to a constant, the problem reduces to one-dimensional optimization of a non-smooth function, and can be solved using simple grid search (Miyajima et al., 2000; Luetttin et al., 2001; Gravier et al., 2002a). For class dependent weights, the problem becomes of higher dimension, and the downhill simplex method (Nelder and Mead, 1965) can be employed. This technique is used by Neti et al. (2000) to estimate exponents for late decision fusion using (18). A different approach is to minimize frame misclassification rate, by using the *maximum entropy* criterion (Gravier et al., 2002a).

In order to capture the effects of varying audio and visual environment conditions to the reliability of each stream, utterance-level, and occasionally frame-level, dependence of the stream weights needs to be considered. In most cases in the literature, exponents are considered as a function of the audio channel *signal-to-noise ratio* (SNR), and each utterance is decoded based on the fusion model parameters at its SNR (Adjoudani and Benoît, 1996; Meier et al., 1996; Cox et al., 1997; Teissier et al., 1999; Gurbuz et al., 2001). This SNR value is either assumed known, or estimated from the audio channel (Cox et al., 1997). A linear dependence between SNR and audio stream weight has been demonstrated by Meier et al. (1996). An alternative technique sets the stream exponents to a linear function of the average conditional *entropy* of the recognizer output, computed using the confusion matrix at a particular SNR for a small-vocabulary isolated word ASR task (Cox et al., 1997). A different approach considers the audio stream exponent as a function of the degree of *voicing* present in the audio channel, estimated as in Berthommier and Glotin (1999). The method was used at the Johns Hopkins summer 2000 workshop (Neti et al., 2000; Glotin et al., 2001), and is referred to in the Experiments section as *AV-MS-UTTER*.

The above techniques do not allow modeling of possible variations in the visual stream reliability, since they concentrate on the audio stream alone. Modeling such variability in the visual signal domain is challenging, and instead it can be achieved using confidence measures of the resulting visual-only classifier. For example, Adjoudani and Benoît (1996) and Rogozan et al. (1997) use the *dispersion* of both audio-only and visual-only class posterior log-likelihoods to model the single-stream classifier confidences, and then compute the utterance-dependent stream exponents as a closed form function of these dispersions. Similarly, Potamianos and Neti (2000) consider various confidence measures, such as entropy and dispersion, to capture the reliability of audio- and visual-only classification at the frame level, and estimate stream exponents on basis of held-out data. Such exponents are held constant within confidence value intervals.

Audio-Visual Speaker Adaptation

Speaker adaptation is traditionally used in practical audio-only ASR systems to improve speaker-independent system performance, when little data from a speaker of interest are available (Gauvain and Lee, 1994; Leggetter and Woodland, 1995; Neumeyer et al., 1995; Anastasakos et al., 1997; Gales, 1999). Adaptation is also of interest across tasks or environments. In the audio-visual ASR domain, adaptation is of great importance, since audio-visual corpora are scarce and their collection expensive.

Given few bimodal adaptation data from a particular speaker, and a baseline speaker-independent HMM, one wishes to estimate adapted HMM parameters that better model the audio-visual observations of the particular speaker. Two popular algorithms for speaker adaptation are *maximum likelihood linear regression* (MLLR - Leggetter and Woodland, 1995) and *maximum-a-posteriori* (MAP) adaptation (Gauvain and Lee, 1994). MLLR obtains a maximum likelihood estimate of a linear transformation of the HMM means, while leaving covariance matrices, mixture weights, and transition probabilities unchanged, and it provides successful adaptation with a small amount of adaptation data (rapid adaptation). On the other hand, MAP follows the Bayesian paradigm for estimating the HMM parameters. MAP estimates of HMM parameters slowly converge to their EM-obtained estimates as the amount of adaptation data becomes large, however such a convergence is slow, and therefore, MAP is not suitable for rapid adaptation. In practice, MAP is often used

in conjunction with MLLR (Neumeyer et al., 1995). Both techniques can be used in feature fusion (Potamianos and Neti, 2001a) and decision fusion models discussed above (Potamianos and Potamianos, 1999), in a straightforward manner. One can also consider feature level (front end) adaptation, by adapting, for example, the audio-only and visual-only LDA and MLLT matrices, and in case HiLDA fusion is used, the joint audio-visual LDA and MLLT matrices (Potamianos and Neti, 2001a). Experiments using these techniques are reported in a later section. Alternative adaptation algorithms also exist, such as speaker adaptive training (Anastasakos et al., 1997) and front end MLLR (Gales, 1999), and can be used in audio-visual ASR (Vanegas et al., 1998).

Summary on Audio-Visual Integration

We have presented a summary of the most common fusion techniques for audio-visual ASR. We first discussed the choice of speech classes and statistical ASR models that influence the design of some fusion algorithms. Subsequently, we described a number of feature and decision integration techniques suitable for bimodal LVCSR, and finally, briefly touched upon the issue of audio-visual speaker adaptation.

Among the fusion algorithms discussed, decision fusion techniques explicitly model the reliability of each source of speech information, by using stream weights to linearly combine audio- and visual-only classifier log-likelihoods. When properly estimated, the use of weights results in improved ASR over feature fusion techniques, as reported in the literature and demonstrated in the Experiments section (Potamianos and Graf, 1998; Neti et al., 2000; Luettin et al., 2001). In most systems reported, such weights are set to a constant value over each modality, possibly dependent on the audio-only channel quality (SNR). However, robust estimation of the weights at a finer level (utterance, or frame level) on basis of both audio and visual channel characteristics has not been sufficiently addressed. Furthermore, the issue of whether speech class dependence of stream weights is desirable, has also not been fully investigated. Although such dependence seems to help in late integration schemes (Neti et al., 2000), or small-vocabulary tasks (Jourlin, 1997; Miyajima et al., 2000), the problem remains unresolved for early integration in LVCSR (Gravier et al., 2002a).

There are additional open questions relevant to decision fusion: The first concerns the stage of measurement level information integration, i.e., the degree of allowed asynchrony between the audio and visual streams. The second has to do with the functional form of stream log-likelihood combination, as integration by means of (14) is not necessarily optimal, and it fails to yield an emission probability distribution. Finally, it is worth mentioning a theoretical shortcoming of the log-likelihood linear combination model used in the decision fusion algorithms considered. In contrast to feature fusion, such combination assumes class conditional independence of the audio and visual stream observations. This appears to be a non-realistic assumption (Yehia et al., 1998). A number of models are being investigated to overcome this drawback (Pavlovic, 1998; Pan et al., 1998).

AUDIO-VISUAL DATABASES

A major contributor to the progress achieved in traditional, audio-only ASR has been the availability of a wide variety of large, multi-subject databases on a number of well-defined recognition tasks of different complexities. These corpora have often been collected using funding from U.S. government agencies (for example, the Defense Advanced Research Projects Agency and the National Science Foundation), or through well-organized European activities, such as the Information System Technology program funded by the European Commission, or the European Language Resources Association. The resulting databases are available to the interested research groups by the Linguistic Data Consortium (LDC), or the European Language resources Distribution Agency (ELDA), for example. Benchmarking research progress in audio-only ASR has been possible on such common databases.

In contrast to the abundance of audio-only corpora, there exist only few databases suitable for audio-visual ASR research. This is because the field is relatively young, but also due to the fact that audio-visual databases pose additional challenges concerning database collection, storage, and distribution, not found in the audio-

only domain. For example, computer acquisition of visual data at full size, frame rate, and high image quality, synchronous to the audio input, requires expensive hardware, whereas even highly compressed visual data storage consumes at least an order of magnitude more storage space than audio, making widespread database distribution a non-trivial task. Although solutions have been steadily improving and becoming available at a lower cost, these issues have seriously hindered availability of large audio-visual corpora. Additional difficulties stem from the proprietary nature of some collected corpora, as well as privacy issues due to the inclusion of the visual modality.

Most existing audio-visual databases are the result of efforts by few university groups or individual researchers with limited resources. Therefore, most of these corpora suffer from one or more shortcomings (Chibelushi et al., 1996, 2002; Hennecke et al., 1996): They contain a single or small number of subjects, affecting the generalizability of developed methods to the wider population; they typically have small duration, often resulting in undertrained statistical models, or non-significant performance differences between various proposed algorithms; and finally, they mostly address simple recognition tasks, such as small-vocabulary ASR of isolated or connected words. These limitations have caused a growing gap in the state-of-the-art between audio-only and audio-visual ASR in terms of recognition task complexity. To help bridge this gap, we have recently completed the collection of a large corpus suitable for audio-visual LVCSR, which we used for experiments during the Johns Hopkins summer 2000 workshop (Neti et al., 2000). Some of these experiments are summarized in the following section.

In the remainder of this section, we give an overview of the most commonly used audio-visual databases in the literature. Some of these sets have been used by multiple sites and researchers, allowing some algorithm comparisons. However, benchmarking on common corpora is not widespread. Subsequently, we describe the IBM ViaVoiceTM audio-visual database, and additional corpora used in the experiments reported in the next section.

Overview of Small- and Medium-Vocabulary Audio-Visual Corpora

The first database used for automatic recognition of audio-visual speech was collected by Petajan (1984). Data of a single subject uttering 2-10 repetitions of 100 isolated English words, including letters and digits, were collected under controlled lighting conditions. Since then, several research sites have pursued audio-visual data collection. Some of the resulting corpora are discussed in the following.

