10.8

Exploiting Visual Information in Automatic Speech Processing

Petar S. Aleksic
Northwestern University

Gerasimos Potamianos
IBM T.J. Watson Research Center

Aggelos K. Katsaggelos
Northwestern University

1 Introduction

With the increasing use of computers in everyday life, the challenging goal of achieving natural, pervasive, and ubiquitous human–computer interaction (HCI) has become very important, affecting, for example, productivity, customer satisfaction, and accessibility, among others. In contrast to the current prevailing HCI paradigm that mostly relies on locally tied, single-modality and computer-centric input/output, future HCI scenarios are envisioned where the computer fades into the background, accepting and responding to user requests in a humanlike behavior, and at the user's location. Not surprisingly, speech is viewed as an integral part of such HCI, conveying not only user linguistic information, but also emotion, identity, location, and computer feedback [1].

However, although great progress has been achieved over the past decades, computer processing of speech still lags significantly compared to human performance levels. For example, automatic speech recognition (ASR) lacks robustness to channel mismatch and environment noise [1, 2], underperforming human speech perception by up to an order of magnitude even in clean conditions [3]. Similarly, text-to-speech (TTS) systems continue to lag in naturalness, expressiveness, and, somewhat less, in intelligibility [4]. Furthermore, typical real-life interaction scenarios, where humans address other humans in addition to the computer, may be located in a variable far-field position compared with the computer sensors, or utilize emotion and nonacoustic cues to convey a message, prove insurmountably challenging to traditional systems that rely on the audio signal alone. In contrast, humans easily master complex communication tasks by using additional channels of information whenever required, most notably the visual sensory channel. It is therefore only natural that significant interest and effort has recently been focused on exploiting the visual modality to improve HCI [5–9]. In this chapter, we review such efforts

with emphasis on the main techniques used in the extraction and integration of the visual signal information into speech processing HCI systems.

Of central importance to human communication is the visual information present in the face. In particular, the lower face plays an integral role in the production of human speech and of its perception, both being audiovisual in nature [9, 10]. Indeed, the visual modality benefit to speech intelligibility has been quantified as far back as in 1954 [11]. Furthermore, bimodal integration of audio and visual stimuli in perceiving speech has been demonstrated by the McGurk effect [12], when for example a person is presented with the audio stimulus "baba" superimposed on a video of moving lips uttering the sound "gaga" the person perceives the sound "dada". Visual speech information is especially critical to the hearing impaired: Mouth movement plays an important role in both sign language and simultaneous communication between the deaf [13].

Face visibility benefits speech perception due to the fact that the visual signal is both correlated to the produced audio signal and contains complementary information to it [14–16]. The former allows the partial recovery of the acoustic signal from visual speech [17], a process akin to speech enhancement when the audio is corrupted by noise [18, 19]. The latter is due to at least the partial visibility of the place of articulation, through the tongue, teeth, and lips, and can help disambiguate speech sounds that are highly confusable from acoustics alone; for example, the unvoiced consonants "p" (a bilabial) and "k" (a velar), among others [16]. Not surprisingly, these observations have motivated significant research over the past 20 years on the automatic recognition of visual speech, also known as automatic speechreading, and its integration with traditional audio-only systems, giving rise to audiovisual ASR [20–47].

In addition to improving speech perception, face visibility provides direct and natural communication between humans. Computers however, typically utilize audio-only text-to-speech synthesis to communicate information back to the user in a manner that lags in naturalness, expressiveness, and intelligibility compared with human speech. To address these shortcomings, much research work has recently focused on augmenting TTS systems by synthesized visual speech [7, 48]. Such systems generate synthetic talking faces that can be directly driven by the acoustic signal or the required text, providing animated or photo-realistic output [5, 47–62]. The resulting systems can have widely varying HCI applications, ranging from assistance to hearing impaired persons, to interactive computer-based learning and entertainment.

It is worth noting that face visibility plays additional important roles in human-to-human communication by providing speech segmental and source localization information, as well as by conveying speaker identity and emotion. All are very important to HCI, with obvious implications for ASR or TTS, among others. A number of recently proposed techniques use visual-only or joint audio-visual signal processing for speech activity detection and source localization [63–67], identity recognition from face appearance or visual speech [8, 68–76], and visual recognition and synthesis of human facial emotional expressions [77, 78]. In all cases, the visual modality can significantly improve audio-only systems.

To automatically process and incorporate the visual information into the above speech-based HCI technologies, a number of steps are required that are surprisingly similar. Central to all technologies is the feature representation of visual speech and its robust extraction. In addition, appropriate integration of the audio and visual representations is required for audiovisual ASR, speaker recognition, speech activity detection, and emotion recognition, to ensure improved performance of the bimodal systems over audio-only baselines. In a number of technologies, this integration occurs by exploiting audiovisual signal correlation: for example, audio enhancement by using visual information, speech-to-video synthesis, and detection of synchronous audiovisual sources (localization). Finally, unique to audiovisual TTS and speech-to-video synthesis is the generation of the final video from the synthesized visual speech representation (facial animation). The similarities between the required processing components is reinforced in Fig. 1, where conversions and interactions between the acoustic, visual, and textual representation of speech are graphically depicted.

In this chapter, we review these main processing components, and we discuss their application to speech-based HCI, with main emphasis on ASR, TTS, and speaker recognition. In particular, in Section 2, we focus on visual feature extraction. Section 3 is devoted to the main audiovisual integration strategies, with Section 4 concentrating on their application to audiovisual ASR. Section 5 addresses audiovisual speech synthesis, whereas Section 6 discusses audiovisual speaker recognition. Finally, Section 7 touches on additional applications such as speaker localization, speech activity detection, and emotion recognition, and provides a summary and a short discussion.

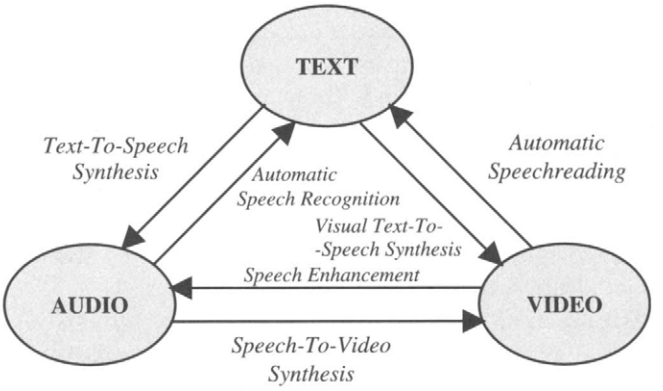


FIGURE 1 Conversions and interactions between the acoustic, visual, and text representations of speech that are the focus of this work (adapted from [5]).

2 Analysis of Visual Signals

The first critical issue in the design and implementation of audiovisual speech systems for HCI is the choice of visual features and their robust extraction from video. Visual speech information is mostly contained in the speaker's mouth region, therefore, typically, the visual features consist of appropriate representations of mouth appearance and/or shape. Indeed, the various sets of visual features proposed in the literature over the last 20 years for visual speech processing applications are generally grouped into three categories [21]: (a) low-level, or appearance-based features, such as transformed vectors of the mouth region pixel intensities using, for example, image compression techniques [22–31, 54, 70, 72]; (b) high-level, or shape-based features, such as geometric or model-based representations of the lip contours [30–41, 52, 54, 68, 73]; and (c) features that are a combination of appearance and shape [29–31].

The choice of visual features clearly mandates the face, lip, or mouth-tracking algorithms required for their extraction, but is also a function of video data quality and resource constraints in the audiovisual speech application. For example, only a crude detection of the mouth region is sufficient to obtain appearance visual features, requiring as little as tracking the face and the two mouth corners. Such steps become even unnecessary if a properly head-mounted video camera is used for data capture, as in [46]. In contrast, a more computationally expensive lip-tracking algorithm is additionally required for shape-based features, being infeasible in videos that contain low-resolution faces. Needless to say, robust tracking of the face, lips, or the mouth region is of paramount importance for using the benefit of visual speech in HCI. In the following, we review such tracking algorithms, before proceeding with a brief description of some commonly used visual features.

2.1 Face Detection, Mouth, and Lip Tracking

Face detection has attracted significant interest in the literature [79–83]. In general, it constitutes a difficult problem, especially in cases where the background, head pose, and lighting are varying. Some reported systems use traditional image processing techniques for face detection, such as color segmentation, edge detection, image thresholding, template matching, or motion information in image sequences [83], taking advantage of the fact that many local facial subfeatures contain strong edges and are approximately rigid.

However, the most widely used techniques follow a statistical modeling approach of face appearance to obtain a binary classification of image regions into the face and nonface classes. Such regions are typically represented as vectors of gray-scale or color image pixel intensities over normalized rectangles of a predetermined size, often projected onto lower-dimensional spaces, and are defined over a "pyramid" of possible locations, scales, and orientations in the image [79]. These regions can be classified using one or more techniques, such as neural networks, clustering algorithms along with distance metrics from the face or nonface spaces, simple linear discriminants, support vector machines, and Gaussian mixture models, for example [79–81]. An alternative popular approach uses a cascade of weak classifiers instead, that are trained using the AdaBoost technique and operate on local appearance features within these regions [82]. Notice that if color information is available, certain image regions that do not contain sufficient number of skin-tone–like pixels can be eliminated from the search [79].

Once face detection is successful, similar techniques can be used in a hierarchic manner to detect a number of interesting facial features such as the mouth corners, eyes, nostrils, chin, etc. The prior knowledge of their relative position on the face can simplify the search task. Such features are needed to determine the mouth region of interest (ROI) and help to normalize it by providing head-pose information. Additional lighting normalization is often applied to the ROI before appearance-based feature extraction (see also Fig. 2A).

Once the ROI is located, a number of algorithms can be used to obtain lip contour estimates. Some popular methods for this task are snakes [84], templates [85], and active shape and appearance models [86]. A snake is an elastic curve represented by a set of control points, and it is used to detect important visual features, such as lines, edges, or contours. The snake control point coordinates are iteratively updated, converging toward a minimum of the energy function, defined on the basis of curve smoothness constraints and a matching criterion to desired features of the image [84]. Templates are parametric curves that are fitted to the desired shape by minimizing an energy function, defined similarly to snakes. Examples of lip contour estimation using a gradient vector field (GVF) snake and two parabolic templates are depicted in Fig. 2B [40].

In contrast, active shape models (ASMs) are statistical models obtained by performing principal component analysis (PCA) on vectors containing the coordinates of a training set of points that lie on the shapes of interest, such as the lip inner and outer contours (see also Fig. 2C). Such vectors are projected onto a lower dimensional space defined by the eigenvectors that correspond to the largest PCA eigenvalues, representing the axes of genuine shape variation. Active appearance models (AAMs) are an extension to ASMs that, in addition to the shape-based model, use two more PCAs: The first captures the appearance variation of the region around the desired shape (for example, of vectors of image pixel intensities within the face contours, as shown in Fig. 2D), whereas the final PCA is built on concatenated weighted vectors of the shape and appearance representations. AAMs thus remove the redundancy due to shape and appearance correlation, and they create a single model that compactly

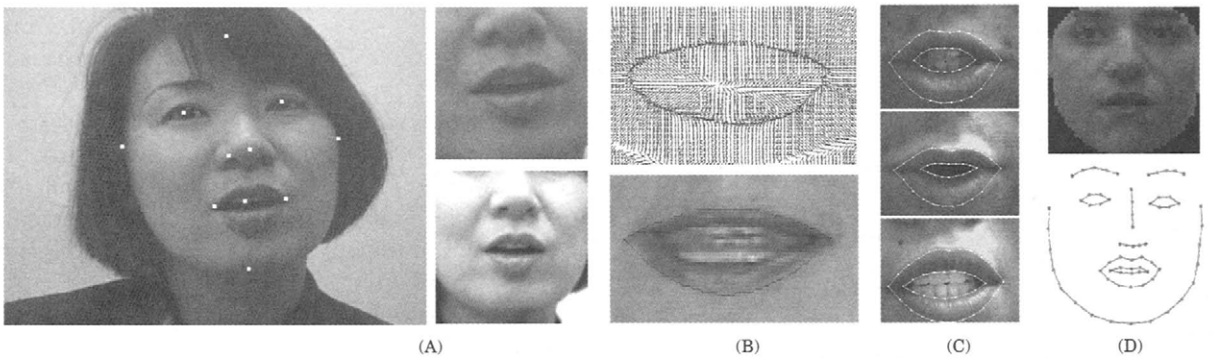


FIGURE 2 Mouth appearance and shape tracking for visual feature extraction. **A:** Eleven detected facial features using the appearance-based approach of [79]. Two corresponding mouth region-of-interests of different sizes and normalization are also depicted [44]. **B:** Lip contour estimation using a gradient vector field snake (**upper:** the snake's external force field is depicted) and two parabolas (**lower**) [40]. **C:** Three examples of lip contour extraction using an active shape model [31]. **D:** Detection of face appearance (**upper**) and shape (**lower**) using active appearance models [29]. (See color insert.)

describes shape and the corresponding appearance deformation. ASMs and AAMs can be used for tracking lips or other shapes by means of the algorithm proposed in [86]. The technique 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. Fitting the models to the image data can be done iteratively, as in [29], or by the downhill simplex method, as in [31]. Examples of lip and face contour estimation by means of ASMs and AAMs are depicted in Figs. 2C and 2D, respectively.

2.2 Visual Features

In the appearance-based approach to visual feature extraction, the pixel-value–based, low-level representation of the mouth ROI is considered as informative for speechreading. Such ROI is extracted by the algorithms discussed in Section 2, and is typically a rectangle containing the mouth, possibly including larger parts of the lower face, such as the jaw and cheeks [44], or could even be the entire face [29] (see also Figs. 2A and 2D). Sometimes, it is extended into a three-dimensional rectangle, containing adjacent frame ROIs, in an effort to capture dynamic speech information [24]. Alternatively, the ROI can correspond to a number of image profiles vertical to the estimated lip contour as in [31], or be just a disk around the mouth center [23]. By concatenating the ROI pixel values, a feature vector $\mathbf{x}_t$ is obtained that is expected to contain most visual speech information (see Fig. 3).

Typically, however, the dimensionality d of the ROI vector $\mathbf{x}_t$ becomes prohibitively large for successful statistical modeling of the classes of interest, such as subphonetic classes via hidden Markov models for audiovisual ASR [87]. For example, in the case of a 64×64-pixel gray-scale ROI, $d = 4{,}096$. Therefore, appropriate, lower-dimensional transformations of $\mathbf{x}_t$ are used as features instead. In general, a $D \times d$-dimensional linear transform matrix $\mathbf{P}$ is sought, such that the transformed data vector $\mathbf{y}_t = \mathbf{x}_t\mathbf{P}$ contains most speechreading information in its $D \ll d$ elements (see also Fig. 3). Matrix $\mathbf{P}$ is typically borrowed from the image compression and pattern classification literatures, and is often obtained based on a number of training ROI vectors. Examples of such transforms are the PCA, also known as "eigenlips," used in the literature for speechreading [22–24, 31], visual TTS [54], and speaker recognition [69], the discrete cosine transform (DCT) [23–26], the discrete wavelet transform (DWT) [24], linear discriminant analysis (LDA) [44, 70], and the maximum likelihood linear transform (MLLT) [44, 72]. Often, such transforms are applied in a cascade [44, 70]. Notice that some are amenable to fast algorithmic implementations. Coupled with the fact that a crude ROI extraction can be achieved by utilizing computationally inexpensive face detection algorithms, appearance-based features allow visual speech representation in real time [46]. Their performance, however, degrades under intense head-pose and lighting variations [46].

In contrast to appearance-based features, high-level shape-based feature extraction assumes that most speechreading information is contained in the shape (inner and outer contours) of the speaker lips, or more generally, in the face contours [29]. As a result, such features achieve a compact representation of visual speech using low-dimensional vectors, and are invariant to head pose and lighting. However, to ensure good performance, their extraction requires robust lip tracking, which often proves difficult and computationally intensive in realistic scenarios.

In general, high-level visual features are divided into geometric and model-based (see also Fig. 3). The former represent features that are meaningful to humans and can be readily extracted from the lip inner and outer contours, such as the height, width, perimeter, and area within the contour.