A number of databases are designed to study audio-visual recognition of consonants (C), vowels (V), or transitions between them. For example, Adjoudani and Benoît (1996) report a single-speaker corpus of 54 /V₁CV₂CV₁/ non-sense words (three French vowels and six consonants are considered). Su and Silsbee (1996) recorded a single-speaker corpus of /aCa/ non-sense words for recognition of 22 English consonants. Robert-Ribes et al. (1998), as well as Teissier et al. (1999) report recognition of ten French oral vowels uttered by a single subject. Czap (2000) considers a single-subject corpus of /V₁CV₁/ and /C₁VC₁/ non-sense words for recognition of Hungarian vowels and consonants.

The most popular task for audio-visual ASR is isolated or connected digit recognition. Various corpora allow digit recognition experiments. For example, the Tulips1 database (Movellan and Chadderdon, 1996) contains recordings of 12 subjects uttering digits “one” to “four”, and has been used for isolated recognition of these four digits in a number of papers (Luettin et al., 1996; Movellan and Chadderdon, 1996; Gray et al., 1997; Vanegas et al., 1998; Scanlon and Reilly, 2001). The M2VTS database, although tailored to speaker verification applications, also contains digit (“0” to “9”) recordings of 37 subjects, mostly in French (Pigeon and Vandendorpe, 1997), and it has been used for isolated digit recognition experiments (Dupont and Luettin, 2000; Miyajima et al., 2000); XM2VTS, an extended version of this database containing 295 subjects has recently been completed in the English language (Messer et al., 1999). Additional single-subject digit databases include the NATO RSG10 digit-triples set, used by Tomlinson et al. (1996) for isolated digit recognition, and two connected-digits databases reported by Potamianos et al. (1998), and Heckmann et al. (2001). Finally, two very recent databases, suitable for multi-subject connected digit recognition, have been collected at the University of Illinois at Urbana-Champaign (a 100-subject set), with results reported in Chu and Huang (2000) and Zhang et al. (2000), and at Clemson University (the 36-subject CUAVE dataset),



Figure 10: Example video frames of ten subjects from the IBM ViaVoiceTM audio-visual database. The database contains approximately 50 hrs of continuous, dictation-style audio-visual speech by 290 subjects, collected with minor face pose, lighting, and background variation (Neti et al., 2000).

as discussed in Patterson et al. (2002).

Isolated or connected letter recognition constitutes another popular audio-visual ASR task. German connected letter recognition on data of up to six subjects has been reported by Bregler et al. (1993), Bregler and König (1994), Duchnowski et al. (1994), and Meier et al. (1996), whereas Krone et al. (1997) work on single-speaker isolated German letter recognition. Single-, or two-subject, connected French letter recognition is considered in Alissali et al. (1996), André-Obrecht et al. (1997), Jourlin (1997), Rogozan et al. (1997), and Rogozan (1999). Finally, for English, a 10-subject isolated letter dataset is used by Matthews et al. (1996) and Cox et al. (1997), whereas a 49-subject connected letter database by Potamianos et al. (1998).

In addition to letter or digit recognition, a number of audio-visual databases have been collected that are suitable for recognition of isolated words. For example, Silsbee and Bovik (1996) have collected a single-subject, isolated word corpus with a vocabulary of 500 words. Recognition of single-subject command words for radio/tape control has been used by Chiou and Hwang (1997), as well as by Gurbuz et al. (2001), and Patterson et al. (2001). A 10-subject isolated word database with a vocabulary size of 78 words is considered by Chen (2001) and Huang and Chen (2001). This corpus has been collected at Carnegie Mellon University (AMP/CMU database), and has also been used by Chu and Huang (2002), Nefian et al. (2002), and Zhang et al. (2002), among others. Single-subject, isolated word recognition in Japanese is reported in Nakamura et al. (2000) and Nakamura (2001), whereas a single-subject German command word recognition is considered by Kober et al. (1997).

Finally, few audio-visual databases are suitable for continuous speech recognition in limited, small-vocabulary domains: Goldschen et al. (1996) have collected single-subject data uttering 150 TIMIT sentences three times. Chan et al. (1998) collected 400 single-subject military command and control utterances. An extended multi-subject version of this database (still with a limited vocabulary of 101 words) is reported in Chu and Huang (2000).

The IBM ViaVoiceTM Audio-Visual Database

To date, the largest audio-visual database collected, and the only one suitable for speaker-independent LVCSR, is the IBM ViaVoiceTM audio-visual database. The corpus consists of full-face frontal video and audio of 290 subjects (see also Figure 10), uttering ViaVoiceTM training scripts, i.e., continuous read speech with mostly verbalized punctuation, dictation style. The database video is of a 704×480 pixel size, interlaced, captured in color at a rate of 30 Hz (i.e., 60 fields per second are available at a resolution of 240 lines), and it is MPEG2 encoded at the relatively high compression ratio of about 50:1. High quality wideband audio is synchronously collected with the video at a rate of 16 kHz and at a relatively clean audio environment (quiet office, with some background computer noise), resulting in a 19.5 dB SNR. The duration of the entire database is approximately 50 hours, and it contains 24,325 transcribed utterances with a 10,403-word vocabulary, from which 21,281 utterances are used in the experiments reported in the next section. In addition to

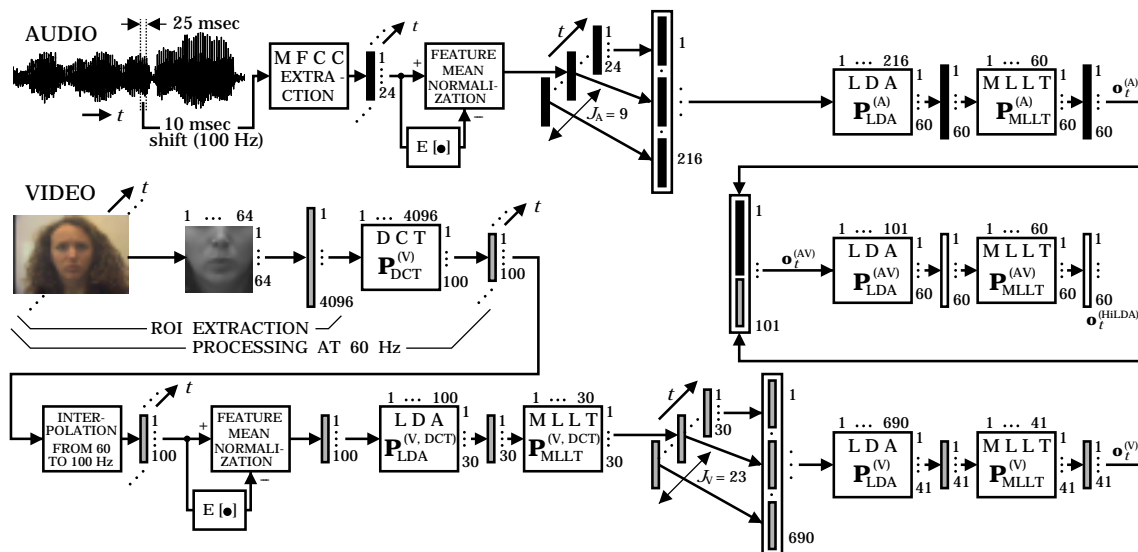


Figure 11: The audio-visual ASR system employed in some of the experiments reported in this chapter. In addition to the baseline system used during the Johns Hopkins summer 2000 workshop, a larger mouth ROI is extracted, within-frame discriminant features are used, and a longer temporal window is considered in the visual front end (compare to Figure 7). HiLDA feature fusion is employed.

LVCSR, a 50-subject connected digit database has been collected at IBM, in order to study the visual modality benefit to a popular small-vocabulary ASR task. This DIGITS corpus contains 6689 utterances of 7- and 10-digit strings (both “zero” and “oh” are used) with a total duration of approximately 10 hrs. Furthermore, to allow investigation of automatic speechreading performance for impaired speech (Potamianos and Neti, 2001a), both LVCSR and DIGITS audio-visual speech data of a single speech impaired male subject with profound hearing loss have been collected. In Table 3, a summary of the above corpora is given, together with their partitioning used in the experiments reported in the following section.

AUDIO-VISUAL ASR EXPERIMENTS

In this section, we present experimental results on visual-only and audio-visual ASR using mainly the IBM ViaVoice™ database, discussed above. Some of these results have been obtained during the Johns Hopkins summer 2000 workshop (Neti et al., 2000). Experiments conducted later on both these data, as well as on the IBM connected digits task (DIGITS) are also reported (Potamianos et al., 2001a; Goecke et al., 2002; Gravier et al., 2002a). In addition, the application of audio-visual speaker adaptation methods on the hearing impaired dataset is also discussed (Potamianos and Neti, 2001a). First, however, we briefly describe the basic audio-visual ASR system, as well as the experimental framework used.