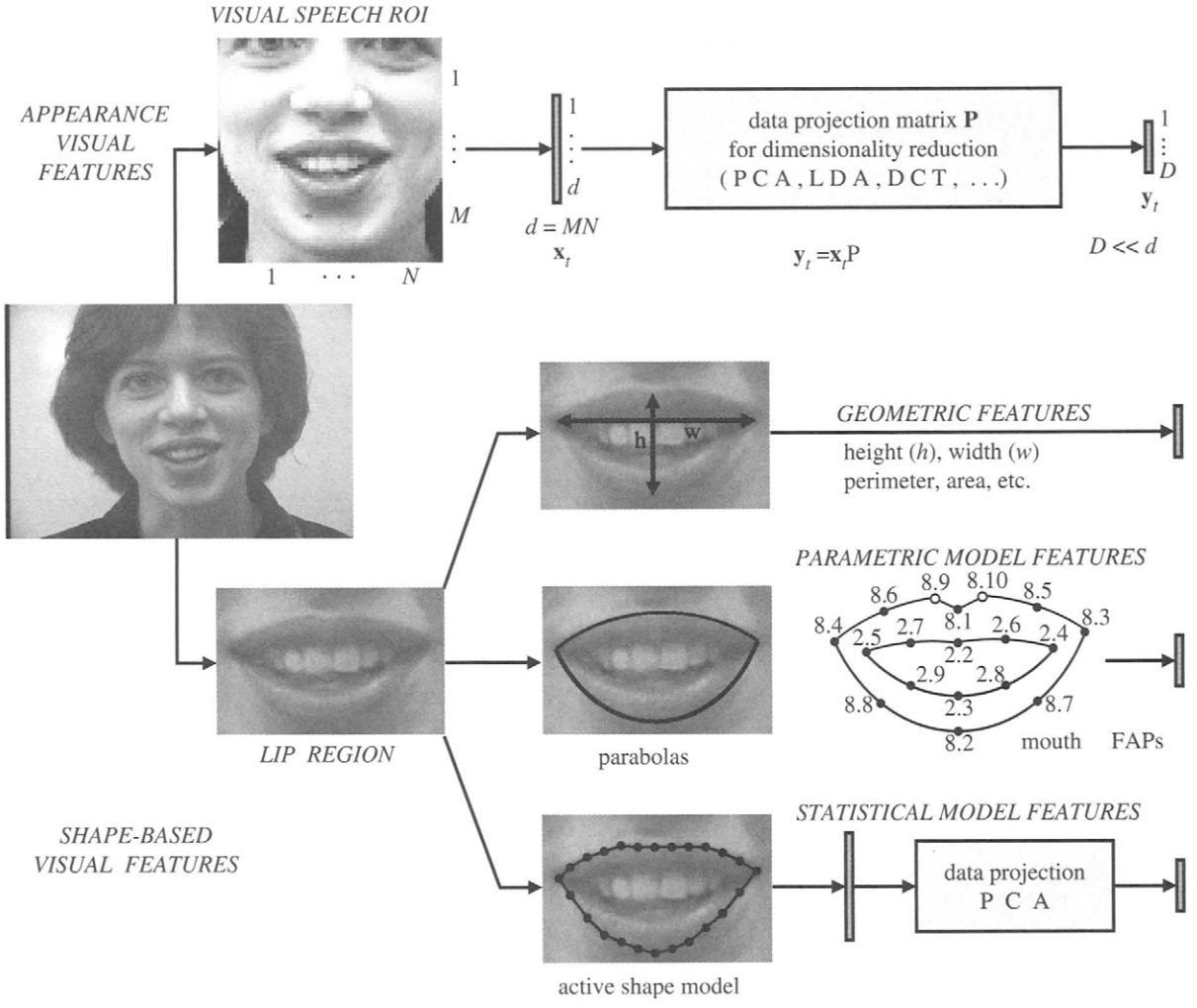


FIGURE 3 Various visual speech feature representation approaches discussed in this section: appearance-based (**upper**) and shape-based features (**lower**) that may utilize lip geometry, parametric, or statistical lip models.

Such features contain significant visual speech information and have been successfully used in speechreading [32–38], visual speech synthesis [5, 54], and speaker recognition [38]. Additional visual features can be derived from the lip contours, such as lip image moments and lip contour Fourier descriptors, that are invariant to affine image transformations [24, 36]. Alternatively, high-level visual features can be model-based, typically obtained in conjunction with one of the parametric or statistical lip-tracking algorithms discussed earlier in Section 2.1. In the parametric approach, the template parameters that track the lips, or in the same manner, the tracking snake's control points or radial vectors, can be directly used as visual speech features [30, 41]. Similarly, ASMs can be used as visual features by applying the model PCA on the vector of point coordinates of the tracked lip contour [31, 68].

In a related, recently introduced approach [40], a standard parametrization of the outer lip contour by means of a subset of facial animation parameters (FAPs) [88] is used to provide visual speech features. FAPs describe facial movement and are used in the MPEG-4 audiovisual object-based video representation standard to control facial animation, together with the so-called facial definition parameters that describe the face shape. There are 68 FAPs, divided into ten groups, depending on the particular region of the face that they are located (see also Fig. 4). Of particular interest to visual speech applications are the "group 8" parameters, which describe outer lip contour movement [40]. Additional speech information is contained in "group 2" parameters, which correspond to inner lip and jaw motion; "group 6" ones, which describe the tongue; and less so, in cheek movement captured by "group 5" FAPs.

Clearly, appearance- and shape-based visual features are quite different in nature, coding low- and high-level information about the speaker's face and lip movements. Not surprisingly, combinations of features from both categories have been suggested in the literature. In most cases, features of each type are just concatenated, as in [30, 31], where PCA appearance features are combined with snake-based features

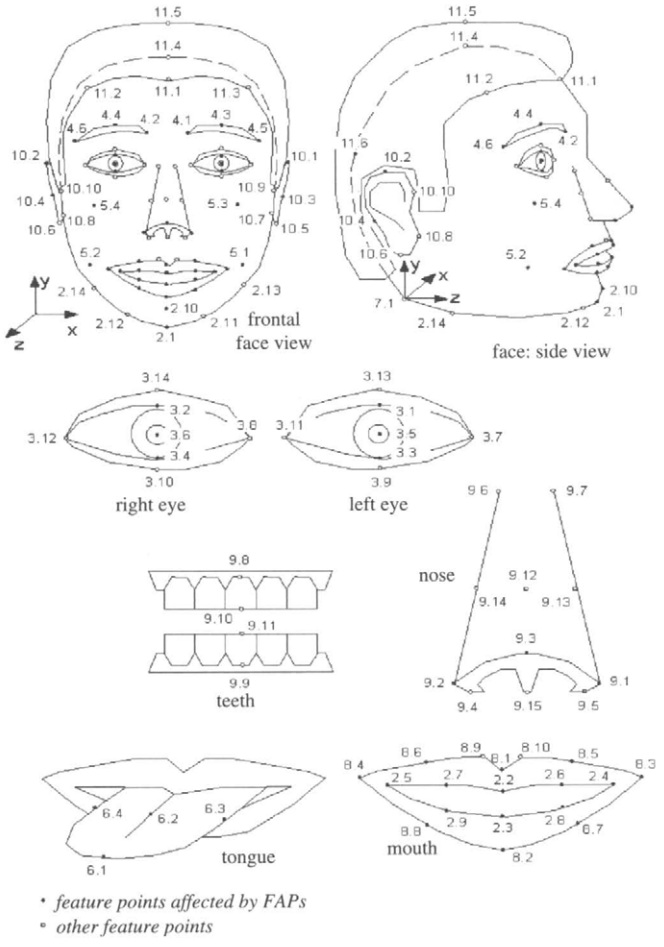


FIGURE 4 The facial animation control points supported by the MPEG-4 video representation standard [88]. Facial animation parameters (FAPs) describe the movement of 68 of these control points. There are ten FAP groups, with groups 8, 2, 6, and 5 being of interest in shape-based visual speech feature extraction [40, 52, 73].

or ASMs, respectively. A different approach to combining the two classes of features is to create a single model of face shape and appearance using the AAM [86], discussed earlier. The final model PCA can be applied on the vector of the tracked shape and its corresponding appearance representations to provide a set of visual features [29]. Finally, it is interesting to note that features from both categories can be used in a hierarchic manner. For example, in the visual text-to-speech synthesis reported in [54], visual unit selection occurs on the basis of the appearance representation of candidate mouth shapes within a set determined by their geometric shape features (see also Section 5.3).

In typical speech-based HCI, visual features are used with audio features obtained from the acoustic waveform. Such features, for example, could be mel-frequency cepstral coefficients (MFCCs) or linear prediction coefficients (LPCs), and are mostly extracted at a 100-Hz rate [1, 87]. In contrast, visual features are generated at the much lower video frame or field rate. They can however be easily postprocessed (up-sampled) by linear interpolation to achieve audiovisual feature synchrony at the audio rate and thus simplify audiovisual integration as discussed in Section 3 [44]. In addition to interpolation, a number of visual feature postprocessing methods play a critical role in enhancing the performance of visual speech processing systems. The most important such techniques concern capturing the visual speech dynamics. Similar to audio-only systems, this can be achieved by augmenting the "static" (frame-based) visual feature vector by its first- and second-order derivatives, which are computed over a short temporal window centered at the current video frame [87]. Alternatively, a "dynamic" feature vector can be obtained by training an LDA matrix to project the concatenation of neighboring visual feature vectors onto a lower-dimensional space [44]. LDA can also be followed by a feature space rotation matrix (MLLT) to improve statistical modeling of the extracted features [44]. Mean normalization of the visual feature vector can also contribute to improved performance, by reducing variability due to illumination, for example. Finally, feature selection within a larger pool of candidate features can also be considered as a form of post-processing. A case of such selection for automatic speechreading appears in [37].

In summary, a number of approaches are viable for extracting and representing visual speech information. Unfortunately however, limited work exists in the literature in comparing their relative performance. Most such comparisons are in the context of automatic speechreading and audiovisual ASR, where features within the same category (appearance- or shape-based) are usually investigated [23, 24, 27, 29, 37]. Occasionally, features across categories are compared, but in most cases with inconclusive results [24, 29, 30, 45]. Thus, the question of what are the most appropriate visual speech features that are sufficiently speaker-independent and robust to visual environment and head-pose variation, remains to a large extent unresolved. Nevertheless, as the results in subsequent sections demonstrate, the specific implementations of both appearance- and shape-based systems, which are considered in this chapter and reviewed next, suffice to benefit a number of speech-related HCI technologies under somewhat constrained visual conditions. In practice, factors such as computational requirements, video quality, and the visual environment could determine the most suitable approach in a particular application.

2.3 Two Visual Feature Extraction Systems

In this chapter, we will be further considering two particular implementations of visual feature extraction when reporting audiovisual speech processing results. The first is the appearance-based system developed at IBM Research. The system is depicted in Fig. 5A, in parallel with its complementary audio processing module, as used for providing time-synchronous bimodal feature vectors for audiovisual ASR

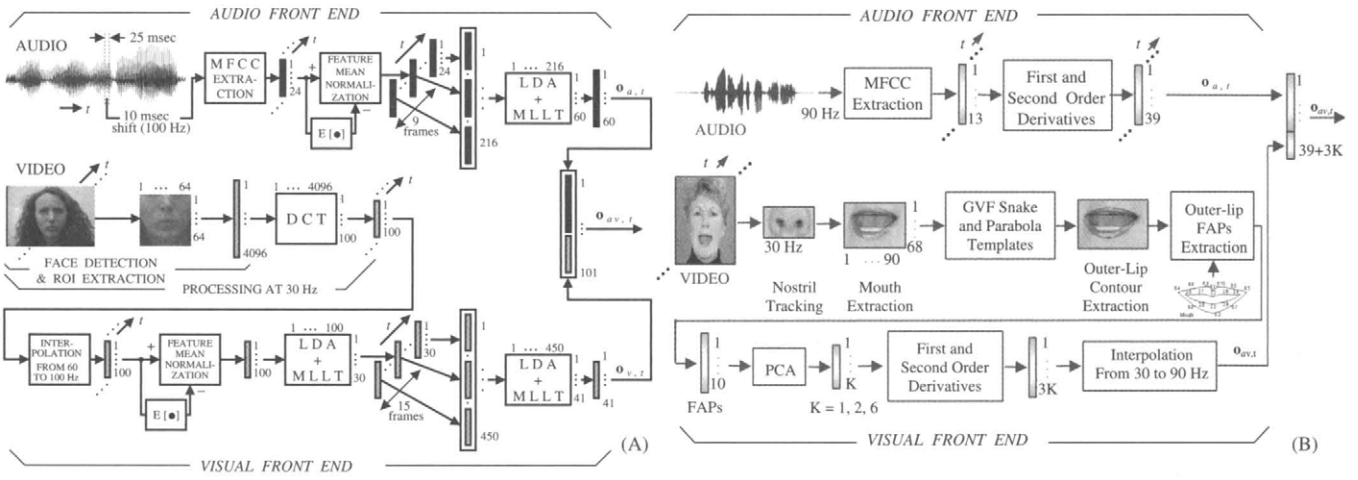


FIGURE 5 Two implementations of visual feature extraction, depicted schematically in parallel with the audio front end, as used for audiovisual automatic speech recognition experiments in this chapter: **A**: The appearance-based visual front end system of IBM Research, also employed for bimodal speaker recognition [72] and audio enhancement [19]; **B**: the shape-based system of Northwestern University [40], also used for speech-to-video synthesis [52] and audiovisual speaker recognition [73]. (See color insert.)

[44]. With minor modifications, it is also used for audiovisual speaker recognition [72] and noisy audio feature enhancement assisted by the visual observations [19]. Given the video of the speaker's face, the system first detects the face and 26 facial landmark points using the statistical tracking algorithm of [79], thus allowing the extraction of a normalized 64×64-pixel gray-scale ROI (see also Fig. 2A). A two-dimensional, separable DCT is subsequently applied on the ROI vector, and the top 100 coefficients (in terms of energy) are retained. The feature vector dimensionality is further reduced to 30 by means of an intraframe LDA/MLLT. Following some of the postprocessing steps discussed above, a 41-dimensional dynamic visual speech vector $\mathbf{o}_{v,t}$ is extracted at each time instant t at a 100-Hz rate, synchronized with 60-dimensional MFCC-based audio features $\mathbf{o}_{a,t}$.

The second system, developed at Northwestern University (NWU), is shape-based and uses a set of FAPs [88] as visual features (see Fig. 5B). The system first uses a template to track the speaker's nostrils, thus determining the approximate mouth location. Subsequently, the outer lip contour is tracked using a combination of a GVF and a parabolic template (see also Fig. 2B). Following the outer lip contour detection and tracking, ten FAPs describing the outer lip shape ("group 8" FAPs [88]) are extracted from the resulting lip contour (see also Figs. 3 and 4). These are placed into a feature vector which is subsequently projected by means of PCA onto a two-dimensional space [40]. The resulting visual features are augmented by their first and second derivatives providing a six-dimensional dynamic visual speech vector $\mathbf{o}_{v,t}$. These features are interpolated to the 90-Hz frame rate of 39-dimensional, MFCC-based audio features [87]. The combined features are used for a number of audiovisual speech applications such as ASR, speech-to-video synthesis, and speaker-recognition [40, 52, 73].

3 Audiovisual Information Fusion

The second critical issue in the design of audiovisual speech processing systems is the integration of the available modality representations. To justify the complexity and cost of incorporating the visual modality into HCI, integration strategies should ensure that the performance of the multimodal system exceeds that of its single-modality counterpart, hopefully by a significant amount. For example, one would expect that the transcription accuracy of an audiovisual ASR system greatly surpasses that of the audio-only system, especially in noisy environments, or that audiovisual TTS is perceived as more friendly, intelligible, and natural than a synthetic voice-only system in subjective evaluation tests. In this section, we review the main concepts and techniques that are essential to successful audiovisual integration in speech-based HCI.

In this chapter, we are interested in a number of diverse bimodal technologies, with main emphasis on ASR, text-to-speech, and speaker recognition. Clearly, the characteristics and requirements of each technology differ significantly, therefore it is natural that, among them, so do the modality integration methods. Nevertheless, a number of themes are similar in at least some of the technologies, thus allowing a common framework in their review. For example, central to ASR, text-to-speech, and text-dependent speaker recognition algorithms is the notion of speech classes underlying the acoustic and visual representations. Of course, different types

of classes are required for a number of other audiovisual applications, such as the general speaker recognition problem, emotion detection, audiovisual localization, and so forth. The second common theme across the technologies of interest is the issue of combining the acoustic and visual feature streams using classifiers designed to outperform their single-modality counterparts. The choice of classifiers and algorithms for feature and classifier fusion are clearly central to the design of audiovisual ASR and speaker recognition systems, among others. Finally, techniques for exploiting the correlation between the two signals are also of interest, and in this chapter are considered in the context of speech-to-video synthesis, discussed in Section 5.4.

3.1 Speech Classes in Audiovisual Integration

The basic unit that describes how speech conveys linguistic information is the phoneme. For American English, there are approximately 42 such units [89], generated by specific positions or movements of the vocal tract articulators. However, since only a small part of the vocal tract is visible, not every phoneme pair can be disambiguated by the video information alone. The number of visually distinguishable units is therefore much smaller. Such units are referred to as *visemes* in the audiovisual speech processing and human perception literature [9, 10, 16].

Importantly, visemes capture "place" of articulation information [14, 16], that is, they describe where the constriction occurs in the mouth and how mouth parts, such as the lips, teeth, tongue, and palate, move during speech articulation. As a result, many consonant phonemes with identical "manner" of articulation, which are therefore difficult to distinguish on the basis of acoustic information alone, may differ in the place of articulation, and thus be visually identifiable; for example, the two nasals "m" (a bilabial) and "n" (an alveolar). In contrast, phonemes "m" and "p" are easier to perceive acoustically than visually, since they are both bilabial, but differ in the manner of articulation, instead.

Various mappings between phonemes and visemes can be found in the literature. In general, they are derived by human speech-reading studies, but they can also be generated using statistical clustering techniques [37]. There is no universal agreement about the exact grouping of phonemes into visemes, although some clusters are well-defined; for example, the bilabial group {"p", "b", "m"}. All its three members are articulated at the same place (lips), thus appearing visually the same. A particular phoneme-to-viseme grouping is depicted in Table 1 [44].

In audio-only speech applications, the set of classes of interest in technologies such as ASR, text-dependent speaker recognition, and TTS most often consist of subphonetic units. Such classes are designed by clustering the possible phonetic contexts (tri-phones, for example) by means of a decision tree, to allow coarticulation modeling [87, 89]. Occasionally, subword units are used in specific, small-vocabulary tasks. Naturally therefore, in visual speech applications one could consider visemic subphonetic classes, obtained for example by decision tree clustering based on visemic context. Indeed, visemic classes have been occasionally used in ASR [32, 43], and of course play a central role in visual synthesis systems (see Section 5). However, the use of different classes for the audio and visual components complicates audiovisual integration, especially in ASR and text-dependent speaker-recognition. For such applications, identical classes are used for both speech modalities, most often subphonetic classes.

TABLE 1 A 42-phoneme to 12-viseme mapping of the HTK phone set [87]

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

3.2 Classifiers in Speech Applications

In typical speech technologies discussed in this chapter, the classes of interest are hidden. We denote such unknown classes by $c \in \mathcal{C}$. For example, in the case of ASR, $\mathcal{C}$ represents a set of subphonetic or subword units, as discussed above. Classification of a sequence of such units gives rise to recognized words, based on a phonetic dictionary for the ASR task vocabulary. In the synthesis systems discussed in Section 5, set $\mathcal{C}$ can contain all candidate concatenative units, or describe a set of quantized representations of the signal to be synthesized. For speaker identification, $\mathcal{C}$ corresponds to the enrolled subject population, possibly augmented by a class denoting the unknown subject, whereas for authentication, $\mathcal{C}$ reduces to a two-member set. In the particular case of text-dependent speaker recognition, $\mathcal{C}$ can be considered as the product space between the set of speakers and the set of phonetic based units.

The hidden classes are observed only through the signal representation, namely a series of extracted feature vectors. We denote such vectors by $\mathbf{o}_{s,t}$, and their sequence over an interval $\mathcal{T}$ by $\mathbf{O}_s = \{\mathbf{o}_{s,t}, t \in \mathcal{T}\}$, where $s \in \mathcal{S}$ denotes the available modality; for example $\mathcal{S} = \{a, v, f\}$, in the speaker-recognition system of [74], that is based on audio, visual-labial, and face-appearance input.

A number of methods can then be used to model the association between the unknown classes and the observed feature vectors. Most such approaches are statistical in nature and provide a conditional probability measure for $Pr(\mathbf{o}_{s,t}|\, c)$ or $Pr(c\,|\,\mathbf{o}_{s,t})$; for example, artificial neural networks (ANNs), used for automatic speechreading in [22, 23, 33] and visual speech synthesis in [51, 61] or support vector machines (SVMs), as in [43]. Alternatively, the space of possible observation vectors is discretized through the process of vector quantization (VQ), as in [39, 62]. Then, the statistical model provides conditional probabilities of the form $Pr(q(\mathbf{o}_{s,t})\,|\,c)$, where $q(\bullet)$ belongs to a discrete set of codebooks.

In most practical cases though, a Gaussian mixture density is assumed, namely

$$Pr(\mathbf{o}_{s,t}|\, c) = \sum_{k=1}^{K_{s,c}} w_{s,c,k}\, \mathcal{N}(\mathbf{o}_{s,t}; \mathbf{m}_{s,c,k}, \mathbf{s}_{s,c,k}), \tag{1}$$

resulting in the Gaussian mixture model (GMM) classifier. In (1), $K_{s,c}$ denotes the number of mixture weights $w_{s,c,k}$, which are positive and add to one, and $\mathcal{N}(\mathbf{o}; \mathbf{m}, \mathbf{s})$ represents a multivariate normal distribution with mean $\mathbf{m}$ and a covariance matrix $\mathbf{s}$, typically considered as diagonal. Emission probability model (1) is therefore described by parameter vector

$$\mathbf{b}_s = [\{[w_{s,c,k}, \mathbf{m}_{s,c,k}, \mathbf{s}_{s,c,k}], k = 1, \ldots, K_{s,c},\ c \in \mathcal{C}\}], \tag{2}$$

for a particular modality s.

This model is sufficient to address problems where a single underlying class c is assumed to generate the entire observation sequence $\mathbf{O}_s$, and conditional independence of the observations holds. This is the case in most text-independent speaker recognition systems, for example. The model then allows maximum-a-posteriori estimation of the unknown class, as

$$\hat{c} = \arg\max_{c \in \mathcal{C}}\ Pr(c) \prod_{t \in \mathcal{T}} Pr(\mathbf{o}_{s,t}|\, c), \tag{3}$$

where $Pr(c)$ denotes the class prior.

Model (3) is however inappropriate for applications where a temporal sequence of interacting states is assumed to generate the series of observations, as is the case in ASR, speech synthesis, and text-dependent speaker recognition. There, hidden Markov models (HMMs) are widely used. In generating the observed sequence in modality s, the HMM assumes a sequence of hidden states sampled according to the transition probability parameter vector $\mathbf{a}_s = [\{Pr(c'|c''),\ c', c'' \in \mathcal{C}\}]$. The states subsequently "emit" the observed features with class-conditional probability given by (1). The HMM parameter vector $\mathbf{p}_s = [\mathbf{a}_s, \mathbf{b}_s]$ is typically estimated iteratively, using the expectation-maximization (EM) algorithm [87, 89], as

$$\mathbf{p}_s^{(j+1)} = \arg\max_{\mathbf{p}} Q(\mathbf{p}_s^{(j)}, \mathbf{p}|\, \mathbf{O}_s),\ j = 0, 1, \ldots. \tag{4}$$

In (4), $\mathbf{O}_s$ consists of all feature vectors in the training set, and $Q(\bullet, \bullet|\bullet)$ represents the EM algorithm auxiliary function, defined as in [89]. Alternatively, discriminative training methods can be used [89]. Once the model parameters are estimated, HMMs can be used to obtain the hidden classes of interest, also known as the "optimal state sequence" $\mathbf{c} = \{c_t,\ t \in \mathcal{T}\}$, given an observation sequence $\mathbf{O}_s$ over interval $\mathcal{T}$; namely,

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c} \in \mathcal{C}^{|\mathcal{T}|}} Pr(\mathbf{O}_s, \mathbf{c})\,, \tag{5}$$

where

$$Pr(\mathbf{O}_s, \mathbf{c}) = \prod_{t \in \mathcal{T}} Pr(c_t|\, c_{t-1})\, Pr(\mathbf{o}_{s,t}|\, c_t). \tag{6}$$

In practice, the Viterbi algorithm is used for solving (5), based on dynamic programming [87, 89].

3.3 Feature and Classifier Fusion

The above presentation assumes that only one observation stream, $\mathbf{o}_{s,t}$, is provided. In practical audiovisual speech applications though, multiple streams are available, which result in multimodal observations $\mathbf{o}_t = \{\mathbf{o}_{s,t}, s \in \mathcal{S}\}$, assuming time-synchronous stream feature representations; for example, in the case of audiovisual ASR, $\mathbf{o}_{av,t} = [\mathbf{o}_{a,t}, \mathbf{o}_{v,t}]$. As already mentioned, integrating such multimodal information into systems that outperform their single-modality counterparts constitutes a major focus of audiovisual speech research.

Indeed, various information fusion algorithms have been considered in the literature, differing both in their basic design and in the terminology used [8, 21, 35, 44, 47]. In this paper, we adopt a broad grouping of such techniques into feature fusion and decision fusion methods [44]. The first are based on training a single classifier on the multimodal feature vector $\mathbf{o}_t$, or on any appropriate transformation of it [34, 35, 44]. In contrast, decision fusion algorithms utilize each single-modality classifier output to jointly estimate the hidden classes of interest. Typically, this is achieved by linearly combining the class-conditional observation log-likelihoods of the individual classifiers into a joint audiovisual classification score, using appropriate weights that capture the reliability of each single-modality classifier or data stream [21, 28, 31–33]. The two approaches are schematically depicted in Fig. 6, in the case of one observation stream available for each of the audio and visual modalities.

Audiovisual feature fusion techniques include plain feature concatenation [34], feature weighting [35, 47], both also

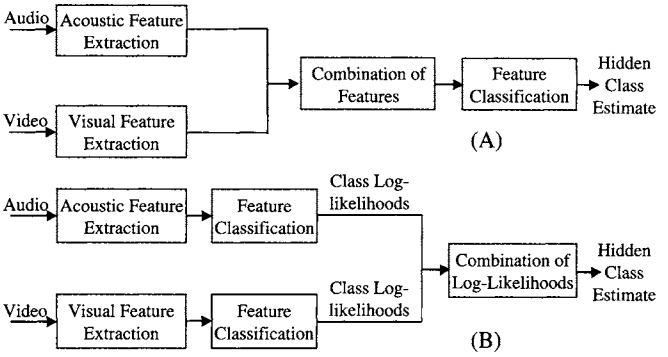


FIGURE 6 Block diagrams of the (**A**) feature fusion and (**B**) decision fusion approaches to audiovisual integration.

known as direct identification fusion [35], as well as the "dominant" and "motor" recording fusion [35]. The latter seek a data-to-data mapping of 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 the basis of audiovisual features (for example, using regression, as in [18]) also falls within this category of fusion. Another interesting feature fusion technique, proposed for audiovisual ASR in [44], seeks a discriminant projection of the concatenated bimodal vector $\mathbf{o}_{av,t}$ onto a lower-dimensional space for improved statistical modeling. The projected vector $\mathbf{o}_{d,t} = \mathbf{o}_{av,t}\mathbf{P}_{av}$ is modeled using the single-stream HMM of (1) and (6), where $\mathbf{P}_{av}$ is a cascade of an LDA projection and MLLT rotation (see also Section 2.3).

Although many feature fusion techniques result in improved system performance [44], they cannot explicitly model the reliability of each modality. Such modeling is extremely important due to the varying speech information content of the audio and visual streams. The decision fusion framework, on the other hand, provides a mechanism for capturing these reliabilities, by borrowing from classifier combination theory, an active area of research with many applications [90].

Various classifier combination techniques have been considered for audiovisual speech applications, including for example a cascade of fusion modules, some of which possibly using only rank-order classifier information about the hidden classes of interest [20, 32]. However, by far the most commonly used decision fusion techniques belong to the paradigm of classifier combination using a parallel architecture, adaptive combination weights, and class score level information. These methods derive the most likely hidden class by linearly combining the log-likelihoods of the single-modality classifier decisions, using appropriate weights [28, 31, 34–36]. This corresponds to the adaptive product rule in the likelihood domain [90], and it is also known as the separate identification model for audiovisual fusion [32, 35].

In the most common application of this approach to audiovisual speech systems, the combination occurs at the observation frame level, resulting in the multimodal class-conditional

$$Pr(\mathbf{o}_t|\, c) = \prod_{s\in\mathcal{S}} Pr(\mathbf{o}_{s,t}|\, c)^{\lambda_{s,c,t}}, \tag{7}$$

for all hidden classes $c \in \mathcal{C}$. Notice that (7) does not represent a probability distribution in general, and should be viewed as a "score", when used in conjunction with (3) and (5). In (7), $\lambda_{s,c,t}$ denote the stream exponents (weights), that are non-negative, and model stream reliability as a function of modality s, state c, and utterance frame (time) t. These are typically constrained to sum to one or $|\mathcal{S}|$, and are often set to global, modality-only dependent values, $\lambda_s \leftarrow \lambda_{s,c,t}$, for all c and t.

Joint model (7) can be used, for example, in audiovisual speaker recognition with the GMM of (1) and (3), as in [72, 73], as well as for audiovisual ASR [25, 28, 31], resulting in the so-called multistream HMM [see also (1) and (6)]. Notice that (7) also provides a framework to incorporate feature fusion; for example, in [44], the discriminant feature vector $\mathbf{o}_{d,t}$ is used as one of two or three streams for audiovisual ASR together with audio and possibly visual features. The approach is referred to as "hybrid" fusion.

Training the parameters of (7) requires additional steps, compared to (4). For example, in the particular case of two observation streams (audio and visual), each modeled by a single-stream HMM classifier with identical set of classes, the multistream HMM parameter vector becomes [see also (1), (2), and (7)]

$$\overline{\mathbf{p}}_{av} = [\mathbf{p}_{av}, \lambda_a, \lambda_v], \quad \text{where} \quad \mathbf{p}_{av} = [\mathbf{a}_{av}, \mathbf{b}_a, \mathbf{b}_v].$$

This consists of the HMM transition probabilities $\mathbf{a}_{av}$ and the emission probability parameters $\mathbf{b}_a$ and $\mathbf{b}_v$ of its single-stream components. The parameters of $\mathbf{p}_{av}$ can be estimated separately for each stream component using the EM algorithm, namely (4) for $s = a, v$, and subsequently, by possibly setting the joint HMM transition probability vector equal to the audio-one, i.e., $\mathbf{a}_{av} = \mathbf{a}_a$. The alternative is to jointly estimate parameters $\mathbf{p}_{av}$, to enforce state synchrony in training. In the latter scheme, the EM-based parameter reestimation becomes [87]

$$\mathbf{p}_{av}^{(j+1)} = \arg\max_{\mathbf{p}} Q(\overline{\mathbf{p}}_{av}^{(j)}, \mathbf{p}|\, \mathbf{O}_{av})$$

[see also (4)]. The two approaches thus differ in the E-step of the EM algorithm. In both separate and joint HMM training, in addition to $\mathbf{p}_{av}$, the stream exponents λ_a and λ_v need to be obtained. This can be performed using discriminative training methods, simple parameter search on a grid, or mappings of signal quality measures to exponent values [25, 28, 33–35, 44].

Finally, of particular interest to audiovisual ASR is the level at which the stream log-likelihoods are combined. The use of

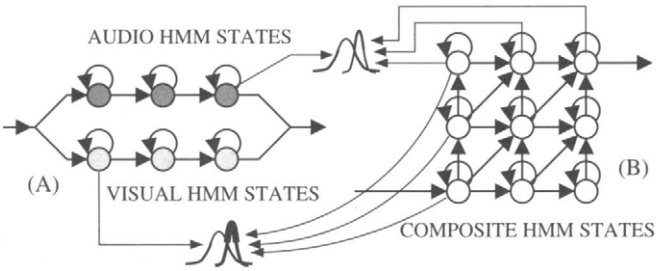


FIGURE 7 **A**: Phone-synchronous two-stream hidden Markov model (HMM) with three states per phone and modality. **B**: Its equivalent product HMM; the single-stream emission probabilities are tied for states along the same row (column) to the corresponding audio (visual) state probabilities of form (1).

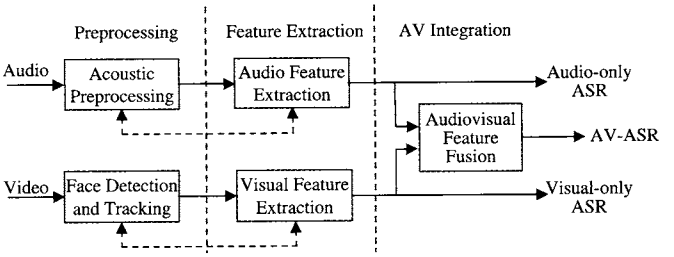


FIGURE 8 Block diagram of an audiovisual automatic speech recognition system.

HMMs allows likelihood recombination at a coarser level than the HMM state, for example at the phone or word boundary. Product or coupled HMMs are typically used for the task, as in [26, 31, 44]. Such models allow state-level asynchrony between the acoustic and visual observations within the phone or word, forcing their synchrony at the unit boundaries instead. Product HMMs consist of composite audiovisual states, as depicted in Fig. 7, thus resulting in a much larger state space compared to multistream models fused at the state level, as in (7). To avoid undertraining, the single-stream emission probability components of the observation class-conditionals are tied along identical visual and audio states (see also Fig. 7).