The Audio-Visual ASR System

Our basic audio-visual ASR system utilizes appearance based visual features, that use a discrete cosine transform (DCT) of the mouth region-of-interest (ROI), as described in Potamianos et al. (2001b). Given the video of the speaker’s face, available at 60 Hz, it first performs face detection and mouth center and size estimation employing the algorithm of Senior (1999), and on basis of these, it extracts a size-normalized, 64×64 greyscale pixel mouth ROI, as discussed in a previous section (see also Figure 2). Subsequently, a two-dimensional, separable, fast DCT is applied on the ROI, and its 24 highest energy coefficients (over the training data) are retained. A number of post-processing steps are applied on the resulting “static” feature

Speech condition	Recognition task	Training set			Held-out set			Adaptation set			Test set		
		Utter.	Dur.	Sub.	Utter.	Dur.	Sub.	Utter.	Dur.	Sub.	Utter.	Dur.	Sub.
Normal	LVCSR	17111	34:55	239	2277	4:47	25	855	2:03	26	1038	2:29	26
	DIGITS	5490	8:01	50	670	0:58	50	670	0:58	50	529	0:46	50
Impaired	LVCSR	N / A			N / A			50	0:11	1	50	0:11	1
	DIGITS	N / A			N / A			80	0:08	1	60	0:06	1

Table 3: The IBM audio-visual databases discussed and used in the experiments reported in this chapter. Their partitioning into training, held-out, adaptation, and test sets is depicted (number of utterances, duration (in hours), and number of subjects are shown for each set). Both large-vocabulary continuous speech (LVCSR) and connected digit (DIGITS) recognition are considered for normal, as well as impaired speech. The IBM ViaVoice™ database corresponds to the LVCSR task in the normal speech condition. For the normal speech DIGITS task, the held-out and adaptation sets are identical. For impaired speech, due to the lack of sufficient training data, adaptation of HMMs trained in the normal speech condition is considered.

vector, namely, linear interpolation to the audio feature rate (from 60 to 100 Hz), feature mean normalization (FMN) for improved robustness to lighting and other variations, concatenation of 15 adjacent features to capture dynamic speech information (see also (5)), and linear discriminant analysis (LDA) for optimal dimensionality reduction, followed by a maximum likelihood data rotation (MLLT) for improved statistical data modeling. The resulting feature vector $\mathbf{o}_t^{(V)}$ has dimension 41. These steps are described in more detail in the Visual front end section of this chapter (see also Figure 7). Improvements to this DCT based visual front end have been proposed in Potamianos and Neti (2001b), including the use of a larger ROI, a within-frame discriminant DCT feature selection, and a longer temporal window (see Figure 11). During the Johns Hopkins summer workshop, and in addition to the DCT based features, joint appearance and shape features by means of active appearance models (AAMs) have also been employed. In particular, 6000-dimensional appearance vectors containing the normalized face color pixel values, and 134-dimensional shape vectors of the face shape coordinates are extracted at 30 Hz, and are passed through two stages of principal components analysis (PCA). The resulting “static” AAM feature vector is 86-dimensional, and it is post-processed similarly to the DCT feature vector (see Figure 7), resulting to 41-dimensional “dynamic” features.

In parallel to the visual front end, traditional audio features are extracted at a 100 Hz rate that consist of mel frequency cepstral coefficients (MFCCs) and its energy (Rabiner and Juang, 1993; Deller et al., 1993; Young et al., 1999). The obtained “static” feature vector is 24-dimensional, and following FMN, LDA on 9 adjacent frames, and MLLT, it gives rise to a 60-dimensional dynamic speech vector, $\mathbf{o}_t^{(A)}$, as depicted in Figure 11. The audio and visual front ends provide time-synchronous audio and visual feature vectors that can be used in a number of fusion techniques discussed in a previous section. The derived concatenated audio-visual vector $\mathbf{o}_t^{(AV)}$ has dimension 101, whereas in the HiLDA feature fusion implementation, the bimodal LDA generates features $\mathbf{o}_t^{(HiLDA)}$ with a reduced dimensionality 60 (see also Figure 11).

In all cases where LDA and MLLT matrices are employed (audio-, visual-only, and audio-visual feature extraction by means of HiLDA fusion), we consider $|\mathcal{C}| = 3367$ context-dependent sub-phonetic classes that coincide with the context-dependent states of an available audio-only HMM, that has been previously developed at IBM for LVCSR, trained on a number of audio corpora (Polymenakos et al., 1998). The forced alignment (Rabiner and Juang, 1993) of the training set audio, based on this HMM and the data transcriptions, produces labels $c(l) \in \mathcal{C}$ for the training set audio-, visual-, and audio-visual data vectors \mathbf{x}_l , $l = 1, \dots, L$. Such labeled vectors can then be used to estimate the required matrices \mathbf{P}_{LDA} , \mathbf{P}_{MLLT} , as described in the Visual front end section of this chapter.

The Experimental Framework

The audio-visual databases discussed above have been partitioned into a number of sets in order to train and evaluate models for audio-visual ASR, as detailed in Table 3. For both LVCSR and DIGITS speech tasks in

Modality	Remarks	WER	Modality	Remarks	WER
Visual	DCT	58.1	Acoustic	MFCC (noisy)	55.0
	DWT	58.8	None	Oracle	31.2
	PCA	59.4		Anti-oracle	102.6
	AAM	64.0		LM best path	62.0

Table 4: Comparisons of various visual features (three appearance based features, and one joint shape and appearance feature representation) for speaker-independent LVCSR (Neti et al., 2000; Matthews et al., 2001). Word error rate (WER), %, is depicted on a subset of the IBM ViaVoiceTM database test set of Table 3. Visual performance is obtained after rescoring of lattices, that have been previously generated based on noisy (at 8.5 dB SNR) audio-only MFCC features. For comparison, characteristic lattice WERs are also depicted (oracle, anti-oracle, and best path based on language model scores alone). Among the visual speech representations considered, the DCT based features are superior and contain significant speech information.

the normal speech condition, the corresponding *training* sets are used to obtain all LDA and MLLT matrices required, the phonetic decision trees that cluster HMM states on basis of phonetic context, as well as to train all HMMs reported. The *held-out* sets are used to tune parameters relevant to audio-visual decision fusion and decoding (such as the multi-stream HMM and language model weights, for example), whereas the *test sets* are used for evaluating the performance of the trained HMMs. Optionally, the *adaptation sets* can be employed for tuning the front ends and/or HMMs to the characteristics of the test set subjects. In the LVCSR case, the subject populations of the training, held-out, and test sets are disjoint, thus allowing for *speaker-independent* recognition, whereas in the DIGITS data partitioning, all sets have data from the same 50 subjects, thus allowing *multi-speaker* experiments. Due to this fact, the adaptation and held-out sets for DIGITS are identical. For the impaired speech data, the duration of the collected data is too short to allow HMM training. Therefore, LVCSR HMMs trained on the IBM ViaVoiceTM dataset are adapted on the impaired LVCSR and DIGITS adaptation sets (see Table 3).

To assess the benefit of the visual modality to ASR in noisy conditions (in addition to the relatively clean audio condition of the database recordings), we artificially corrupt the data audio with additive, non-stationary, speech “babble” noise at various SNRs. ASR results are then reported at a number of SNRs, ranging within $[-1.5, 19.5]$ dB for LVCSR and $[-3.5, 19.5]$ dB for DIGITS, with all corresponding front end matrices and HMMs trained in the *matched* condition. In particular, during the Johns Hopkins summer 2000 workshop, only two audio conditions were considered for LVCSR: The original 19.5 dB SNR audio and a degraded one at 8.5 dB SNR. Notice that, in contrast to the audio, no noise is added to the video channel or features. Many cases of “visual noise” could have been considered, such as additive noise on video frames, blurring, frame rate decimation, and extremely high compression factors, among others. Some preliminary studies on the effects of video degradations to visual recognition can be found in the literature (Davoine et al., 1997; Williams et al., 1997; Potamianos et al., 1998). These studies find automatic speechreading performance to be rather robust to video compression for example, but to degrade rapidly for frame rates below 15 Hz.

The ASR experiments reported next follow two distinct paradigms. The results on the IBM ViaVoiceTM data obtained during the Johns Hopkins summer 2000 workshop employ a *lattice rescoring* paradigm, due to the limitations in large-vocabulary decoding of the HTK software used there (Young et al., 1999); namely, lattices were first generated prior to the workshop using the IBM Research decoder (Hark) with HMMs trained at IBM, and subsequently rescored during the workshop, by trained tri-phone context-dependent HMMs on various feature sets or fusion techniques using HTK. Three sets of lattices were generated for these experiments, and were based on clean audio-only (19.5 dB), noisy audio-only, and noisy audio-visual (at the 8.5 dB SNR condition) HiLDA features. In the second experimental paradigm, *full decoding* results obtained by directly using the IBM Research recognizer are reported. For the LVCSR experiments, 11-phone context-dependent HMMs with 2,808 context-dependent states and 47 k Gaussian mixtures are used, whereas for DIGITS recognition in normal speech the corresponding numbers are 159 and 3.2 k (for single-stream models). Decoding using the closed set vocabulary (10,403 words) and a trigram language model is employed for LVCSR (this is the case also for the workshop results), whereas the 11 digit (“zero” to “nine”, including

Audio Condition:	Clean	Noisy	Audio Condition:	Clean	Noisy
Audio-only	14.44	48.10	AV-MS-Joint (DF)	14.62	36.61
AV-Concat (FF)	16.00	40.00	AV-MS-Sep (DF)	14.92	38.38
AV-HiLDA (FF)	13.84	36.99	AV-MS-PROD (DF)	14.19	35.21
AV-DMC (DF)	13.65 → 12.95	—	AV-MS-UTTER (DF)	13.47	35.27

Table 5: Test set speaker-independent LVCSR audio-only and audio-visual WER, %, for the clean (19.5 dB SNR) and a noisy audio (8.5 dB) condition. Two feature fusion (FF) and five decision fusion (DF) based audio-visual systems are evaluated using the lattice rescoring paradigm (Neti et al., 2000; Glotin et al., 2001; Luettin et al., 2001).