Audio and visual recognition log-likelihoods can also be combined at the utterance level. This approach can easily be applied on small-vocabulary tasks, where likelihoods can be calculated for each word, based on the acoustic and visual observations. However, the number of possible hypotheses on large-vocabulary and continuous speech recognition tasks becomes prohibitively large. In such cases, recombination is usually limited to N-best hypotheses, generated either by the audio-only system or obtained as the union of audio- and visual-only N-best hypotheses. These are then rescored by combining the log-likelihoods generated using audio and visual HMMs [34, 36].

4 Audiovisual Automatic Speech Recognition

As discussed in the Introduction, visual speech plays an important role in human speech perception, improving speech intelligibility especially in noise [9–11]. A number of reasons were cited there, the most important being the fact that visual speech information contains both correlated and complementary information to the acoustic signal. Not surprisingly, such information can be beneficial to automatic speech recognition as well. Incorporating visual speech information into ASR is generally viewed as a very promising approach for improving speech recognition robustness to noise [1, 2] and bridging the gap between human and automatic performance [3]. Naturally therefore, significant research has recently focused in this area.

In 1984, Petajan [20] developed the first audiovisual ASR system. He used image thresholding to obtain binary mouth images from the input video, which were subsequently analyzed to derive mouth height, width, perimeter, and area, to be used as visual features in speech recognition. He first reported visual-only speech recognition results of isolated words within a 100-word vocabulary, using dynamic time warping [89]. In addition, he combined the acoustic and visual speech recognizers in a serial fashion to improve ASR performance: The visual speech system was used to rescore several N-best word hypotheses, as obtained by audio-only ASR, to generate the final bimodal recognition result.

A number of researchers have developed audiovisual ASR systems since [20–47]. Their systems vary in a number of areas, which have been discussed in detail in Sections 2 and 3. In summary, variations can be found in: the visual front end design, with some works adopting appearance-based features [22–31], whereas other researchers considering shape-based techniques [30–39], or even combinations of the two approaches [29–31]; the choice of classes used in the recognition process, for example subphonetic [29, 40, 44], subword [28, 31, 44], or viseme-based [32, 43]; the employed recognition method, such as ANNs [22, 23, 33], SVMs [43], simple weighted distances used with VQ [20], and HMMs with various emission probability models [29, 31, 36, 39, 44]; and finally, the approach of integrating the audio and visual observation streams, generally grouped into feature fusion [34, 35, 44, 47] and decision fusion methods [25, 26, 28, 31–36]. Overall, the reported bimodal systems show improved performance compared to audio-only ASR for the recognition tasks considered: Such are typically small-vocabulary tasks, for example isolated words [31], connected digits [28], or closed-set sentences [37], with large-vocabulary tasks recently reported [40, 44].

In the remainder of the section, we briefly review corpora commonly used for audiovisual ASR research, and we present experimental results on some of them using the IBM and NWU systems, previously discussed in Section 2.3. These results clearly demonstrate the benefit of incorporating the visual modality into ASR.

4.1 Bimodal Corpora for Automatic Speech Recognition

In contrast to the abundance of audio-only corpora, there are only a few databases suitable for audiovisual ASR research. This is because the field is relatively young but also because audiovisual corpora pose additional challenges concerning database collection, storage, distribution, and privacy. Most commonly used databases in the literature are the product of efforts by few university groups or individual researchers with limited resources, and as a result, they contain a small number of subjects, have relatively short duration, and mostly address simple recognition tasks, such as small-vocabulary ASR of isolated or connected words [8, 21]. Examples of such popular datasets in audiovisual ASR research are the Clemenson University audiovisual experiments (CUAVE) corpus containing connected digit strings [36], the Advanced multimedia processing lab, Carnegie Mellon University (AMP/CMU) database of 78 isolated words [47], the Tulips1 set of four isolated digits [27], and the digit portion of the extended Multi-Modal verification for teleservices and security applications ((X)M2VTS) corpora, more often used in speaker recognition experiments (see Section 6). Additional datasets are suitable for recognition of isolated nonsense words consisting of vowel–consonant combinations [34], connected letter strings in English [28] and German [22, 23], as well as continuous large-vocabulary speech [37] (see also [40, 91]).

A number of proprietary corpora have also been recorded by many groups, including recent work at IBM Research [44, 46]. There, a number of databases have been collected containing large subject populations (50–290 subjects), uttering both large-vocabulary speech and connected-digit strings [46]. The corpora have been recorded in four different audiovisual conditions to benchmark the performance of the IBM appearance-based visual front end. Three of the sets contain frontal full-face videos and correspond to increasingly more challenging visual domains: The first was collected in a quiet studiolike environment, using a high-quality camera, uniform lighting and background, and relatively stable frontal subject head pose. The second corpus was recorded using a portable collection system on a laptop, with quarter-frame resolution video captured via an inexpensive Webcam and audio by the built-in personal computer (PC) microphone. The database subjects were typically recorded in their own offices with varying lighting, background, and head pose. The third set was recorded in an automobile, both stationary and moving at approximately 30 or 60 mph, that was equipped with a wideband microphone and a lipstick-style camera. Compared with the previous two databases, the lighting, background, and head pose vary significantly, therefore this database represents the most challenging set. Finally, to study the benefits of direct visual ROI capture, a fourth set was recorded by means of a specially designed audiovisual wearable headset with an infrared camera housed inside its boom. This device provides high-quality visual data of the mouth ROI, being relatively insensitive to head-pose and lighting variations [46]. The video frame rate of all corpora is 30 Hz, and—with the exception of the office data—the resolution is 704 × 480 pixels. Typical frames of all four sets are depicted in Fig. 9.

FIGURE 9 Example frames from the four IBM audiovisual automatic speech recognition corpora discussed in Sections 4.1 and 4.2. **Top-to-bottom:** Full-face data collected in the studiolike, office, and car environments; **bottom line:** Region of interest–only data captured by a specially designed headset [46]. (See color insert.)

4.2 Experimental Results

We now proceed to experimentally demonstrate the benefit of visual speech to ASR. In the first set of experiments, the IBM appearance-based audiovisual ASR system (see also Fig. 5A) with two-stream HMM-based decision fusion is applied to the four corpora depicted in Fig. 9. A number of connected-digits recognition results are reported in Table 2 in terms of word error rate (WER), percentage (%), using a multis-peaker training-testing scenario. In addition to visual-only recognition, audio-only and audiovisual ASR results are depicted for two acoustic conditions: the original recorded audio and

TABLE 2 Connected-digit recognition on the four IBM databases of Fig. 9 [46]

		Clean			Noisy		
Database	VI	AU	AV	%	AU	AV	%
Studio	27.44	0.84	0.66	21	24.56	10.66	58
Office	43.33	2.51	1.96	22	24.91	14.73	41
Car	68.75	2.83	2.38	16	25.89	16.22	37
Headset	21.35	1.33	0.94	29	25.23	7.92	69

Audio-visual (AV) vs. audio-only (AU) word error rate (WER), % , is depicted for clean and artificially corrupted data using HMMs trained on clean data. The approximate% relative improvement due to the visual modality is also shown for each condition (%), as well as the visual-only (VI) WER.

the artificially corrupted audio by nonstationary babble speech noise. The noise level varies per database, with the experiment designed to result in audio-only WER of about 25% for all four corpora. In addition to the WER results, the approximate relative percentage reduction in WER, achieved by incorporating the visual modality into ASR, is shown for both acoustic conditions.

Table 2 demonstrates two major points [46]: First, that the ASR gains due to visual speech are large, even for the relatively clean acoustic conditions of the original data. Such benefits become dramatic at high noise levels, reaching for example a relative 69% WER reduction for the headset data. It is interesting to note that these gains hold even though the visual-only performance is significantly worse than audio-only ASR (for example, 15 to 25 times worse in WER for the particular tasks). Second, as the visual environment becomes more challenging, due to head-pose and lighting variation, both visual-only performance and ASR gains degrade. For example, the visual-only WER is only 21.3% for the headset corpus, but 68.7% in the automobile data. Clearly, under challenging visual conditions, the performance of appearance-level visual features suffers.

The above observations carry through to large-vocabulary ASR as well. This is partially demonstrated in Fig. 10A, where speaker-independent, large vocabulary (>10,000 words) continuous speech recognition results are depicted for the studio-quality database using three fusion techniques for audiovisual ASR over a wide range of acoustic signal-to-noise-ratio (SNR) conditions. The best results are obtained by a hybrid fusion approach that uses the two-stream HMM (AV-MS) framework to combine audio features with fused audiovisual discriminant features (AV-Discr.), achieving for example an 8-dB "effective SNR" performance gain at 10 dB, as depicted in Fig. 10A [44].

Similar conclusions are reached when using the NWU audiovisual ASR system that uses shape-based visual features, obtained by PCA on FAPs of the outer and inner lip contours [40, 45] (see also Section 2.2). A summary of single-speaker, large-vocabulary ($\approx$ 1,000 words) recognition experiments using the Bernstein lipreading corpus [91] is depicted in Fig. 10B. There, audio-only WER, %, is compared to audiovisual ASR performance over a wide range of acoustic SNR conditions (0–30 dB), obtained by corrupting the original signal with white Gaussian noise. It can be clearly seen in Fig. 10B that considerable ASR improvement is achieved, compared to the audio-only performance, for all noise levels tested when visual speech information is utilized. Of particular interest is to compare such gains when using FAPs extracted from the inner and outer lips. Figure 10B demonstrates that inner-lip FAPs, when used as visual features do not provide as much speechreading information as the outer-lip FAPs. However, when both inner- and outer-lip FAPs are used as visual features, the performance of the audiovisual ASR system improves as compared with when

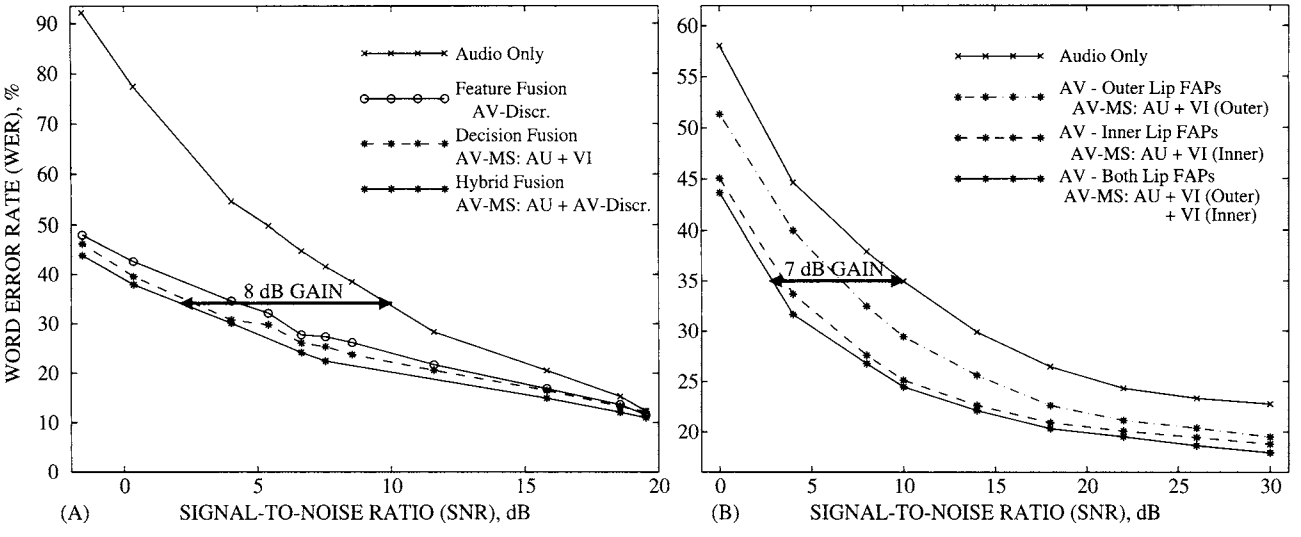


FIGURE 10 Large-vocabulary, audiovisual automatic speech recognition results using the IBM (**left**) and Northwestern University (NWU) (**right**) systems. In both cases, audio-only and audiovisual word error rate (WER), %, are depicted vs. audio channel signal-to-noise ratio (SNR) for hidden Markov models (HMMs) trained in matched noise conditions. The effective SNR gains are also shown with reference to the audio-only WER at 10 dB. Notice that the axes ranges in the two plots differ. In more detail: (**A**) In the IBM system, appearance-based visual features are combined with audio using three different techniques. Reported resuts are on the IBM, studio-quality database [44]. (**B**): In the NWU system, shape-based visual features using outer-, inner-only, or both lip-contour facial animation parameters (FAPs) are combined with audio features by means of decision fusion. Reported resuts are on the Bernstein lip-reading corpus [91].

only the outer-lip FAPs are used [40]. Note that these results are consistent with investigations of inner versus outer lip geometric visual features for automatic speechreading [24].

5 Audiovisual Speech Synthesis

Audiovisual speech synthesis is a topic at the intersection of a number of areas including computer graphics, computer vision, image and video processing, speech processing, physiology, and psychology. Audiovisual speech synthesis systems automatically generate either voice and facial animation from arbitrary text (audiovisual TTS, or visual TTS (VTTS)), or facial animation from arbitrary speech (speech-to-video synthesis). A view of an animated face, be it text- or speech-driven, can significantly improve intelligibility of both natural and synthetic speech, especially under nonideal acoustic conditions. Moreover, facial expressions and prosodic information can signal emotions, add emphasis to speech, and support dialog interaction.

Audiovisual speech synthesis systems have numerous applications related to human communication and perception, including tools for the hearing impaired, multimodal virtual agent–based user interfaces (desktop assistants, e-mail messengers, newscasters, online shopping agents, etc.), computer-based learning, net-gaming, advertising, and entertainment. For example, facial animation generated from telephone speech by a speech-to-video synthesis system could greatly benefit the hearing impaired, while an e-mail service that transforms text and emoticons (facial expressions coded into a certain series of keystrokes) from text into an animated talking face could personalize and improve the e-mail experience. Audiovisual speech synthesis is also suitable for wireless communication applications. Indeed, some face animation technologies have very low-bandwidth transmission requirements, utilizing a small number of animation control parameters. New mobile technology standards allow large-bandwidth multimedia applications, thus enabling the transmission of full synthetic video, if desired.

Two critical topics in the design and performance of audiovisual speech synthesis systems are modeling the speech coarticulation and the animation of the face. Various approaches exist in the literature for addressing these issues, and are presented in detail in the next two sections. Following their review, VTTS and speech-to-video synthesis are discussed, and evaluation results of the visual speech synthesis system developed at NWU are presented.

5.1 Coarticulation Modeling

Coarticulation refers to changes in speech articulation (acoustic or visual) of the current speech segment (phoneme or viseme) due to neighboring speech. In the visual domain, this phenomenon arises because the visual articulator movements are affected by the neighboring visemes. Addressing this issue is crucial to visual speech synthesis, since, to achieve realistic facial animation, the dynamic properties and timing of the articulatory movements need to be proper. A number of methods have been suggested in the literature to model coarticulation. In general, they can be classified into rule-based and data-based approaches and are reviewed next.

Techniques in the first category define rules to control the visual articulators for each speech segment of interest, which could be phonemes or bi- or tri-phones. For example, Löfquist [92] proposed an "articulatory gesture" model. He suggested utilizing dominance functions, defined for each phoneme, which increase and decrease over time during articulation, to model the influence of the phoneme on the movement of articulators. Dominance functions corresponding to the neighboring phonemes will overlap, therefore, articulation at the current phoneme will depend not only on the dominance function corresponding to the current phoneme, but also on the ones of the previous and following phonemes. In addition, it is proposed that each phoneme has a set of dominance functions, one for each articulator (lips, jaw, velum, larynx, tongue, etc.), because the effect of different articulators on neighboring phonemes is not the same. Dominance functions corresponding to various articulators may differ in offset, duration, and magnitude. In [49], Cohen and Massaro implemented Löfqvist's gestural theory of speech production, using negative exponential functions as a general form for dominance functions. In their system, the movement of articulators that correspond to a particular phoneme is obtained by spatially and temporally blending (using dominance functions) the effect of all neighboring phonemes under consideration. In other rule-based coarticulation modeling approaches, Pelachaud et al. [56] clustered phonemes into visemes with different deformability ranks, while Breen et al. [57] directly used context in the units used for synthesis, by utilizing static context-dependent visemes. Overall, rule-based methods allow for incremental improvements by refining the articulation models of particular phonemes, which can be advantageous in certain scenarios.

In contrast to rule-based techniques, data-based coarticulation models are derived after training (optimizing) a number of model parameters on an available audiovisual database. Various such models have been considered for this purpose in the literature, for example ANNs and HMMs [51, 52]. Data-based coarticulation models can also be obtained using a concatenative approach [53, 54], where a database of video segments corresponding to context-dependent visemes is created using the phoneme-level transcription of a training audiovisual database. The main advantage of data-driven methods is that they can capture subtle details and patterns in the data, which are generally difficult to model by rules. In addition, retraining for a different speaker or language can be automated. Several approaches for generating visual speech parameters from the acoustic speech representation using

data-driven coarticulation models have been investigated in the literature [51–54].

5.2 Facial Animation

The face is a complex structure consisting of bones, muscles, blood vessels, skin, cartilage, and so forth. Developing a facial animation system is therefore an involved task, requiring a framework for describing the geometric surfaces of the face, its skin color, texture, and animation capabilities. Several computer facial animation systems have been reported in the literature, which can be classified as model-based (also known as knowledge-based) or image-based.

In the model-based facial animation approach, a face is modeled as a three-dimensional (3D) object, and its structure is controlled by a set of parameters. The approach has become popular due to the MPEG-4 facial animation standard [88], and it consists of the following three steps: designing the 3D facial model; digitizing a 3D mesh; and animating the 3D mesh to simulate facial movements. In the first step, a 3D model that captures the facial geometry is created. Most models describe the facial surface using a polygonal mesh (see also Fig. 11A). This method is frequently used due to its simplicity and availability of graphics hardware for efficient rendering of polygon surfaces. The facial surface should not be oversampled, since that would lead to computationally expensive facial animation. The polygons must also be laid out in a way that permits the face to flex and change shape naturally. In the second step, a digitized 3D facial mesh is constructed. This is typically achieved by obtaining the subject's facial geometry using 3D photogrammetry or a 3D scanner. Finally, in the third step, the 3D mesh is animated to simulate facial movements. During animation, the face surface is deformed by moving the vertices of the polygonal mesh, keeping the network topology unchanged.

The motion of the vertices is driven by a set of control parameters. These are mapped to vertex displacements based on interpolation, direct parameterization, pseudo-muscular deformation, or physiologic simulation. In the interpolation approach, a number of key frames, usually corresponding to visemes and facial expressions, are defined, and their vertex positions are stored. The frames in between key frames are generated by interpolation, since all possible linear combinations of key frames are represented by the control parameter space. The main advantages of this approach are simplicity and its support by commercial animation packages. However, the disadvantages lie in the fact that facial feature motion is typically nonlinear, and that the number of achievable facial

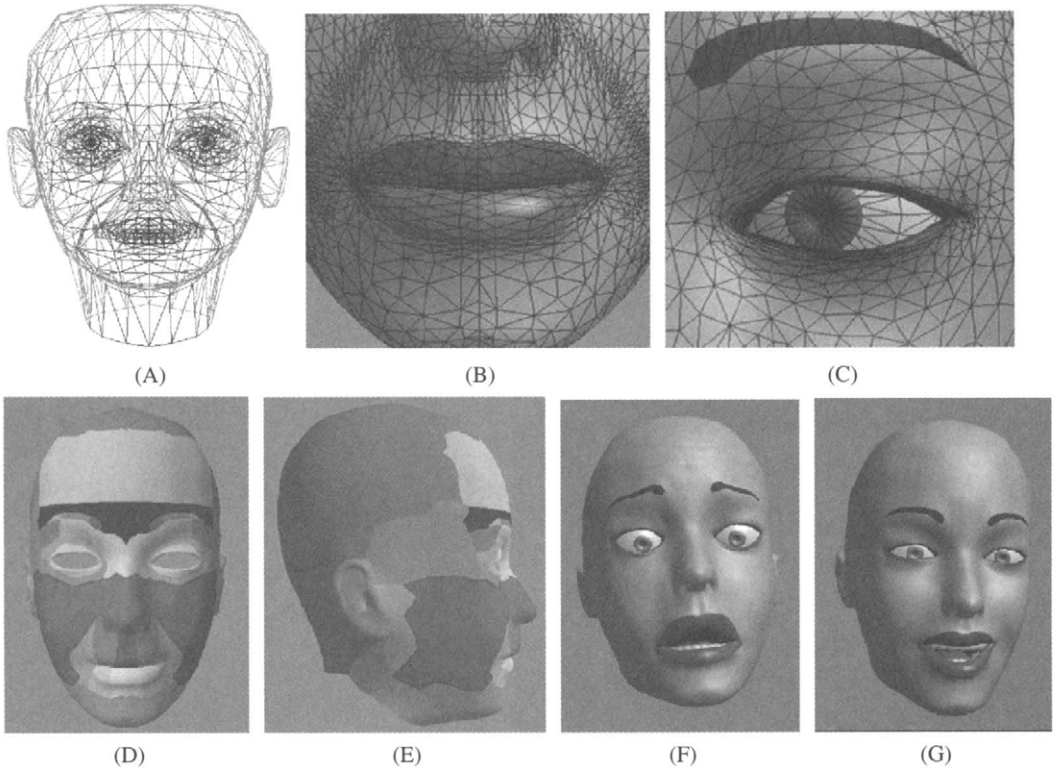


FIGURE 11 MPEG-4 compliant facial animation. (**A**): A polygonal mesh [93]; (**B,C**) detailed structure of the most expressive face regions; (**D,E**) three-dimensional surface is divided into areas corresponding to feature points affected by facial animation parameters and (**F,G**) synthesized expressions of fear and joy. (**B–G**) Correspond to model *Greta* (reproduced with permission from [59]). (See color insert.)

expressions is limited by the number of key frames used. In the direct parameterization approach, basic geometric transformations, such as translation, rotation, and scaling, are used to describe vertex displacements. Pseudo-muscular models, on the other hand, use facial muscle structure to model deformations. The space of allowable deformations is reduced by knowledge of the human face anatomic limitations. Muscles are modeled with one end affixed to the bone structure of the skull and the other end attached to the skin. Finally, modeling the skin with three spring-mass layers has also been used to develop more detailed physiologic models. The main advantage of this approach is the improved realism over purely geometric facial modeling techniques.

Most model-based facial animation systems used today are extensions to Parke's work [58]. His model utilizes a parametrically controlled polygon topology, where the face is constructed from a network of approximately 900 surfaces, arranged and sized to match the facial contours. Large polygons are used in flattered regions of the face, while small ones are used in high curvature areas. Face animation is controlled by a set of about 50 parameters, ten of which drive the articulatory movements involved in speech production. In related work [59], Pasquariello and Pelachaud developed a 3D facial model, named "Greta," consisting of 15,000 polygons (see Figs. 11B and 11C). Greta is compliant with the MPEG-4 standard [88], and able to generate, animate, and render in real-time the structure of a proprietary 3D model. The model uses the pseudo-muscular approach to describe face behavior, and it includes features such as wrinkles, bulges, and furrows to enhance its realism. In particular, a great level of detail is devoted to the facial regions that contain most speechreading and expression information, such as the mouth, eyes, forehead, and the nasolabial furrow (see Figs. 11B and 11C). Furthermore, to achieve more control on the polygonal lattice, the 3D model surface is divided into areas that correspond to feature points affected by FAPs (see also Figs. 11D, and 11E). Examples of facial animation employing the Greta model to display fear and joy expressions are shown in Figs. 11F and 11G.

In contrast to the model-based techniques discussed above, the image-based facial animation approach relies mostly on image processing algorithms [53–55]. There, most of the work is performed during a training process, through which a database of video segments is created. Thus, unlike model-based approaches that use a static facial image, image-based techniques use multiple facial images, being able to capture subtle face deformations that occur during speech. Image-based facial animation consists of the following steps: recording of the video of the subject; video segmentation into animation groups; and animation of the model by concatenating various animation groups. In the first step, video of the subject uttering nonsense syllables or sentences in a controlled environment is recorded. In the second step, the recorded video is analyzed, and video segments consisting of phone, triphone, or word boundaries are identified. Finally, in the third step, the video segments are concatenated to realize the animation. Interpolation and morphing are usually employed to smooth transitions between boundary frames of the video segments.

Several examples of image-based facial animation can be found in the literature. For example, in [55], Ezzat and Poggio report a visual TTS system, named "MikeTalk," which uses 52 viseme images, representing 24 consonants, 12 monophthongs, and 16 diphthongs. To generate smooth transitions between the viseme images, morphing is used. However, the system processes the mouth area only, and does not synthesize head movements or facial expressions. In [53], Bregler et al. report a speech-to-video synthesis system that also uses image-based facial animation. Their system utilizes existing footage to create video of a subject uttering words that were not spoken in the original footage. In the analysis stage, time-alignment of the speech is performed (using HMMs trained on the TIMIT database) to obtain phonetic labels, which are consequently used to segment the video into tri-phones. Only the mouth area is processed and then reimposed with new articulation into the original video sequence. Tri-phone videos and the phoneme labels are stored in the video model. In the synthesis stage, morphing and stitching are used to perform time-alignment of tri-phone videos, time-alignment of the lips to the utterance, illumination matching, and combination of lips and the background. The main disadvantages of this approach lie in the size of the tri-phone video database and in the fact that only the mouth area is processed. Other facial parts such as eyes and eyebrows that carry important conversational information are not considered. To overcome these shortcomings, Cosatto and Graf [54] decompose their facial model into separate parts. The decomposed head model contains a "base face," which covers the area of the whole face and serves as a substrate onto which the facial parts are integrated. The facial parts are the mouth with cheeks and jaw, the eyes, and the forehead with eyebrows. Each part is modeled separately, therefore the number of stored image samples in the database is kept at a manageable level. This allows for independent animation of various areas of the face, therefore increasing the number of free parameters in the animation system and the amount of conversational information contained in the facial animation.

5.3 Visual Text-to-Speech

A general block diagram of a text-driven facial animation system is shown in Fig. 12 [94]. The input text is first processed by a natural language processor (NLP), which analyzes it at various linguistic levels to produce phonetic and prosodic information. The latter refers to speech properties, such as stress and accent (at the syllable or word level), and intonation and rhythm, which describe changes in pitch and timing across words and utterances. NLP can also produce visual prosody, which conveys information about facial expressions

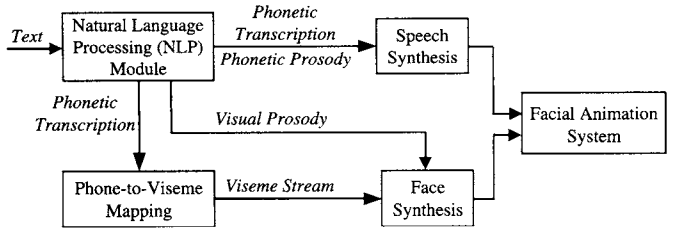


FIGURE 12 Components of typical visual text-to-speech synthesis systems.

(e.g., anger, disgust, happiness, sadness, etc.). The generated phonetic and prosodic information can then be used by the speech synthesis module to produce an acoustic speech signal. Similarly, the face synthesizer uses viseme information, obtained through phoneme-to-viseme mapping, and visual prosody to produce a visual speech signal.

A number of researchers have used this approach for VTTS with their techniques for coarticulation modeling and facial animation. For example, in [60], Cohen and Massaro used Parke's facial model as the basis for their text-driven speech synthesizer. Their main improvements over Parke's model were the inclusion of tongue and their extensive study of coarticulation, which they integrated into face animation [49]. For the "MikeTalk" system [55], discussed earlier, Ezzat and Poggio manually extracted a set of viseme images from the recordings of a subject enunciating 40 to 50 words. They assumed a one-to-one mapping between phonemes and visemes, and modeled a viseme with a static lip shape image instead of using a sequence of images. Subsequently, they reduced the number of visemes to 16 and constructed a database of 256 optical flow vectors that specified the transitions between all possible viseme images. Finally, they used a TTS system to translate text into a phoneme stream with duration information, and used it to generate a sequence of viseme images for face animation, synchronized with TTS-produced speech.

In other work [54], Cossato and Graf developed a TTS system using their decomposed facial model, discussed in the previous section. They first recorded a video database consisting of common tri-phones and quadri-phones uttered by a subject. Then, they extracted and processed mouth images from the video, obtaining both geometric and PCA visual features, and subsequently parametrized and stored them in bins located on a multidimensional grid within the geometric feature space. To reduce storage requirements, within each bin, they discarded mouth images "close" to others in the PCA space, using a vector quantization scheme based on Euclidean distance. For synthesis, they used a coarticulation model, similar to the one in [49], to obtain a smooth trajectory in the geometric feature space, based on the target phonetic sequence and a mapping between visemes and their "average" geometric feature representation. To generate the mouth region animation, they sampled the resulting trajectory at the video rate, and, at each time instant, they chose the closest grid point, providing in this manner a set of candidate mouth bitmaps located within the corresponding bin. Next, they utilized the Viterbi algorithm to compute the lowest-cost path through a graph, having as nodes the candidate images at each time instant. For the transition cost between nodes (mouth images) at consecutive times, they used their Euclidean distance in the PCA space, setting this cost to zero in case the images correspond to neighboring frames of the original video. The resulting path provided the final mouth sequence animation.

5.4 Speech-to-Video Synthesis

Speech-to-video synthesis systems exploit the correlation between acoustic and visual speech, in order to synthesize a visual signal from the available acoustic signal (see Fig. 13). Several approaches for speech-to-video synthesis have been reported in the literature, using methods, such as VQ, ANNs, or HMMs (see also Section 3.2). In general, these techniques can be classified into regression- and symbol-based approaches and are briefly reviewed in this section.

Regression-based methods establish a direct continuous association between acoustic and visual features. VQ and ANNs are commonly used for this task, with the former constituting a simpler approach. In the training phase of VQ-based speech-to-video synthesis, an acoustic codebook is first constructed using clustering techniques. The codebook allows classifying audio features into a small number of classes, with the visual features associated with each class averaged to produce a centroid to be used in synthesis. At the synthesis stage, the acoustic parameters at a given instant are compared against all possible acoustic classes. The class located closest to the given parameters is selected, and the corresponding visual centroid is used to drive the facial animation. In ANN-based speech-to-video synthesis, the acoustic and visual speech features correspond to the input and output network nodes, respectively. In the training phase, the network weights are adjusted using the back-propagation algorithm. For synthesis, at each time instant, the speech features are presented to the network input, with the visual speech parameters generated at the output nodes of the ANN.

The work reported in [51] constitutes a typical example of the regression-based approach. There, Morishima and Harashima investigated the use of VQ and ANNs for predicting facial features from audio. They considered 16-dimensional LPC vectors and eight facial feature points (located on the lips, jaw, and ears) as the representations of the acoustic and visual signals, respectively. In their VQ scheme, they created a 5-bit codebook to allow mapping of the acoustic to the visual parameters, while in their ANN-based algorithm, they used a three-layer ANN architecture. In related work, Lavagetto [61] proposed using six independent time-delay neural networks with four layers, each accepting identical

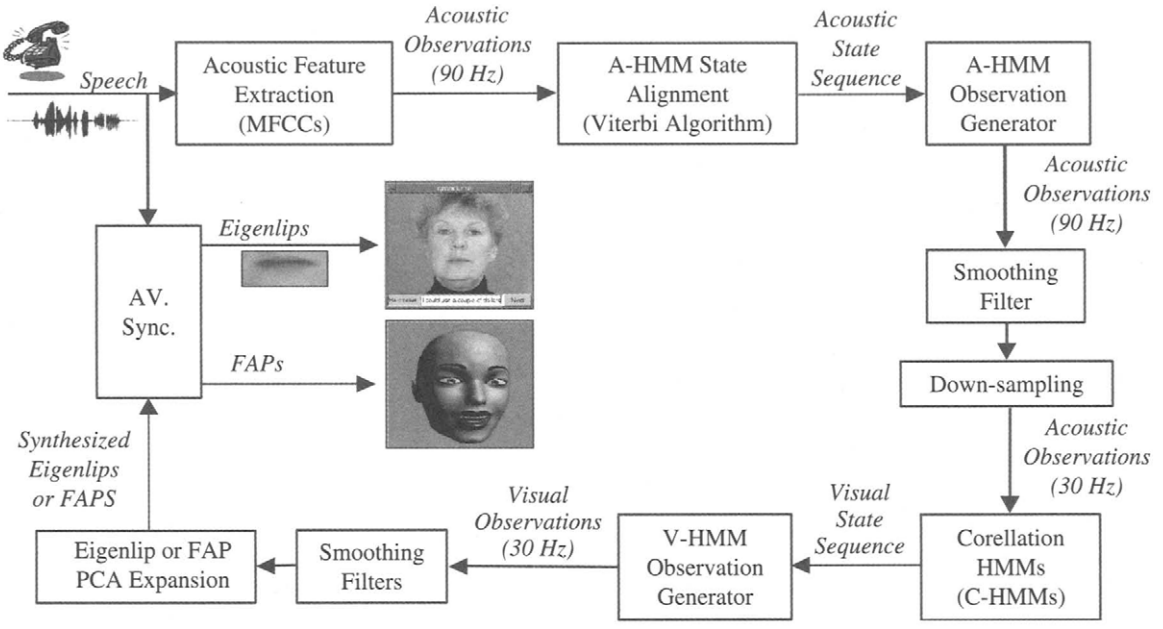


FIGURE 13 The speech-to-video synthesis systems developed in [50, 52] utilize narrowband speech to generate two possible visual representations: eigenlips that can be superimposed on frontal face videos for animation, or facial animation parameters that can be used to drive an MPEG-4–compliant facial animation model. (See color insert.)

acoustic feature input (12-dimensional LPC vectors), but generating as output individual parameters of a geometric visual speech representation.