“oh”) word vocabulary is used for DIGITS (with unknown digit string length).

Visual-Only Recognition

The suitability for LVCSR of a number of appearance based visual features and AAMs was studied during and after the Johns Hopkins summer workshop (Neti et al., 2000; Matthews et al., 2001). For this purpose, noisy audio-only lattices were rescored by HMMs trained on the various visual features considered, namely 86-dimensional AAM features, as well as 24-dimensional DCT, PCA (on 32×32 pixel mouth ROIs), and DWT based features. All features were post-processed as previously discussed to yield 41-dimensional feature vectors (see Figure 7). For the DWT features, the Daubechies class wavelet filter of approximating order 3 is used (Daubechies, 1992; Press et al., 1995). LVCSR recognition results are reported in Table 4, depicted in *word error rate* (WER), %. The DCT outperformed all other features considered. Notice however that these results cannot be interpreted as visual-only recognition, since they correspond to cascade audio-visual fusion of audio-only ASR, followed by visual-only rescoring of a network of recognized hypotheses. For reference, a number of characteristic lattice WERs are also depicted in Table 4, including the audio-only (at 8.5 dB) result. All feature performances are bounded by the lattice *oracle* and *anti-oracle* WERs. It is interesting to note that all appearance based features considered attain lower WERs (e.g., 58.1% for DCT features) than the WER of the best path through the lattice based on the language model alone (62.0%). Therefore, such visual features do convey significant speech information. AAMs on the other hand did not perform well, possibly due to severe undertraining of the models, resulting in poor fitting to unseen facial data.

As expected, visual-only recognition based on full decoding (instead of lattice rescoring) is rather poor. The LVCSR WER on the speaker-independent test set of Table 3, based on per-speaker MLLR adaptation is reported at 89.2% in Potamianos and Neti (2001b), using the DCT features of the workshop. Extraction of larger ROIs and the use of within frame DCT discriminant features and longer temporal windows (as depicted in Figure 11) result in the improved WER of 82.3%. In contrast to LVCSR, DIGITS visual-only recognition constitutes a much easier task. Indeed, on the multi-speaker test set of Table 3, a 16.8% WER is achieved after per-speaker MLLR adaptation.

Audio-Visual ASR

A number of audio-visual integration algorithms presented in the fusion section of this chapter were compared during the Johns Hopkins summer 2000 workshop. As already mentioned, two audio conditions were considered: The original clean database audio (19.5 dB SNR) and a noisy one at 8.5 dB SNR. In the first case, fusion algorithm results were obtained by rescoring pre-generated clean audio-only lattices; at the second condition, HiLDA noisy audio-visual lattices were rescored. The results of these experiments are summarized in Table 5. Notice that every fusion method considered outperformed audio-only ASR in the noisy case, reaching up to a 27% relative reduction in WER (from 48.10% noisy audio-only to 35.21% audio-visual). In the clean audio condition, among the two feature fusion techniques considered, HiLDA fusion (Potamianos et al., 2001a) improved ASR from 14.44% audio-only to a 13.84% audio-visual WER, however concatenative fusion degraded performance to 16.0%. Among the decision fusion algorithms used, the product HMM (AV-MS-PROD) with jointly trained audio-visual components (Luettin et al., 2001) improved performance

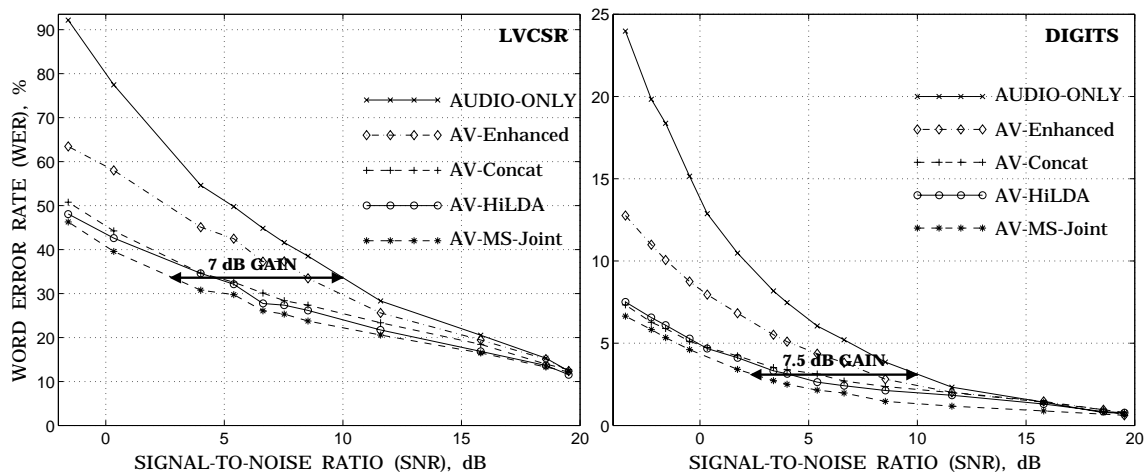


Figure 12: Comparison of audio-only and audio-visual ASR by means of three feature fusion (AV-Concat, AV-HiLDA, and AV-Enhanced) algorithms and one decision fusion (AV-MS-Joint) technique, using the full decoding experimental paradigm. WERs vs. audio channel SNR are reported on both the IBM ViaVoice™ test set (speaker-independent LVCSR - left), as well as on the multi-speaker DIGITS test set (right) of Table 3. HiLDA feature fusion outperforms alternative feature fusion methods, whereas decision fusion outperforms all three feature fusion approaches, resulting in an effective SNR gain of 7 dB for LVCSR and 7.5 dB for DIGITS, at 10 dB SNR (Potamianos et al., 2001a; Goecke et al., 2002; Gravier et al., 2002a). Notice that the WER range in the two graphs differs.

to a 14.19% WER. In addition, utterance-based stream exponents for a jointly trained multi-stream HMM (AV-MS-UTTER), estimated using an average of the voicing present at each utterance, further reduced WER to 13.47% (Glotin et al., 2001), achieving a 7% relative WER reduction over audio-only performance. Finally, a late integration technique based on discriminative model combination (AV-DMC) of audio and visual HMMs (Beyerlein, 1998; Vergyri, 2000; Glotin et al., 2001) produced a WER of 12.95%, amounting to a 5% reduction from its clean audio-only baseline of 13.65% (this differs from the 14.44% audio-only result due to the rescoring of n -best lists instead of lattices). Notice, that for both clean and noisy audio conditions, the best decision fusion method outperformed the best feature fusion technique considered. In addition, for both conditions, joint multi-stream HMM training outperformed separate training of the HMM stream components, something not surprising, since joint training forces state synchrony between the audio and visual streams.

To further demonstrate the differences between the various fusion algorithms and to quantify the visual modality benefit to ASR, we review a number of full decoding experiments recently conducted for both the LVCSR and DIGITS tasks, and at a large number of SNR conditions (Potamianos et al., 2001a; Goecke et al., 2002; Gravier et al., 2002a). All three feature fusion techniques discussed in the relevant section of this chapter are compared to decision fusion by means of a jointly trained multi-stream HMM. The results are depicted in Figure 12. Among the feature fusion methods considered, HiLDA feature fusion is superior to both concatenative fusion and the enhancement approach. In the clean audio case for example, HiLDA fusion reduces the audio-only LVCSR WER of 12.37% to 11.56% audio-visual, whereas feature concatenation degrades performance to 12.72% (the enhancement method obviously provides the original audio-only performance in this case). Notice that these results are somewhat different to the ones reported in Table 5, due to the different experimental paradigm considered. In the most extreme noisy case considered for LVCSR (-1.5 dB SNR), the audio-only WER of 92.16% is reduced to 48.63% using HiLDA, compared to 50.76% when feature concatenation is employed and 63.45% when audio feature enhancement is used. Similar results hold for DIGITS recognition, although the difference between HiLDA and concatenative feature fusion ASR is small, possibly due to the fact that HMMs with significantly less Gaussian mixtures are used, and the availability of sufficient data to train on high dimensional concatenated audio-visual vectors. The comparison

Task →		LVCSR			DIGITS		
↓ Method	Modality →	AU	VI	AV	AU	VI	AV
Unadapted		116.022	136.359	106.014	52.381	48.016	24.801
MLLR		52.044	110.166	42.873	3.770	16.667	0.992
MAP		52.376	101.215	44.199	3.373	12.103	1.190
MAP+MLLR		47.624	95.027	41.216	2.381	10.516	0.992
Mat+MAP		52.928	98.674	46.519	3.968	8.730	1.190
Mat+MAP+MLLR		50.055	93.812	41.657	2.381	8.531	0.992

Table 6: Adaptation results on the speech impaired data. WER, %, of the audio-only (AU), visual-only (VI), and audio-visual (AV) modalities, using HiLDA feature fusion, is reported on both the LVCSR (*left* table part) and DIGITS test sets (*right* table) of the speech impaired data using unadapted HMMs (trained in normal speech), as well as a number of HMM adaptation methods. All HMMs are adapted on the joint speech impaired LVCSR and DIGITS adaptation sets of Table 3. For the continuous speech results, decoding using the test set vocabulary of 537 words is reported. MAP followed by MLLR adaptation, and possibly preceded by front end matrix adaptation (Mat), achieves the best results for all modalities and for both tasks considered (Potamianos and Neti, 2001a).

between multi-stream decision fusion and HiLDA fusion reveals that the jointly trained multi-stream HMM performs significantly better. For example, at -1.5 dB SNR, LVCSR WER is reduced to 46.28% (compared to 48.63% for HiLDA). Similarly, for DIGITS recognition at -3.5 dB, the HiLDA WER is 7.51%, whereas the multi-stream HMM WER is significantly lower, namely 6.64%. This is less than one third of the audio-only WER of 23.97%.