In contrast to regression-based techniques, in the symbol-based approach, the acoustic signal is first transformed into an intermediate discrete representation consisting of a sequence of subphonetic or subword units. HMMs are typically used for this purpose, since they provide explicit phonetic information, which can help in the analysis of coarticulation effects. Reported HMM-based systems vary in two basic aspects: the units used for recognition (i.e., what do the HMM states represent; see also Sections 3.1 and 3.2) and the method for synthesizing the visual parameter trajectories from the recognized HMM state sequence. Examples of such systems are provided next.

Simons and Cox [62] developed an HMM-based speech-driven synthetic head. They analyzed a small number of phonetically rich sentences to obtain several acoustic and visual training vectors. They used VQ to produce audio and visual codebooks of sizes 64 and 16, respectively. Then, they created a fully-connected 16-state discrete HMM, each state representing a particular vector quantized mouthshape, and producing the 64 possible audio code words. The HMM transition and observation probabilities were trained on the basis of the joint audiovisual vector-quantized data. Subsequently, the trained HMM was used in synthesis by means of the Viterbi algorithm, generating the most likely visual state sequence (hence, visual representation) given the input audio observations.

Chen and Rao [5] trained continuous whole-word HMMs using audiovisual observations (henceforth such HMMs are referred to as AV-HMMs). They used the width and height of the outer lip contour as visual features, and 13 MFCCs as acoustic features. Subsequently, they built for each word an acoustic HMM (A-HMM), which had the same transition matrix and initial state distribution as the corresponding AV-HMM [see Equations (1), (2), and (4)]. The state acoustic observation pdf for each particular A-HMM state was derived by integrating the AV-HMM observation pdf over the visual parameters. In the synthesis phase, the A-HMMs and the acoustic observations were first used, employing the Viterbi algorithm, to obtain the optimal acoustic state sequence. Next, assuming that the AV-HMM state sequence is the same as the A-HMM state sequence, they estimated for each state the cor-responding visual feature vector, using AV pdfs and the acoustic observations.

Bregler et al. [53] created an HMM-based speech-driven facial animation system called "Video Rewrite." They first trained an A-HMM system on the TIMIT database, and used it to segment the audio portion of a joint audiovisual database into tri-phones. The visual segments, time-synchronous to the resulting tri-phones, were then stored into a video database, indexed by the corresponding tri-visemes. At the synthesis stage, given the input acoustic signal, they first obtained its phonetic level transcription using the A-HMM system. Subsequently, they used the concatenative approach to synthesis with visual segments selected from the created video database. Various cost metrics were considered for the segment

selection, and the Viterbi algorithm was used to obtain the optimal sequence of video segments. Finally, they used warping techniques to smooth the selected video segments and synchronize them with the speech signal.

Finally, two systems were developed at NWU [50, 52], using two different visual speech representations, eigenlips and FAPs. PCA was performed on both FAPs and mouth images to obtain visual features of lower dimensionality. MFCCs were used as acoustic features in both systems. The block diagram of the developed systems is shown in Fig. 13. The two systems utilized continuous A-HMMs, visual HMMs (V-HMMs), and correlation HMMs (C-HMMs). In this approach, the A-HMMs and the Viterbi algorithm were used to realize the audio state sequence that best described the acoustic observations extracted from the input narrowband speech signal. The A-HMM observation generator then used the means corresponding to each resulting A-HMM state to produce speaker-independent observations [see also (1), (2), and (6) in Section 3.2]. Smoothing and down-sampling were subsequently used to obtain acoustic observations at the video rate (30 Hz), while the C-HMM system mapped the generated acoustic observations, using the Viterbi algorithm, into a visual state sequence. Finally, the visual state sequence and the V-HMM observation generator were employed to produce visual observations.

Two key elements of the NWU systems were the C-HMM training procedure and model architecture. To ensure that the C-HMMs were capable of approximating the optimal visual state sequence given the acoustic observations, they were built with the same topology and identical state transition and initial probabilities as the V-HMMs. As a result of the above constraints, only the C-HMM observation pdfs had to be estimated during training (see also Section 3.2). In more detail, the C-HMMs were trained using the following procedure [50, 52]: In the first step, A-HMMs and V-HMMs were independently trained using the TIMIT corpus and the visual part of the Bernstein database, respectively. The two HMMs had different topologies to account for the unequal audio and video observation rates. Next, in the second step, the trained A-HMMs and V-HMMs in conjunction with the Viterbi algorithm were used to force-align acoustic and visual training data, respectively, and generate corresponding acoustic and visual state sequences. In the third step, down-sampled acoustic observation sequences were generated using the acoustic state sequences obtained in the second step. In the fourth step, the visual state sequence generated in the second step was utilized as a constraint to distribute the down-sampled acoustic observations among the C-HMM states. Finally, reestimation of the C-HMM observation pdfs was carried out. This training procedure generated C-HMMs capable of producing, in conjunction with V-HMMs, visual state sequences and estimates of the visual articulatory movements from down-sampled acoustic observations.

5.5 Visual Speech Synthesis Evaluation

Evaluating visual synthesis systems is extremely important to benchmark algorithmic improvements, assess the suitability of specific databases for training data-driven techniques, and quantify the benefit of incorporating the visual modality over traditional audio-only synthesis, for example. There are both objective and subjective methods to evaluate visual speech synthesis. The former typically compare the difference between a set of synthesized and recorded test sequences, in terms of mean squared error or other distance metrics in the visual speech representation space, or, alternatively, report ASR performance on the synthesized test set [52]. Although relatively easy to perform, objective evaluation does not necessarily indicate how two systems will be relatively received by human users in practice. Subjective testing is instead required for such assessments [7, 48].

In general, subjective evaluation of visual speech synthesis performance should be application-dependent. Such tests should be developed with the goal of evaluating a number of issues, for example the degree of realism in the animation, user satisfaction, and the effectiveness in communicating the intended message. In particular, effectiveness in communicating and especially intelligibility of a talking head should be of primary importance. Intelligibility evaluation approaches aim at measuring either phoneme identification performance (recognition of vowels and consonants) or speechreading performance (recognition of isolated words or sentences), by human subjects.

In this section, we provide an example of an intelligibility subjective evaluation in the case of the eigenlips-based, speech-to-video synthesis system developed at NWU and discussed earlier. In these tests [48, 50], several subjects have been presented with three types of stimuli: (a) audio-only signal; (b) audio, supplemented by the synthesized video signal; and (c) audio, together with the original video footage. In all cases, the audio (and video, where applicable) utterances were from the Bernstein lipreading corpus [91], and the intelligibility experiments were conducted with the audio corrupted by additive, white Gaussian noise, resulting in speech signals of -5 dB, -10 dB, and -15 dB SNRs (see Fig. 14). The audio-only word recognition accuracy achieved by the subjects was 92.2%, 66.8%, and 11.7%, at the three SNR levels, respectively. The word-recognition accuracies improved significantly, when synthesized video was also presented to the subjects, reaching 97.9%, 87.5%, and 46.1%. Subjective tests under scenarios (a) and (b) were also performed using sets that contained certain number of repeated utterances used throughout all the tests. The word-recognition accuracies improved when repeated utterances were used, especially for the SNR of -15 dB, indicating that the subjects used prior knowledge to assist the transcription of the repeated utterances. However, these results were still inferior to human speech perception word-recognition accuracies obtained using natural instead of

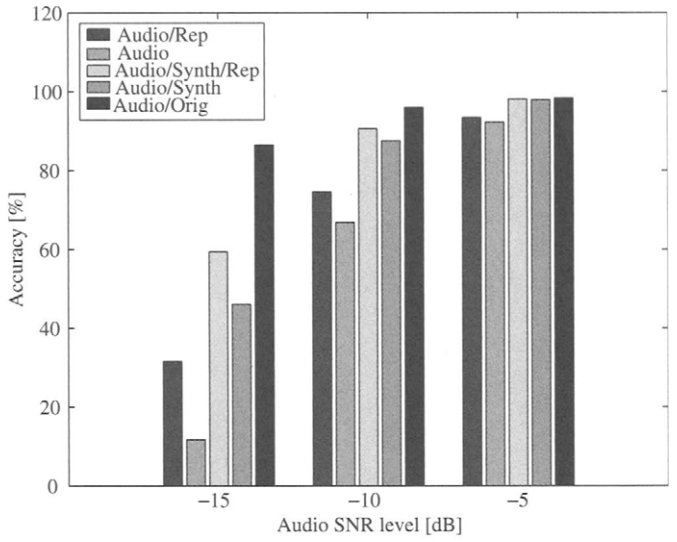


FIGURE 14 Intelligibility-based subjective evaluation of the speech-to-video synthesis system developed at Northwestern University [48, 50]. Human speech perception is compared using audio-only vs. audio with synthesized video and vs. audio with natural video of the lip region. For the first two conditions, results for repetitive presentation of the stimuli to the subjects are also given ("Rep"). Experiments are reported over three acoustic noise conditions. (See color insert.)

synthesized video, namely 98.3%, 95.9%, and 86.5%, respectively. Clearly, these subjective tests suggest that animated faces obtained using visual speech synthesizers can improve speech intelligibility, especially under noisy conditions (see also [7]).

6 Audiovisual Speaker Recognition

Audiovisual speaker recognition (also known as audiovisual biometric) systems utilize acoustic and visual information to perform automatic person recognition. A person recognition system should be capable of rejecting claims from impostors, persons not registered with the system, and accepting claims from the clients, persons registered with the system. Person recognition can be classified into two problems: person identification and person verification (authentication) [68]. Person identification is the problem of determining the identity of a person (who the person is) from a closed set of candidates, while person verification refers to the problem of determining whether a person is who s/he claims to be. There are a number of systems that require person recognition to reliably determine the identity of persons requesting their services. Applications that can use person recognition systems include automatic banking, computer network security, information retrieval, secure building access, and so forth. Personal property, such as cell phones, PDAs, laptops, cars, and so forth, could also have built-in person recognition systems which would prevent impostors from using them.

Biometrics, or biometric recognition, refers to utilizing physiologic and behavioral characteristics for automatic person recognition. Traditional person identification methods, including knowledge-based [e.g., passwords, personal identification number (PINs)] and token-based (e.g., ATM or credit cards, and keys) do not provide reliable performance. Passwords can be compromised, while keys and cards can be stolen or duplicated. Identity theft is one of the fastest growing crimes in the United States. Unlike knowledge- and token-based information, biometric characteristics cannot be forgotten or easily stolen. There are many different biometric characteristics that can be used in person recognition systems, including fingerprints, palm prints, hand and finger geometry, hand veins, iris and retinal scans, infrared thermograms, DNA, ears, faces, gait, voice, signature, and so forth (see Fig. 15) [71, 72, 76, 95].

Each biometric characteristic has its own advantages and disadvantages and there is no single modality which performs the best for all applications. The choice of biometric characteristics depends on many factors including the best achievable performance, uniqueness, robustness to noise, cost of biometric sensors, invariance of characteristics with time, robustness to attacks, population coverage, scalability, and so forth. All of these factors are usually considered when choosing the most appropriate biometric characteristics for a certain application. In addition, there are a number of biometric applications for which it is desirable to use nonintrusive and user-friendly methods for extraction of biometric features. Developing such biometric systems makes biometric technology more socially acceptable and accelerates its integration into every day life.

A person's voice and face are biometric characteristics that are easily collected and natural to the user. These characteristics can be utilized for nonintrusive person recognition. The recent technology advances decreased the cost of audio and video biometric sensors and opened a door to audiovisual biometrics. Acoustic and visual biometric characteristics can contain static and dynamic information. LPCs, MFCCs, and their derivatives, are commonly used as acoustic features in speaker recognition systems. Visual features can describe only the mouth region (visual-labial features) or the whole face (visual-facial features). Both mouth and face can be represented using shape-based features or appearance-based features (see also Section 2). Shape-based labial features include lip-contour shape and geometric features, while shape-based facial features include active shape models, facial feature geometry, elastic graphs, etc. Labial and facial appearance-based features are obtained using image projections such as LDA, PCA, DCT, and so forth, on mouth or face images. Facial features can also be classified as global or local if the face is represented by only one feature vector or by multiple vectors each representing local information. Face images used for extraction of visual features can be visible or infrared, two-dimensional (2D) or 3D, and so forth.

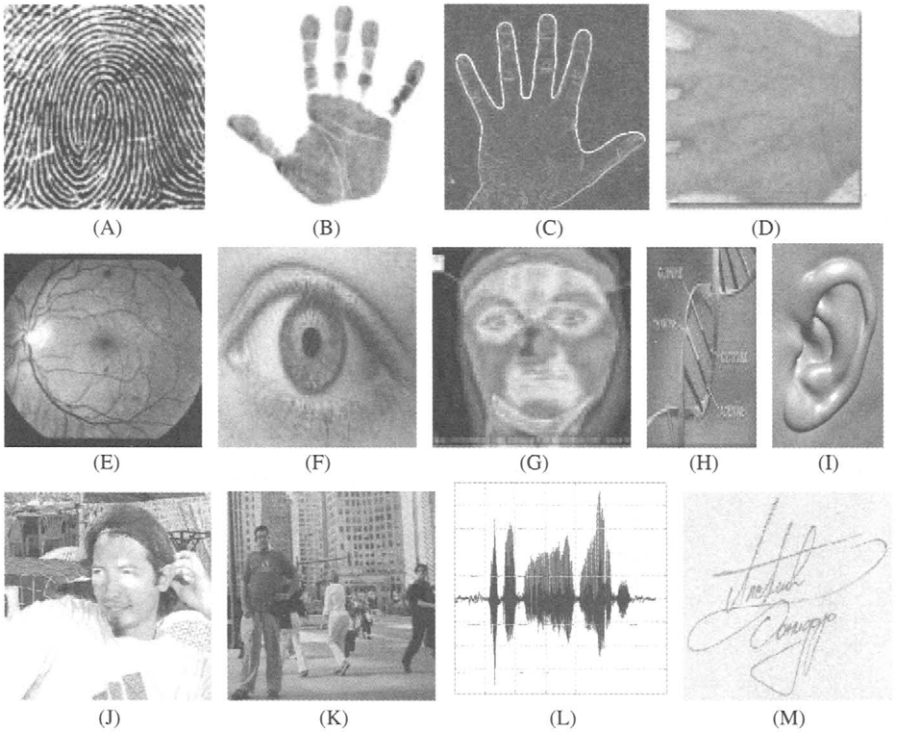


FIGURE 15 Biometric characteristics: (**A**) fingerprints; (**B**) palm print; (**C**) hand and finger geometry; (**D**) hand veins; (**E**) retinal scan; (**F**) iris; (**G**) infrared thermogram; (**H**) DNA; (**I**) ears; (**J**) face; (**K**) gait; (**L**) speech; (M) signature. (See color insert.)

Although single-modality biometric systems can achieve high performance in some cases, they are usually not robust to noise and do not meet the needs of many potential person recognition applications. Speaker recognition systems that rely only on audio data are sensitive to microphones (headset, desktop, telephone, etc.), acoustic environment (car, plane, factory, etc.), and channel noise (telephone lines, VoIP, etc.). On the other hand, systems that rely only on visual data can be sensitive to visual noise (lightning changes, poor video quality, occlusion, segmentation errors, etc.). To improve the robustness of biometric systems, multisamples (multiple samples of the same biometric characteristic), multialgorithms (multiple algorithms with the same biometric sample), and multimodal (different biometric characteristics) biometric systems have been developed. The advantage of multimodal biometric systems lies in their robustness, since different modalities can provide independent (complementary) information. Different modalities are combined to eliminate problems characteristic of single modalities. It has been shown that using multiple biometric modalities improves the performance of a biometric system [71–73, 76]. In audiovisual speaker recognition systems, speech is utilized together with either static video frames of faces (face recognition) or video sequences of the face (or the mouth area) in order to improve speaker recognition performance (see Fig. 16). Audio-visual speaker recognition systems can also utilize all three modalities [74].