A useful indicator when comparing fusion techniques and establishing the visual modality benefit to ASR is the *effective SNR gain*, measured here with reference to the audio-only WER at 10 dB. To compute this gain, we need to consider the SNR value where the audio-visual WER equals the reference audio-only WER (see Figure 12). For HiLDA fusion, this gain equals approximately 6 dB for both LVCSR and DIGITS tasks. Jointly trained multi-stream HMMs improve these gains to 7 dB for LVCSR and 7.5 dB for DIGITS, at 10 dB SNR. Full decoding experiments employing additional decision fusion techniques are currently in progress. In particular, intermediate fusion results by means of the product HMM are reported in Gravier et al. (2002b).

Audio-Visual Adaptation

We now describe recent experiments on audio-visual adaptation in a case study of single-subject audio-visual ASR of impaired speech (Potamianos and Neti, 2001a). As already indicated, the small amount of speech impaired data collected (see Table 3) is not sufficient for HMM training, thus calling for speaker adaptation techniques instead. A number of such methods, described in a previous section, are used for adapting audio-only, visual-only, and audio-visual HMMs suitable for LVCSR. The results on both speech impaired LVCSR and DIGITS tasks are depicted in Table 6. Notice, that due to poor accuracy on impaired speech, decoding on the LVCSR task is performed using the 537 word test set vocabulary of the dataset. Clearly, the mismatch between the normal and impaired speech data is dramatic, as the “Unadapted” table entries demonstrate. Indeed, the audio-visual WER in the LVCSR task reaches 106.0% (such large numbers occur due to word insertions), whereas the audio-visual WER in the DIGITS task is 24.8% (in comparison, the normal speech, per subject adapted audio-visual LVCSR WER is 10.2%, and the audio-visual DIGITS WER is only 0.55%, computed on the test sets of Table 3).

We first consider MLLR and MAP HMM adaptation using the joint speech impaired LVCSR and DIGITS adaptation tests. Audio-, visual-only, and audio-visual performances improve dramatically, as demonstrated in Table 6. Due to the rather large adaptation set, MAP performs similarly well to MLLR. Applying MLLR after MAP improves results, and it reduces the audio-visual WER to 41.2% and 0.99% for the LVCSR and DIGITS tasks, respectively, amounting to a 61% and 96% relative WER reduction over the audio-visual

unadapted results, and to a 13% and 58% relative WER reduction over the audio-only MAP+MLLR adapted results. Clearly therefore, the visual modality dramatically benefits the automatic recognition of impaired speech. We also apply front end adaptation, possibly followed by MLLR adaptation, with the results depicted in the Mat+MAP(+MLLR) entries of Table 6. Although visual-only recognition improves, the audio-only recognition results fail to do so. As a consequence, audio-visual ASR degrades, possibly also due to the fact that, in this experiment, audio-visual matrix adaptation is only applied to the second stage of LDA/MLLT.

SUMMARY AND DISCUSSION

In this chapter, we provided an overview of the basic techniques for automatic recognition of audio-visual speech, proposed in the literature over the past twenty years. The two main issues relevant to the design of audio-visual ASR systems are: First, the visual front end that captures visual speech information and, second, the integration (fusion) of audio and visual features into the automatic speech recognizer used. Both are challenging problems, and significant research effort has been directed towards finding appropriate solutions.

We first discussed extracting visual features from the video of the speaker's face. The process requires first the detection and tracking of the face, mouth region, and possibly the speaker's lip contours. A number of mostly statistical techniques, suitable for the task were reviewed. Various visual features proposed in the literature were then presented. Some are based on the mouth region appearance and employ image transforms or other dimensionality reduction techniques borrowed from the pattern recognition literature, in order to extract relevant speech information. Others capture the lip contour and possibly face shape characteristics, by means of statistical, or geometric models. Combinations of features from these two categories are also possible.

Subsequently, we concentrated on the problem of audio-visual integration. Possible solutions to it differ in various aspects, including the classifier and classes used for automatic speech recognition, the combination of single-modality features vs. single-modality classification decisions, and in the latter case, the information level provided by each classifier, the temporal level of the integration, and the sequence of such decision combination. We concentrated on HMM based recognition, based on sub-phonetic classes, and, assuming time-synchronous audio and visual feature generation, we reviewed a number of feature and decision fusion techniques. Within the first category, we discussed simple feature concatenation, discriminant feature fusion, and a linear audio feature enhancement approach. For decision based integration, we concentrated in linear log-likelihood combination of parallel, single-modality classifiers at various levels of integration, considering the state-synchronous multi-stream HMM for "early" fusion, the product HMM for "intermediate" fusion, and discriminative model combination for "late" integration, and we discussed training the resulting models.

Developing and benchmarking feature extraction and fusion algorithms requires available audio-visual data. A limited number of corpora suitable for research in audio-visual ASR have been collected and used in the literature. A brief overview of them was also provided, followed by a description of the IBM ViaVoiceTM database, suitable for speaker-independent audio-visual ASR in the large-vocabulary, continuous speech domain. Subsequently, a number of experimental results were reported using this database, as well as additional corpora recently collected at IBM. Some of these experiments were conducted during the summer 2000 workshop at the Johns Hopkins University, and compared both visual feature extraction and audio-visual fusion methods for LVCSR. More recent experiments, as well as a case study of speaker adaptation techniques for audio-visual recognition of impaired speech were also presented. These experiments showed that a visual front end can be designed that successfully captures speaker-independent, large-vocabulary continuous speech information. Such a visual front end uses discrete cosine transform coefficients of the detected mouth region of interest, suitably post-processed. Combining the resulting visual features with traditional acoustic ones results in significant improvements over audio-only recognition in both clean and of course degraded acoustic conditions, across small and large vocabulary tasks, as well as for both normal and impaired speech. A successful combination technique is the multi-stream HMM based decision fusion approach, or the simpler, but inferior, discriminant feature fusion (HiLDA) method.

This chapter clearly demonstrates that, over the past twenty years, much progress has been accomplished in capturing and integrating visual speech information into automatic speech recognition. However, the visual modality has yet to become utilized in mainstream ASR systems. This is due to the fact that issues

of both practical and research nature remain challenging. On the practical side of things, the high quality of captured visual data, which is necessary for extracting visual speech information capable of enhancing ASR performance, introduces increased cost, storage, and computer processing requirements. In addition, the lack of common, large audio-visual corpora that address a wide variety of ASR tasks, conditions, and environments, hinders development of audio-visual systems suitable for use in particular applications.

On the research side, the key issues in the design of audio-visual ASR systems remain open and subject to more investigation. In the visual front end design, for example, face detection, facial feature localization, and face shape tracking, robust to speaker, pose, lighting, and environment variation constitute challenging problems. A comprehensive comparison between face appearance and shape based features for speaker-dependent vs. speaker-independent automatic speechreading is also unavailable. Joint shape and appearance three-dimensional face modeling, used for both tracking and visual feature extraction has not been considered in the literature, although such an approach could possibly lead to the desired robustness and generality of the visual front end. In addition, when combining audio and visual information, a number of issues relevant to decision fusion require further study, such as the optimal level of integrating the audio and visual log-likelihoods, the optimal function for this integration, as well as the inclusion of suitable, local estimates of the reliability of each modality into this function.

Further investigation of these issues is clearly warranted, and it is expected to lead to improved robustness and performance of audio-visual ASR. Progress in addressing some or all of these questions can also benefit other areas where joint audio and visual speech processing is suitable (Chen and Rao, 1998), such as speaker identification and verification (Jourlin et al., 1997; Wark and Sridharan, 1998; Fröba et al., 1999; Jain et al., 1999; Maison et al., 1999; Chibelushi et al., 2002; Zhang et al., 2002), visual text-to-speech (Cohen and Massaro, 1994; Chen et al., 1995; Cosatto et al., 2000), speech event detection (De Cuetos et al., 2000), video indexing and retrieval (Huang et al., 1999), speech enhancement (Girin et al., 2001b), coding (Foucher et al., 1998), signal separation (Girin et al., 2001a), and speaker localization (Bub et al., 1995; Wang and Brandstein, 1999; Zotkin et al., 2002). Improvements in these areas will result in more robust and natural human-computer interaction.

ACKNOWLEDGEMENTS

We would like to acknowledge a number of people for particular contributions to this work: Giridharan Iyengar and Andrew Senior (IBM) for their help with face and mouth region detection on the IBM ViaVoiceTM and other audio-visual data discussed in this chapter; Rich Wilkins and Eric Helmuth (formerly with IBM) for their efforts in data collection; Guillaume Gravier (currently at IRISA/INRIA Rennes) for the joint multi-stream HMM training and full decoding on the connected digits and LVCSR tasks; Roland Goecke (currently at the Australian National University) for experiments on audio-visual based enhancement of audio features during a summer internship at IBM; Hervé Glotin (ERSS-CNRS) and Dimitra Vergyri (SRI) for their work during the summer 2000 Johns Hopkins workshop on utterance-dependent multi-stream exponent estimation based on speech voicing and late audio-visual fusion within the discriminative model combination framework, respectively; and the remaining summer workshop student team members for invaluable help.

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.
- Adjoudani, A., Guiard-Marigny, T., Le Goff, B., Reveret, L., and Benoît, C. (1997). A multimedia platform for audio-visual speech processing. *Proc. European Conference on Speech Communication and Technology*, Rhodes, Greece, pp. 1671–1674.
- Alissali, M., Deléglise, P., and Rogozan, A. (1996). Asynchronous integration of visual information in an automatic speech recognition system. *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 34–37.
- Anastasakos, T., McDonough, J., and Makhoul, J. (1997). Speaker adaptive training: A maximum likelihood approach to speaker normalization. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, pp. 1043–1046.
- André-Obrecht, R., Jacob, B., and Parlangeau, N. (1997). Audio visual speech recognition and segmental master slave HMM. *Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, pp. 49–52.

- Bahl, L.R., Brown, P.F., DeSouza, P.V., and Mercer, L.R. (1986). Maximum mutual information estimation of hidden Markov model parameters for speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Tokyo, Japan, pp. 49–52.
- Barker, J.P. and Berthommier, F. (1999). Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models. *Proc. International Conference on Auditory-Visual Speech Processing*, Santa Cruz, CA, pp. 112–117.
- Basu, S., Oliver, N., and Pentland, A. (1998). 3D modeling and tracking of human lip motions. *Proc. International Conference on Computer Vision*, Mumbai, India, pp. 337–343.
- Bernstein, L.E., Demorest, M.E., and Tucker, P.E. (1998). What makes a good speechreader? First you have to find one. In Campbell, R., Dodd, B., and Burnham, D. (Eds.), *Hearing by Eye II*. Hove, United Kingdom: Psychology Press Ltd. Publishers, pp. 211–227.
- Berthommier, F. and Glotin, H. (1999). A new SNR-feature mapping for robust multistream speech recognition. *Proc. International Congress on Phonetic Sciences*, San Francisco, CA, pp. 711–715.
- Berthommier, F. (2001). Audio-visual recognition of spectrally reduced speech. *Proc. International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark, pp. 183–189.
- Beyerlein, P. (1998). Discriminative model combination. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, pp. 481–484.
- Boulevard, H. and Dupont, S. (1996). A new ASR approach based on independent processing and recombination of partial frequency bands. *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 426–429.
- Brand, M., Oliver, N., and Pentland, A. (1997). Coupled hidden Markov models for complex action recognition. *Proc. Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 994–999.
- Bregler, C., Hild, H., Manke, S., and Waibel, A. (1993). Improving connected letter recognition by lipreading. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Minneapolis, MN, pp. 557–560.
- Bregler, C. and Konig, Y. (1994). “Eigenlips” for robust speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, 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.
- Bub, U., Hunke, M., and Waibel, A. (1995). Knowing who to listen to in speech recognition: Visually guided beamforming. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Detroit, MI, pp. 848–851.
- Campbell, R., Dodd, B., and Burnham, D., Eds. (1998). *Hearing by Eye II*. Hove, United Kingdom: Psychology Press Ltd. Publishers.
- Chan, M.T., Zhang, Y., and Huang, T.S. (1998). Real-time lip tracking and bimodal continuous speech recognition. *Proc. Workshop on Multimedia Signal Processing*, Redondo Beach, CA, pp. 65–70.
- Chan, M.T. (2001). HMM based audio-visual speech recognition integrating geometric- and appearance-based visual features. *Proc. Workshop on Multimedia Signal Processing*, Cannes, France, pp. 9–14.
- Chandramohan, D. and Silsbee, P.L. (1996). A multiple deformable template approach for visual speech recognition. *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 50–53.
- Chatfield, C. and Collins, A.J. (1991). *Introduction to Multivariate Analysis*. London, United Kingdom: Chapman and Hall.
- Chen, T., Graf, H.P., and Wang, K. (1995). Lip synchronization using speech-assisted video processing. *IEEE Signal Processing Letters*, 2(4):57–59.
- Chen, T. and Rao, R.R. (1998). Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–851.
- Chen, T. (2001). Audiovisual speech processing. Lip reading and lip synchronization. *IEEE Signal Processing Magazine*, 18(1):9–21.
- Chibelushi, C.C., Deravi, F., and Mason, J.S.D. (1996). *Survey of Audio Visual Speech Databases*. Technical Report. Swansea, United Kingdom: Department of Electrical and Electronic Engineering, University of Wales.
- Chibelushi, C.C., Deravi, F., and Mason, J.S.D. (2002). A review of speech-based bimodal recognition. *IEEE Transactions on Multimedia*, 4(1):23–37.
- Chiou, G. and Hwang, J.-N. (1997). Lipreading from color video. *IEEE Transactions on Image Processing*, 6(8):1192–1195.
- Chou, W., Juang, B.-H., Lee, C.-H., and Soong, F. (1994). A minimum error rate pattern recognition approach to speech recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 8(1):5–31.
- Chu, S. and Huang, T. (2000). Bimodal speech recognition using coupled hidden Markov models. *Proc. International Conference on Spoken Language Processing*, Beijing, China, vol. II, pp. 747–750.
- Chu, S.M. and Huang, T.S. (2002). Audio-visual speech modeling using coupled hidden Markov models. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, pp. 2009–2012.
- Cohen, M.M. and Massaro, D.W. (1994). What can visual speech synthesis tell visual speech recognition? *Proc. Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA.

- Cootes, T.F., Taylor, C.J., Cooper, D.H., and Graham, J. (1995). Active shape models - their training and application. *Computer Vision and Image Understanding*, 61(1):38–59.
- Cootes, T.F., Edwards, G.J., and Taylor, C.J. (1998). Active appearance models. *Proc. European Conference on Computer Vision*, Freiburg, Germany, pp. 484–498.
- Cosatto, E., Potamianos, G., and Graf, H.P. (2000). Audio-visual unit selection for the synthesis of photo-realistic talking-heads. *Proc. International Conference on Multimedia and Expo*, New York, NY, pp. 1097–1100.
- Cox, S., Matthews, I., and Bangham, A. (1997). Combining noise compensation with visual information in speech recognition. *Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, pp. 53–56.
- Czap, L. (2000). Lip representation by image ellipse. *Proc. International Conference on Spoken Language Processing*, Beijing, China, vol. IV, pp. 93–96.
- Dalton, B., Kaucic, R., and Blake, A. (1996). Automatic speechreading using dynamic contours. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 373–382.
- Daubechies, I. (1992). *Wavelets*. Philadelphia, PA: S.I.A.M.
- Davoine, F., Li, H., and Forchheimer, R. (1997). Video compression and person authentication. In Bigün, J., Chollet, G., and Borgefors, G. (Eds.), *Audio-and Video-based Biometric Person Authentication*. Berlin, Germany: Springer, pp. 353–360.
- De Cuetos, P., Neti, C., and Senior, A. (2000). Audio-visual intent to speak detection for human computer interaction. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, pp. 1325–1328.
- Deligne, S., Potamianos, G., and Neti, C. (2002). Audio-visual speech enhancement with AVCDCN (audio-visual codebook dependent cepstral normalization). In Press: *Proc. International Conference on Spoken Language Processing*, Denver, CO.
- Deller, Jr., J.R., Proakis, J.G., and Hansen, J.H.L. (1993). *Discrete-Time Processing of Speech Signals*. Englewood Cliffs, NJ: Macmillan Publishing Company.
- 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.
- Dougherty, E.R. and Giardina, C.R. (1987). *Image Processing – Continuous to Discrete, Vol. 1. Geometric, Transform, and Statistical Methods*. Englewood Cliffs, NJ: Prentice Hall.
- Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis*. London, United Kingdom: John Wiley and Sons.
- Duchnowski, P., Meier, U., and Waibel, A. (1994). See me, hear me: Integrating automatic speech recognition and lip-reading. *Proc. International Conference on Spoken Language Processing*, Yokohama, Japan, pp. 547–550.
- Dupont, S. and Luettin, J. (2000). Audio-visual speech modeling for continuous speech recognition. *IEEE Transactions on Multimedia*, 2(3):141–151.
- Foucher, E., Girin, L., and Feng, G. (1998). Audiovisual speech coder: Using vector quantization to exploit the audio/video correlation. *Proc. Workshop on Audio Visual Speech Processing*, Terrigal, Australia, pp. 67–71.
- Fröba, B., Küblbeck, C., Rothe, C., and Plankensteiner, P. (1999). Multi-sensor biometric person recognition in an access control system. *Proc. International Conference on Audio and Video-based Biometric Person Authentication*, Washington, DC, pp. 55–59.
- Gales, M.J.F. (1997). “Nice” model based compensation schemes for robust speech recognition. *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, pp. 55–59.
- Gales, M.J.F. (1999). *Maximum Likelihood Multiple Projection Schemes for Hidden Markov Models*. Technical Report. Cambridge, United Kingdom: Cambridge University.
- Gauvain, J.-L. and Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.
- Ghitza, O. (1986). Auditory nerve representation as a front end for speech recognition in noisy environments. *Computer, Speech and Language*, 1:109–130.
- Girin, L., Feng, G., and Schwartz, J.-L. (1995). Noisy speech enhancement with filters estimated from the speaker’s lips. *Proc. European Conference on Speech Communication and Technology*, Madrid, Spain, pp. 1559–1562.
- Girin, L., Allard, A., and Schwartz, J.-L. (2001a). Speech signals separation: A new approach exploiting the coherence of audio and visual speech. *Proc. Workshop on Multimedia Signal Processing*, Cannes, France, pp. 631–636.
- Girin, L., Schwartz, J.-L., and Feng, G. (2001b). Audio-visual enhancement of speech in noise. *Journal of the Acoustical Society of America*, 109(6):3007–3020.
- Glotin, H. and Berthommier, F. (2000). Test of several external posterior weighting functions for multiband full combination ASR. *Proc. International Conference on Spoken Language Processing*, Beijing, China, vol. I, pp. 333–336.
- Glotin, H., Vergyri, D., Neti, C., Potamianos, G., and Luettin, J. (2001). Weighting schemes for audio-visual fusion in speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, pp. 173–176.
- Goecke, R., Potamianos, G., and Neti, C. (2002). Noisy audio feature enhancement using audio-visual speech data. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, pp. 2025–2028.