Audiovisual speaker recognition systems can be either text-dependent, where speech used for training and testing is constrained to be the same, or text-independent, where speech used for testing is unconstrained. The methods for modeling speakers based on their audiovisual biometric data are usually statistical in nature. Such approaches include, ANNs, SVMs, GMMs, HMMs, etc. (see also Section 3.2). HMMs represent the most commonly used approach for speaker recognition.

In speaker identification systems, the objective is to determine the class $\hat{c}$, corresponding to the enrolled person or the impostor, that best matches the unknown person's audiovisual biometric data $\mathbf{o}_{av,t}$, that is

$$\hat{c} = \arg\max_{c \in C} Pr(c \mid \mathbf{o}_{av,t}),$$

where C denotes the set of classes corresponding to all speakers in the database and the impostor, and $Pr(c \mid \mathbf{o}_{av,t})$ the conditional probability that biometric observations $\mathbf{o}_{av,t}$ were generated by the statistical model for the class c.

In speaker verification systems there are only two classes, and it is necessary to determine whether the class corresponding to the general population (w) or the class corresponding to the true claimant (c) best matches the claimant's biometric observations. The similarity measure (D) can be defined as

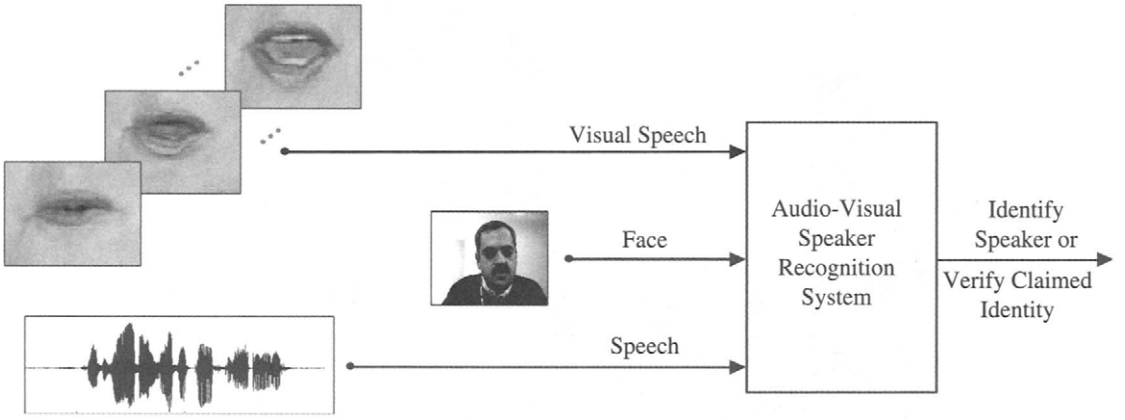


FIGURE 16 Block diagram of an audiovisual speaker recognition system that utilizes static (face image) and dynamic (visual speech) visual information together with acoustic information. (See color insert.)

the likelihood ratio between the speaker set and the world set, that is

$$\mathrm{D} = \log Pr(c \mid \mathbf{o}_{av,t}) - \log Pr(w \mid \mathbf{o}_{av,t}).$$

If D is larger than an *a priori* defined verification threshold the claim is accepted; otherwise it is rejected.

In text-dependent speaker recognition systems, subphonetic units are usually modeled. In that case, the set of classes can be considered as the product space between the set of speakers and the set of phonetic based units. HMMs are commonly used for text-dependent speaker recognition, through modeling the phonetic units by Gaussian mixture densities [see (1) in Section 3.2]. In text-independent systems single-state HMMs (GMMs) can be used to model speakers. In this case a single GMM is assumed to generate the entire audiovisual observation sequence.

Two commonly used error measures for verification performance are false acceptance (FA)—an impostor is accepted—and false rejection (FR)—a client is rejected. They are defined by

$$\mathrm{FA} = I_A/I \times 100\% \quad \mathrm{FR} = C_R/C \times 100\%,$$

where I_A denotes the number of accepted impostors, I the number of impostor claims, C_R the number of rejected clients, and C the number of client claims. There is a trade-off between FA and FR, which is controlled by the choice of the verification threshold. It is usually chosen according to certain FA and FR requirements, based on results obtained through experiments on the evaluation set. The choice of the verification threshold clearly depends on the application and costs assigned to each of the error measures. For example, systems that control access to a highly secure area or manage banking transactions would require very low FA at the expense of increased FR. On the other hand, systems that control tolls or gym access would avoid putting their legitimate customers in inconvenient situations by requiring low FR, at the expense of increased FA. Verification system performance can also be measured using an equal error rate (EER) measure. It is determined after the verification experiments are performed, by choosing the verification threshold for which FA and FR are equal.

Performance of audiovisual speaker recognition systems strongly depends on the choice and accurate extraction of the visual features, and the information fusion approach. Acoustic and visual observations can either be combined to form joint audiovisual observations or utilized as separate observation streams. Information fusion approaches commonly used for fusion of audio and visual biometric information are discussed in more detail in Section 3.3 and in [71, 72]. A number of audiovisual speaker recognition systems that utilize various types of visual features and audiovisual information fusion strategies, have been reported in the literature [72–76].

Brunelli and Falavigna [76] developed a text-independent speaker identification system that combines audio-only speaker identification and face recognition systems. The two systems provide five classifiers, two acoustic and three visual. Two acoustic classifiers correspond to two sets of acoustic features (static and dynamic) derived from the short time spectral analysis of the speech signal. Their audio-only speaker identification system is based on VQ. Three visual classifiers correspond to the visual classifying features extracted from three regions of the face: eyes, nose, and mouth. The individually obtained classification scores are combined using a weighted geometric average. The identification rate of the integrated system is 98%, compared with the 88% and 91% rates obtained by the audio-only speaker recognition and face recognition systems, respectively.

Aleksic and Katsaggelos [73] developed an audiovisual speaker recognition system that utilized 13 MFCC coefficients and their first- and second-order derivatives as acoustic features. A visual feature vector consisting of ten FAPs that describe the movement of the outer-lip contour [88] was projected by means of the PCA onto a 3D space. The resulting visual features were augmented with first- and second-order

TABLE 3 Speaker identification and verification errors obtained when audio-only (AU) or audiovisual (AV) biometric data was utilized

	Identification Error, %		Verification Error (EER), %	
SNR	AU	AV	AU	AV
Clean	5.13	5.13	2.56	1.71
20	19.51	7.69	3.99	2.28
10	38.03	10.26	4.99	2.71
0	53.10	12.82	8.26	3.13

SNR, signal-to-noise.

derivatives providing nine-dimensional dynamic visual feature vectors. They used a feature fusion integration approach and single-stream HMMs to integrate acoustic and visual information. Speaker verification and identification experiments were performed using audio-only and audiovisual information, under both clean and noisy audio conditions at SNRs ranging from 0 to 20 dB. Speaker identification and verification results obtained, expressed in terms of the identification error and EER, are shown in Table 3. Significant improvement in performance over audio-only speaker recognition system was achieved, especially under noisy acoustic conditions. For instance, the identification error was reduced from 53.1%, when audio-only information was utilized, to 12.82%, when audiovisual information was employed at 0 dB SNR.

Jourlin et al. [75] developed an audiovisual speaker verification system that utilizes both acoustic and visual dynamic information. Their 39-dimensional acoustic features consist of LPC coefficients and their first- and second-order derivatives. They use 14 lip shape parameters, ten intensity parameters, and the scale as visual features, resulting in a 25-dimensional visual feature vector. They utilize HMMs and the decision fusion integration approach to perform audio-only, visual-only, and audiovisual experiments. The audiovisual score is computed as a weighted sum of the audio and visual scores. Their results demonstrate a reduction of FA from 2.3% when the audio-only system is used to 0.5% when the multimodal system is used.

Chaudhari et al. [72] developed an audiovisual speaker identification and verification system that modeled reliability of the audio and video information streams with parameters which were time-varying and context dependent. The acoustic features consisted of 23 MFCC coefficients, while visual features consisted of 24 DCT coefficients obtained by applying DCT on the ROI extracted by means of a face tracking algorithm. They utilized GMMs to model speakers and parameters that depended on time, modality, and speaker to model stream reliability. The system that utilized time dependent stream weights achieved an EER of 1.04%, compared with, 1.71%, 1.51%, and 1.22%, of the audio-only, video-only, and audiovisual (feature fusion) systems, respectively.

Dieckmann et al. [74] developed a system that used visual features obtained from all three modalities, face, voice, and lip movement. The identification error decreased to 7% when all three modalities were used, compared with 10.4%, 11%, and 18.7%, when voice, lip movements, and face visual features were used individually.

In summary, there is a need for resources for advancing and accessing audiovisual speaker recognition systems. Publicly available multimodal corpora that better reflects realistic conditions, such as acoustic noise and lighting changes would help in investigating robustness of audiovisual systems. In addition, standard experiments and evaluation procedures should be defined in order to enable fair comparison of different systems. Baseline algorithms and systems could also be chosen and made available to facilitate separate investigation of effects that factors, such as, the choice of acoustic and visual features, the information fusion approach, and classification algorithms, have on system performance.

7 Summary and Discussion

In this chapter, we have focused on how the joint processing of visual and audio signals, both generated by a talking person, can provide valuable speech information to benefit a number of audiovisual speech processing applications crucial to human–computer interaction. We first concentrated on the analysis of visual signals, and described various possible ways of representing and extracting the speech information available in them. We then discussed how the obtained visual features can complement features extracted (by well-studied methods) from the acoustic signal and how the two modality representations can be fused together to allow joint audiovisual speech processing. The general bimodal integration framework was subsequently applied to three problems, namely automatic speech recognition, talking face synthesis, and speaker identification and authentication. In all three cases, we discussed issues specific to the particular application, reviewed several relevant systems that have been reported in the literature, and presented results using the implementations developed at IBM Research and/or Northwestern University. The experimental results demonstrated the importance of utilizing visual speech information, especially in the presence of acoustic noise.

As we mentioned in the Introduction, there are a number of additional applications that can benefit from the joint processing of audio and visual signals. Examples of such are emotion recognition, speaker detection and localization, speech activity detection, and enhancement of the acoustic signal or of its corresponding audio features. Due to lack of space, we only briefly address some of them in the following.

Automatic emotion recognition has many potential applications in human-computer interaction, for example by indirectly providing valuable user input to dialogue

management. Both facial expression and voice reflect the emotional state of a person, thus bimodal processing is a sensible approach. In the visual domain, there are six basic facial expressions: happiness, anger, sadness, fear, surprise, and disgust. Their study is enabled by the facial action coding system (FACS), which provides standardized coding of changes in facial motion through 46 action units that describe basic facial movements [96]. The FACS is based on muscle activity and captures in detail the effect of each action unit on the visual face features. Commonly used visual features for automatic emotion recognition include lip and eyebrow movements, whole face images, optical flow, and so forth [77, 78]. For the classification process, both spatial and spatiotemporal approaches can be used. In the former, visual features obtained from single face images are used, while spatiotemporal approaches utilize features extracted from each frame of the video sequence of interest. Typically, in facial expression recognition systems, artificial neural networks are used to perform spatial classification, whereas hidden Markov models are frequently used in the spatiotemporal approach [77]. Visual systems can of course be combined with audio-only emotion recognizers, using the audiovisual integration framework of Section 3. In this case, typically used audio features include the acoustic signal energy, pitch contour statistics, and so forth.

Among additional joint audiovisual processing applications, speaker detection and tracking are especially useful in environments such as conference rooms, where multiple persons are present, and signals from both video cameras and microphone arrays are available. In such occasions, speaker detection and tracking can be performed using acoustically guided cameras, visually guided microphone arrays, or through joint audiovisual tracking [63–65]. Of particular importance to speech applications is the detection of synchronous audio-visual sources in the presence of multiple speakers in the scene, as is often the case in broadcast videos. Joint audiovisual speech activity localization can benefit from the fact that the two modalities are correlated, and, for example, can be quantified by using mutual information of the two signals [66]. Furthermore, visual information, such as user pose and proximity to a computer or kiosk, as well as mouth movement, can be used to flag speech intent [67] or augment acoustic cues for speech activity detection. The resulting systems will be robust to environmental noise and are expected to eventually make the "push-to-talk" button in present automatic speech recognizers obsolete. Another application that exploits the correlation between the audio and visual speech signals is the bimodal enhancement of audio. There, acoustic information is restored using the video of the speaker's mouth region with the corrupted audio signal. The enhancement can occur in the signal space or the audio feature space utilizing linear or nonlinear techniques [18, 19]. Such an approach is beneficial, for example, when the amount of visual data available for training is insufficient to obtain visual-only speech models, thus not allowing audiovisual automatic speech recognition by means of the fusion techniques discussed in Section 3.

Clearly, the field of joint audiovisual signal processing is a very new, active, and exciting topic of research and development. Indeed, there are a number of major accomplishments, some of which have been described in this chapter in the context of speech applications for human–computer interaction. Concerning the practical deployment of these technologies, several obstacles have been slowly lifting, with audiovisual speech processing systems starting to exhibit real-time performance and improved robustness [46]. Nevertheless, various research issues remain open to further investigation. For example, the design of a truly speaker-independent, high-performing visual feature representation with improved robustness to the visual environment and user behavior, possibly using 3D face information, as well as the development of improved audiovisual integration algorithms that will allow unconstrained audiovisual asynchrony modeling and robust, localized reliability estimation of the signal information content, to name a few. Clearly, further research is required to advance the field and for audiovisual signal processing to become widespread in practice. The ground is fertile for additional major accomplishments and revolutionary future multimodal technologies and applications, promising to improve human–computer interaction and, with that, life quality.

8 References

[1] D. O'Shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis," *Proc. IEEE*, 91, 1272–1305 (2003).

[2] R. Stern, A. Acero, F.-H. Liu, and Y. Ohshima, "Signal processing for robust speech recognition," in *Automatic Speech and Speaker Recognition. Advanced Topics*, C.-H. Lee, F. K. Soong, and Y. Ohshima, eds. (Norwell, MA: Kluwer Academic, Norwell, MA, 1997), 357–384.

[3] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Commun.* 22, 1–15 (1997).

[4] R. van Bezooijen and V. J. Heuven, "Assessment of synthesis systems," in *Handbook of Standards and Resources for Spoken Language Systems*, D. Gibbon, R. Moore, and R. Winski, eds. (Mouton de Gruyter, New York, 1997), 481–563.

[5] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, 86, 837–852 (1998).

[6] S. Oviatt, P. Cohen, L. Wu, et al., "Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and research directions," *Human–Computer Inter.* 15, 263–322 (2000).

[7] J. Schroeter, J. Ostermann, H. P. Graf, et al., "Multimodal speech synthesis," in *Proc. Int. Conf. Multimedia Expo*, New York, NY, July 30–Aug. 2 (2000), 571–574.

[8] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A review of speech-based bimodal recognition," *IEEE Trans. Multimedia*, 4, 23–37 (2002).

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

[10] R. Campbell, B. Dodd, and D. Burnham, eds., *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech* (Psychology Press Ltd., Hove, United Kingdom, 1998).

[11] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *J. Acoust. Soc. Am.* 26, 212–215 (1954).

[12] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, 264, 746–748 (1976).

[13] M. Marschark, D. LePoutre, and L. Bement, "Mouth movement and signed communication," in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech*, R. Campbell, B. Dodd, and D. Burnham, eds. (Psychology Press Ltd., Hove, United Kingdom, 1998) 245–266.

[14] A. Q. Summerfield, "Some preliminaries to a comprehensive account of audiovisual speech perception," in *Hearing by Eye: The Psychology of Lip-Reading*, R. Campbell and B. Dodd, eds. (Lawrence Erlbaum Associates, London, United Kingdom, 1987) 3–51.

[15] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Commun.* 26, 23–43 (1998).

[16] D. W. Massaro and D. G. Stork, "Speech recognition and sensory integration," *American Scientist*, 86, 236–244 (1998).

[17] J. P. Barker and F. Berthommier, "Estimation of speech acoustics from visual speech features: A comparison of linear and non-linear models," in *Proc. Conf. Audio-Visual Speech Processing* (Santa Cruz, CA, Aug. 7–9, 1999) 112–117.

[18] L. Girin, J.-L. Schwartz, and G. Feng, "Audio-visual enhancement of speech in noise," *J. Acoust. Soc. Am.* 109, 3007–3020 (2001).

[19] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (audiovisual codebook dependent cepstral normalization)," in *Proc. Int. Conf. Spoken Lang. Processing* (Denver, CO, Sept. 16–20, 2002) 1449–1452.

[20] E. Petajan, "Automatic lipreading to enhance speech recognition," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, IL, 1984.