- Goldschen, A.J., Garcia, O.N., and Petajan, E.D. (1996). Rationale for phoneme-viseme mapping and feature selection in visual speech recognition. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 505–515.
- Golub, G.H. and Van Loan, C.F. (1983). *Matrix Computations*. Baltimore, MD: The Johns Hopkins University Press.
- Gonzalez, R.C. and Wintz, P. (1977). *Digital Image Processing*. Reading, MA: Addison-Wesley Publishing Company.
- Gopinath, R.A. (1998). Maximum likelihood modeling with Gaussian distributions for classification. *Proc. International Conference on Acoustics, Speech and Signal Processing*, 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. *Proc. International Conference on Systems, Man, and Cybernetics*, Orlando, FL, pp. 2034–2039.
- Grant, K.W. and Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. *Proc. International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark, pp. 132–137.
- Gravier, G., Axelrod, S., Potamianos, G., and Neti, C. (2002a). Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, pp. 853–856.
- Gravier, G., Potamianos, G., and Neti, C. (2002b). Asynchrony modeling for audio-visual speech recognition. *Proc. Human Language Technology Conference*, San Diego, CA.
- 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*. Cambridge, MA: MIT Press, vol. 9, pp. 751–757.
- Gurbuz, S., Tufekci, Z., Patterson, E., and Gowdy, J.N. (2001). Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, pp. 177–180.
- Heckmann, M., Berthommier, F., and Kroschel, K. (2001). A hybrid ANN/HMM audio-visual speech recognition system. *Proc. International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark, pp. 190–195.
- 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.
- Hermansky, H. and Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589.
- Hermansky, H., Tibrewala, S., and Pavel, M. (1996). Towards ASR on partially corrupted speech. *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 462–465.
- Hernando, J., Ayarte, J., and Monte, E. (1995). Optimization of speech parameter weighting for CDHMM word recognition. *Proc. European Conference on Speech Communication and Technology*, Madrid, Spain, pp. 105–108.
- Huang, F.J. and Chen, T. (2001). Consideration of Lombard effect for speechreading. *Proc. Workshop on Multimedia Signal Processing*, Cannes, France, pp. 613–618.
- Huang, J., Liu, Z., Wang, Y., Chen, Y., and Wong, E.K. (1999). Integration of multimodal features for video scene classification based on HMM. *Proc. Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, pp. 53–58.
- Iyengar, G. and Neti, C. (2001). Detection of faces under shadows and lighting variations. *Proc. Workshop on Multimedia Signal Processing*, Cannes, France, pp. 15–20.
- Iyengar, G., Potamianos, G., Neti, C., Faruque, T., and Verma, A. (2001). Robust detection of visual ROI for automatic speechreading. *Proc. Workshop on Multimedia Signal Processing*, Cannes, France, pp. 79–84.
- Jain, A., Bolle, R., and Pankanti, S. (1999). Introduction to Biometrics. In Jain, A., Bolle, R., and Pankanti, S. (Eds.), *Biometrics. Personal Identification in Networked Society*. Norwell, MA: Kluwer Academic Publishers, pp. 1–41.
- 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.
- Jiang, H., Soong, F., and Lee, C. (2001). Hierarchical stochastic feature matching for robust speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, pp. 217–220.
- Jourlin, P. (1997). Word dependent acoustic-labial weights in HMM-based speech recognition. *Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, pp. 69–72.
- Jourlin, P., Luetin, J., Genoud, D., and Wassner, H. (1997). Acoustic-labial speaker verification. *Pattern Recognition Letters*, 18(9):853–858.
- Juang, B.H. (1991). Speech recognition in adverse environments. *Computer, Speech and Language*, 5:275–294.
- Kass, M., Witkin, A., and Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331.
- Kittler, J., Hatef, M., Duin, R.P.W., and Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239.
- Kober, R., Harz, U., and Schiffers, J. (1997). Fusion of visual and acoustic signals for command-word recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, pp. 1495–1497.

- Krone, G., Talle, B., Wichert, A., and Palm, G. (1997). Neural architectures for sensorfusion in speech recognition. *Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, pp. 57–60.
- Leggetter, C.J. and Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185.
- Li, N., Dettmer, S., and Shah, M. (1995). Lipreading using eigensequences. *Proc. International Workshop on Automatic Face- and Gesture-Recognition*, Zurich, Switzerland, pp. 30–34.
- Lippmann, R.P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.
- Liu, F., Stern, R., Huang, X., and Acero, A. (1993). Efficient cepstral normalization for robust speech recognition. *Proc. ARPA Workshop on Human Language Technology*, Princeton, NJ.
- Luettin, J., Thacker, N.A., and Beet, S.W. (1996). Speechreading using shape and intensity information. *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 58–61.
- Luettin, J. and Thacker, N.A. (1997). Speechreading using probabilistic models. *Computer Vision and Image Understanding*, 65(2):163–178.
- Luettin, J., Potamianos, G., and Neti, C. (2001). Asynchronous stream modeling for large vocabulary audio-visual speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, pp. 169–172.
- Maison, B., Neti, C., and Senior, A. (1999). Audio-visual speaker recognition for broadcast news: some fusion techniques. *Proc. Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, pp. 161–167.
- Marschark, M., LePoutre, D., and Bement, L. (1998). Mouth movement and signed communication. In Campbell, R., Dodd, B., and Burnham, D. (Eds.), *Hearing by Eye II*. Hove, United Kingdom: Psychology Press Ltd. Publishers, pp. 245–266.
- Mase, K. and Pentland, A. (1991). Automatic lipreading by optical flow analysis. *Systems and Computers in Japan*, 22(6):67–75.
- Massaro, D.W. (1996). Bimodal speech perception: A progress report. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 79–101.
- Massaro, D.W. and Stork, D.G. (1998). Speech recognition and sensory integration. *American Scientist*, 86(3):236–244.
- Matthews, I., Bangham, J.A., and Cox, S. (1996). Audio-visual speech recognition using multiscale nonlinear image decomposition. *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 38–41.
- Matthews, I., Cootes, T., Cox, S., Harvey, R., and Bangham, J.A. (1998). Lipreading using shape, shading and scale. *Proc. Workshop on Audio Visual Speech Processing*, Terrigal, Australia, pp. 73–78.
- Matthews, I., Potamianos, G., Neti, C., and Luettin, J. (2001). A comparison of model and transform-based visual features for audio-visual LVCSR. *Proc. International Conference on Multimedia and Expo*, Tokyo, Japan.
- McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264:746–748.
- Meier, U., Hürst, W., and Duchnowski, P. (1996). Adaptive bimodal sensor fusion for automatic speechreading. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, pp. 833–836.
- Messer, K., Matas, J., Kittler, J., Luettin, J., and Maitre, G. (1999). XM2VTS: The extended M2VTS database. *Proc. International Conference on Audio and Video-based Biometric Person Authentication*, Washington, DC, pp. 72–76.
- Miyajima, C., Tokuda, K., and Kitamura, T. (2000). Audio-visual speech recognition using MCE-based HMMs and model-dependent stream weights. *Proc. International Conference on Spoken Language Processing*, Beijing, China, vol. II, pp. 1023–1026.
- Movellan, J.R. and Chadderdon, G. (1996). Channel separability in the audio visual integration of speech: A Bayesian approach. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 473–487.
- Nadas, A., Nahamoo, D., and Picheny, M. (1989). Speech recognition using noise adaptive prototypes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37:1495–1503.
- Nakamura, S., Ito, H., and Shikano, K. (2000). Stream weight optimization of speech and lip image sequence for audio-visual speech recognition. *Proc. International Conference on Spoken Language Processing*, Beijing, China, vol. III, pp. 20–23.
- Nakamura, S. (2001). Fusion of audio-visual information for integrated speech processing. In Bigun, J. and Smeraldi, F. (Eds.), *Audio-and Video-Based Biometric Person Authentication*. Berlin, Germany: Springer-Verlag, pp. 127–143.
- Nefian, A.V., Liang, L., Pi, X., Liu, X., and Murphy, K. (2002). Dynamic Bayesian networks for audio-visual speech recognition. In Press: *EURASIP Journal on Applied Signal Processing*.
- Nelder, J.A. and Mead, R. (1965). A simplex method for function minimisation. *Computing Journal*, 7(4):308–313.
- Neti, C. (1994). Neuromorphic speech processing for noisy environments. *Proc. International Conference on Neural Networks*, Orlando, FL, pp. 4425–4430.
- Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J., Mashari, A., and Zhou, J. (2000). *Audio-Visual Speech Recognition*. Final Workshop 2000 Report. Baltimore, MD: Center for Language and Speech Processing, The Johns Hopkins University.

- Neumeyer, L., Sankar, A., and Digalakis, V. (1995). A comparative study of speaker adaptation techniques. *Proc. European Conference on Speech Communication and Technology*, Madrid, Spain, pp. 1127–1130.
- Okawa, S., Nakajima, T., and Shirai, K. (1999). A recombination strategy for multi-band speech recognition based on mutual information criterion. *Proc. European Conference on Speech Communication and Technology*, Budapest, Hungary, pp. 603–606.
- Pan, H., Liang, Z.-P., Anastasio, T.J., and Huang, T.S. (1998). A hybrid NN-Bayesian architecture for information fusion. *Proc. International Conference on Image Processing*, Chicago, IL, vol. I, pp. 368–371.
- Patterson, E.K., Gurbuz, S., Tufekci, Z., and Gowdy, J.N. (2001). Noise-based audio-visual fusion for robust speech recognition. *Proc. International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark, pp. 196–199.
- Patterson, E.K., Gurbuz, S., Tufekci, Z., and Gowdy, J.N. (2002). CUAVE: A new audio-visual database for multimodal human-computer interface research. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Orlando, FL, pp. 2017–2020.
- Pavlovic, V. (1998). Multimodal tracking and classification of audio-visual features. *Proc. International Conference on Image Processing*, Chicago, IL, vol. I, pp. 343–347.
- Petajan, E.D. (1984). Automatic lipreading to enhance speech recognition. *Proc. Global Telecommunications Conference*, Atlanta, GA, pp. 265–272.
- Pigeon, S. and Vandendorpe, L. (1997). The M2VTS multimodal face database. In Bigün, J., Chollet, G., and Borgefors, G. (Eds.), *Audio-and Video-based Biometric Person Authentication*, Berlin, Germany: Springer, pp. 403–409.
- 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. *Proc. International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, pp. 901–904.
- Potamianos, G. and Graf, H.P. (1998). Discriminative training of HMM stream exponents for audio-visual speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, pp. 3733–3736.
- Potamianos, G., Graf, H.P., and Cosatto, E. (1998). An image transform approach for HMM based automatic lipreading. *Proc. International Conference on Image Processing*, Chicago, IL, vol. I, pp. 173–177.
- Potamianos, G. and Potamianos, A. (1999). Speaker adaptation for audio-visual speech recognition. *Proc. European Conference on Speech Communication and Technology*, Budapest, Hungary, pp. 1291–1294.
- Potamianos, G. and Neti, C. (2000). Stream confidence estimation for audio-visual speech recognition. *Proc. International Conference on Spoken Language Processing*, Beijing, China, vol. III, pp. 746–749.
- Potamianos, G. and Neti, C. (2001a). Automatic speechreading of impaired speech. *Proc. International Conference on Auditory-Visual Speech Processing*, Aalborg, Denmark, pp. 177–182.
- Potamianos, G. and Neti, C. (2001b). Improved ROI and within frame discriminant features for lipreading. *Proc. International Conference on Image Processing*, Thessaloniki, Greece, vol. III, pp. 250–253.
- Potamianos, G., Luettin, J., and Neti, C. (2001a). Hierarchical discriminant features for audio-visual LVCSR. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, pp. 165–168.
- Potamianos, G., Neti, C., Iyengar, G., Senior, A.W., and Verma, A. (2001b). A cascade visual front end for speaker independent automatic speechreading. *International Journal of Speech Technology*, 4(3–4):193–208.
- Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. (1995). *Numerical Recipes in C. The Art of Scientific Computing*. Cambridge, United Kingdom: 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.
- Robert-Ribes, J., Piquemal, M., Schwartz, J.-L., and Escudier, P. (1996). Exploiting sensor fusion architectures and stimuli complementarity in AV speech recognition. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 193–210.
- Robert-Ribes, J., Schwartz, J.-L., Lallouache, T., and Escudier, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual, and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America*, 103(6):3677–3689.
- Rogozan, A., Deléglise, P., and Alissali, M. (1997). Adaptive determination of audio and visual weights for automatic speech recognition. *Proc. European Tutorial Workshop on Audio-Visual Speech Processing*, Rhodes, Greece, pp. 61–64.
- Rogozan, A. and Deléglise, P. (1998). Adaptive fusion of acoustic and visual sources for automatic speech recognition. *Speech Communication*, 26(1–2):149–161.
- Rogozan, A. (1999). Discriminative learning of visual data for audiovisual speech recognition. *International Journal on Artificial Intelligence Tools*, 8(1):43–52.
- Rosenblum, L.D. and Saldaña, H.M. (1998). Time-varying information for visual speech perception. In Campbell, R., Dodd, B., and Burnham, D. (Eds.), *Hearing by Eye II*. Hove, United Kingdom: Psychology Press Ltd. Publishers, pp. 61–81.

- Rowley, H.A., Baluja, S., and Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38.
- Scanlon, P. and Reilly, R. (2001). Feature analysis for automatic speechreading. *Proc. Workshop on Multimedia Signal Processing*, Cannes, France, pp. 625–630.
- Senior, A.W. (1999). Face and feature finding for a face recognition system. *Proc. International Conference on Audio and Video-based Biometric Person Authentication*, Washington, DC, pp. 154–159.
- Silsbee, P.L. (1994). Motion in deformable templates. *Proc. International Conference on Image Processing*, Austin, TX, vol. 1, pp. 323–327.
- Silsbee, P.L. and Bovik, A.C. (1996). Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(5):337–351.
- Smeele, P.M.T. (1996). Psychology of human speechreading. In Stork, D.G. and Hennecke, M.E. (Eds.), *Speechreading by Humans and Machines*. Berlin, Germany: Springer, pp. 3–15.
- Stork, D.G. and Hennecke, M.E., Eds. (1996). *Speechreading by Humans and Machines*. Berlin, Germany: Springer.
- Su, Q. and Silsbee, P.L. (1996). Robust audiovisual integration using semicontinuous hidden Markov models. *Proc. International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 42–45.
- Sumby, W.H. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2):212–215.
- 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*. London, United Kingdom: Lawrence Erlbaum Associates, pp. 3–51.
- 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.
- Sung, K. and Poggio, T. (1998). Example-based learning for view-based human face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):39–51.
- Swets, D.L. and Weng, J. (1996). Using discriminant eigenfaces for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):831–836.
- Teissier, P., Robert-Ribes, J., and Schwartz, J.L. (1999). Comparing models for audiovisual fusion in a noisy-vowel recognition task. *IEEE Transactions on Speech and Audio Processing* 7(6):629–642.
- Tomlinson, M.J., Russell, M.J., and Brooke, N.M. (1996). Integrating audio and visual information to provide highly robust speech recognition. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, pp. 821–824.
- Vanegas, O., Tanaka, A., Tokuda, K., and Kitamura, T. (1998). HMM-based visual speech recognition using intensity and location normalization. *Proc. International Conference on Spoken Language Processing*, Sydney, Australia, pp. 289–292.
- Varga, P. and Moore, R.K. (1990). Hidden Markov model decomposition of speech and noise. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Albuquerque, NM, pp. 845–848.
- Vergyri, D. (2000). *Integration of Multiple Knowledge Sources in Speech Recognition Using Minimum Error Training*. PhD Thesis. Baltimore, MD: Center for Speech and Language Processing, The Johns Hopkins University.
- Wang, C. and Brandstein, M.S. (1999). Multi-source face tracking with audio and visual data. *Proc. Workshop on Multimedia Signal Processing*, Copenhagen, Denmark, pp. 475–481.
- Wark, T. and Sridharan, S. (1998). A syntactic approach to automatic lip feature extraction for speaker identification. *Proc. International Conference on Acoustics, Speech and Signal Processing*, Seattle, WA, pp. 3693–3696.
- Williams, J.J., Rutledge, J.C., Garstecki, D.C., and Katsaggelos, A.K. (1997). Frame rate and viseme analysis for multimedia applications. *Proc. Workshop on Multimedia Signal Processing*, Princeton, NJ, pp. 13–18.
- Xu, L., Krzyżak, A., and Suen, C.Y. (1992). Methods of combining multiple classifiers and their applications in handwritten character recognition. *IEEE Transactions on Systems Man and Cybernetics*, 22(3):418–435.
- Yehia, H., Rubin, P., and Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1–2):23–43.
- Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (1999). *The HTK Book*. Cambridge, United Kingdom: Entropic Ltd.
- Yuille, A.L., Hallinan, P.W., and Cohen, D.S. (1992). Feature extraction from faces using deformable templates. *International Journal of Computer Vision*, 8(2):99–111.
- Zhang, X., Broun, C.C., Mersereau, R.M., and Clements, M. (2002). Automatic speechreading with applications to human-computer interfaces. In Press: *EURASIP Journal on Applied Signal Processing*.
- Zhang, Y., Levinson, S., and Huang, T. (2000). Speaker independent audio-visual speech recognition. *Proc. International Conference on Multimedia and Expo*, New York, NY, pp. 1073–1076.
- Zotkin, D.N., Duraiswami, R., and Davis, L.S. (2002). Joint audio-visual tracking using particle filters. In Press: *EURASIP Journal on Applied Signal Processing*.