[21] M. E. Hennecke, D. G. Stork, and K. V. Prasad, "Visionary speech: looking ahead to practical speechreading systems," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, eds. (Springer, Berlin, Germany, 1996) 331–349.

[22] C. Bregler and Y. Konig, "Eigenlips' for robust speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Adelaide, Australia, Apr. 19–22, 1994) 669–672.

[23] P. Duchnowski, U. Meier, and A. Waibel, "See me, hear me: Integrating automatic speech recognition and lip-reading," in *Proc. Int. Conf. Spoken Lang. Processing* (Yokohama, Japan, Sept. 18–22, 1994) 547–550.

[24] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. Int. Conf. Image Process.* (Chicago, IL, Oct. 4–7) 1, 173–177 (1998).

[25] S. Nakamura, H. Ito, and K. Shikano, "Stream weight optimization of speech and lip image sequence for audiovisual speech recognition," in *Proc. Int. Conf. Spoken Lang. Processing* (Beijing, China, Oct. 16–20, 2000) 20–23.

[26] A. V. Nefian, L. Liang, X. Pi, et al., "Dynamic Bayesian networks for audio-visual speech recognition," *EURASIP J. Appl. Signal Process.* 2002, 1274–1288 (2002).

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

[28] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Seattle, WA, May 12–15, 1998) 3733–3736.

[29] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Proc. Int. Conf. Multimedia Expo* (Tokyo, Japan, Aug. 22–25, 2001).

[30] G. Chiou and J.-N. Hwang, "Lipreading from color video," *IEEE Trans. Image Process.* 6, 1192–1195 (1997).

[31] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. Multimedia*, 2, 141–151 (2000).

[32] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Commun.* 26, 149–161 (1998).

[33] M. Heckmann, F. Berthommier, and K. Kroschel, "Noise adaptive stream weighting in audio-visual speech recognition," *EURASIP J. Appl. Signal Process.* 2002, 1260–1273 (2002).

[34] A. Adjoudani and C. Benôt, "On the integration of auditory and visual parameters in an HMM-based ASR," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, eds.(Springer, Berlin, Germany, 1996) 461–471.

[35] P. Teissier, J. Robert-Ribes, and J. L. Schwartz, "Comparing models for audiovisual fusion in a noisy-vowel recognition task," *IEEE Trans. Speech Audio Process.* 7, 629–642 (1999).

[36] S. Gurbuz, Z. Tufekci, E. Patterson, and J. N. Gowdy, "Application of affine-invariant Fourier descriptors to lipreading for audiovisual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Salt Lake City, UT, May 7–11, 2001) 177–180.

[37] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Rationale for phoneme-viseme mapping and feature selection in visual speech recognition," in *Speechreading by Humans and Machines*, D. G. Stork and M. E. Hennecke, eds. (Springer, Berlin, Germany, 1996) 505–515.

[38] X. Zhang, C. C. Broun, R. M. Mersereau, and M. Clements, "Automatic speechreading with applications to human-computer interfaces," *EURASIP J. Appl. Signal Process.* 2002, 1228–1247 (2002).

[39] P. L. Silsbee and A. C. Bovik, "Computer lipreading for improved accuracy in automatic speech recognition," *IEEE Trans. Speech Audio Process.* 4, 337–351 (1996).

[40] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features," *EURASIP J. Appl. Signal Process.* 2002, 1213–1227 (2002).

[41] D. Chandramohan and P. L. Silsbee, "A multiple deformable template approach for visual speech recognition," in *Proc. Int.*

Conf. Spoken Lang. Processing (Philadelphia, PA, Oct. 3–6, 1996) 50–53.

[42] S. M. Chu and T. S. Huang, "Audio-visual speech modeling using coupled hidden Markov models," in *Proc. Int. Conf. Acoust.*, Speech, Signal Processing (Orlando, FL, May 13–17, 2002) 2009–2012.

[43] M. Gordan, C. Kotropoulos, and I. Pitas, "A support vector machine-based dynamic network for visual speech recognition applications," *EURASIP J. Appl. Signal Process.* 2002, 1248–1259 (2002).

[44] G. Potamianos, C. Neti, G. Gravier, et al., "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, 91, 1306–1326 (2003).

[45] P. S. Aleksic and A. K. Katsaggelos, "Comparison of low- and high-level visual features for audio-visual continuous automatic speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Montreal, Canada, May 17–21, 2004) 917–920.

[46] G. Potamianos, C. Neti, J. Huang, et al., "Towards practical deployment of audio-visual speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Montreal, Canada, May 17–21, 2004) 777–780.

[47] T. Chen, "Audiovisual speech processing. Lip reading and lip synchronization," *IEEE Signal Process.* 18, 9–21 (2001).

[48] J. J. Williams, A. K. Katsaggelos, and D. C. Garstecki, "Subjective analysis of an HMM-based visual speech synthesizer," in *Proc. SPIE Conf. Human Vision Electronic Imaging* (San Jose, CA, Jan. 21–25, 2001) 544–555.

[49] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, M. Magnenat-Thalmann and D. Thalmann, eds. (Springer-Verlag, Tokyo, Japan, 1993) 141–155.

[50] J. J. Williams and A. K. Katsaggelos, "An HMM-based speech-to-video synthesizer," *IEEE Trans. Neural Networks*, 13, 900–915 (2002).

[51] S. Morishima and H. Harashima, "A media conversion from speech to facial image for intelligent man-machine interface," *IEEE J. Select. Areas Commun.* 9, 594–600 (1991).

[52] P. S. Aleksic and A. K. Katsaggelos, "Speech-to-video synthesis using MPEG-4 compliant visual features," *IEEE Trans. Circuits Syst. Video Technol.* 14, 682–692 (2004).

[53] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proc. Int. Conf. Computer Graphics Interact. Techniques* (Los Angeles, CA, Aug. 3–8, 1997) 353–360.

[54] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Trans. Multimedia*, 2, 152–163 (2000).

[55] T. Ezzat and T. Poggio, "MikeTalk: A talking facial display based on morphing visemes," in *Proc. Computer Animation* (Philadelphia, PA, June 1998) 96–98.

[56] C. Pelachaud, N. Badler, and M. Steedman, "Linguistic issues in facial animation," in *Computer Animation*, N. Magnenat-Thalmann and D. Thalmann, eds. (Springer-Verlag, Berlin, Germany, 1991) 15–30.

[57] A. P. Breen, E. Bowers, and W. Welsh, "An investigation into the generation of mouth shapes for a talking head," in *Proc. Int. Conf. Spoken Lang. Processing* (Philadelphia, PA, Oct. 3–6, 1996) 2159–2162.

[58] F. I. Parke, "Parameterized models for facial animation," *IEEE Comput. Graph. Appl.* 2, 61–68 (1982).

[59] S. Pasquariello and C. Pelachaud, "Greta: A simple facial animation engine," in *6th Online World Conference on Soft Computing in Industrial Appications* (Sept. 2001).

[60] D. W. Massaro and M. M. Cohen, "Perception of synthesized audible and visible speech," in *Psychol. Science*, 1, 55–63 (1990).

[61] F. Lavagetto, "Time-delay neural networks for estimating lip movements from speech analysis: A useful tool in audio/video synchronization," *IEEE Trans. Circuits Syst. Video Technol.* 7, 786–800 (1997).

[62] A. D. Simons and S. J. Cox, "Generation of mouthshapes for a synthetic talking head," *Proc. Inst. Acoustics*, 12, 475–482 (1990).

[63] U. Bub, M. Hunke, and A. Waibel, "Knowing who to listen to in speech recognition: Visually guided beamforming," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Detroit, MI, May 9–12, 1995) 848–851.

[64] C. Wang and M. S. Brandstein, "Multi-source face tracking with audio and visual data," in *Proc. Works. Multimedia Signal Processing* (Copenhagen, Denmark, Sept. 13–15, 1999) 475–481.

[65] D. N. Zotkin, R. Duraiswami, and L. S. Davis, "Joint audio-visual tracking using particle filters," *EURASIP J. Appl. Signal Process.* 2002, 1154–1164 (2002).

[66] G. Iyengar, H. J. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in *Proc. Int. Conf. Acoust.*, Speech, Signal Processing (Hong Kong, China, 2003) 772–775.

[67] P. De Cuetos, C. Neti, and A. Senior, "Audio-visual intent to speak detection for human computer interaction," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Istanbul, Turkey, June 5–9, 2000) 1325–1328.

[68] J. Luettin, "Speaker verification experiments on the XM2VTS database," IDIAP Research Institute, Martigny, Switzerland, Research Report 99-02, Jan. 1999.

[69] B. Maison, C. Neti, and A. Senior, "Audio-visual speaker recognition for broadcast news: some fusion techniques," in *Proc. Works. Multimedia Signal Processing* (Copenhagen, Denmark, Sept. 13–15, 1999) 161–167.

[70] T. Wark, S. Sridharan, and V. Chandran, "Robust speaker verification via fusion of speech and lip modalities," in *Proc. Int. Conf. Acoust., Speech, Signal Processing* (Phoenix, AZ, Mar. 15–19, 1999) 3061–3064.

[71] C. Sanderson and K. K. Paliwal, "Information fusion and person verification using speech and face information," IDIAP Research Institute, Martigny, Switzerland, Research Report 02–33, Sept. 2002.

[72] U. V. Chaudhari, G. N. Ramaswamy, G. Potamianos, and C. Neti, "Information fusion and decision cascading for audiovisual speaker recognition based on time-varying stream reliability prediction," in *Proc. Int. Conf. Multimedia Expo* (Baltimore, MD, July 6–9, 2003) 9–12.

[73] P. S. Aleksic and A. K. Katsaggelos, "An audiovisual person identification and verification system using FAPs as visual features," in *Proc. Works. Multimedia User Authentication* (Santa Barbara, CA, Dec. 11–12, 2003) 80–84.

[74] U. Dieckmann, P. Plankensteiner, and T. Wagner, "SESAM: A biometric person identification system using sensor fusion," *Pattern Recognition Lett.* 18, 827–833 (1997).

[75] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, "Acoustic-labial speaker verification," *Pattern Recognition Lett.* 18, 853–858 (1997).

[76] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. Pattern Anal. Machine Intell.* 17, 955–966 (1995).

[77] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision Image Understanding*, 91, 160–187 (2003).

[78] P. S. Aleksic and A. K. Katsaggelos, "Automatic facial expression recognition using facial animation parameters," *IEEE Trans. Signal Process.* 2005 Supplement on secure media, Oct. 2005.

[79] A. W. Senior, "Face and feature finding for a face recognition system," in *Proc. Int. Conf. Audio Video-based Biometric Person Authentication* (Washington, DC, Mar. 22–23, 1999) 154–159.

[80] H. A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *IEEE Trans. Pattern Anal. Machine Intell.* 20, 23–38 (1998).

[81] K.-K. Sung and T. Poggio, "Example-based learning for view-based human face detection," *IEEE Trans. Pattern Anal. Machine Intell.* 20, 39–51 (1998).

[82] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. Conf. Computer Vision Pattern Recognition* (Kauai, HI, Dec. 11–13, 2001) 511–518.

[83] H. P. Graf, E. Cosatto, and G. Potamianos, "Robust recognition of faces and facial features with a multi-modal system," in *Proc. Int. Conf. Systems, Man, Cybernetics* (Orlando, FL, 1997) 2034–2039.

[84] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.* 4, 321–331 (1988).

[85] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature extraction from faces using deformable templates," *Int. J. Comput. Vis.* 8, 99–111 (1992).

[86] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. Europ. Conf. Computer Vision* (Freiburg, Germany, 1998) 484–498.

[87] S. Young, D. Kershaw, J. Odell, et al., *The HTK Book* (Entropic Ltd., United Kingdom, 1999).

[88] I. S. Pandzic and R. Forchheimer, Eds., *MPEG-4 Facial Animation: The Standard, Implementation and Applications* (John Wiley and Sons, United Kingdom, 2002).

[89] J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals* (Macmillan Publishing, Englewood Cliffs, NJ, 1993).

[90] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Machine Intell.* 22, 4–37 (2000).

[91] L. E. Bernstein, *Lipreading Corpus V-VI: Disc 3,* (Gallaudet University, Washington, DC, 1991).

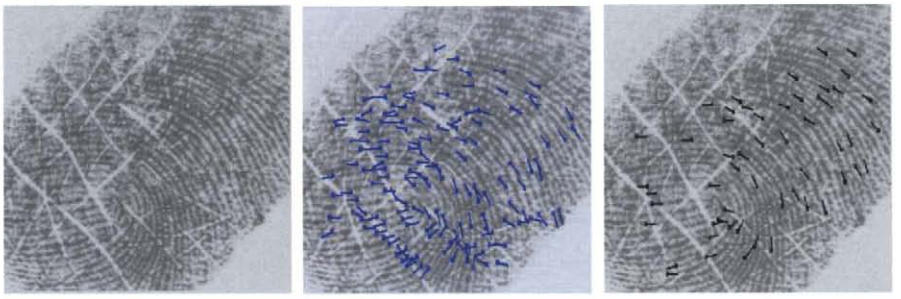
[92] A. Löfqvist, "Speech as audible gestures," in *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, eds. (Kluwer Academic, Dordrecht, The Netherlands, 1990) 289–322.

[93] G. A. Abrantes, "FACE - Facial Animation System, version 3.3.1," Instituto Superior Tecnico, 1997–98.

[94] T. Dutoit, *An Introduction to Text-To-Speech Synthesis.* (Kluwer Academic, Dordrecht, The Netherlands, 1997).

[95] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.* 14, 4–20 (2004).

[96] P. Ekman and W. Friesen, *Facial Action Coding System* (Consulting Psychologists Press Inc., Palo Alto, CA, 2003).



(a) (b) (c)

FIGURE 10.5.8 Fingerprint Enhancement Results: (a) a poor quality fingerprint; (b) minutiae extracted without image enhancement; and (c) minutiae extracted after image enhancement [11].

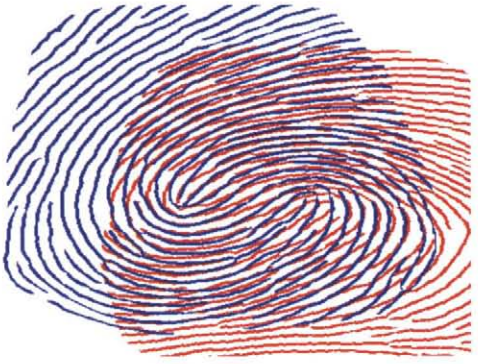


FIGURE 10.5.13 Aligned ridge structures of mated pairs. Note that the best alignment in one part (mid-left) of the image results in large displacements between the corresponding minutiae in the other regions (bottom right) [18]. ©IEEE.

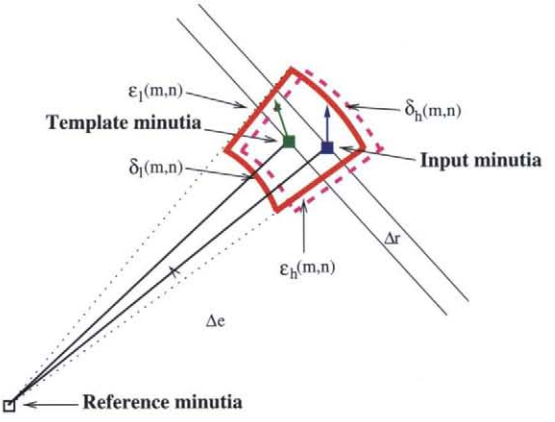
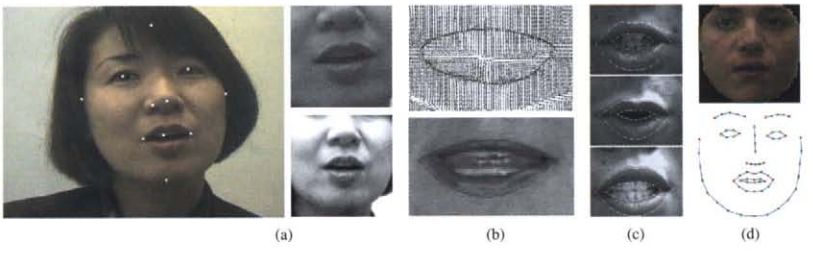


FIGURE 10.5.14 Bounding box and its adjustment [18]. ©IEEE.



(a) (b) (c) (d)

FIGURE 10.8.2 Mouth appearance and shape tracking for visual feature extraction. **A**: Eleven detected facial features using the appearance-based approach of [79]. Two corresponding mouth region-of-interests of different sizes and normalization are also depicted [44]. **B**: Lip contour estimation using a gradient vector field snake (**upper:** the snake's external force field is depicted) and two parabolas (**lower**) [40]. **C**: Three examples of lip contour extraction using an active shape model [31]. **D**: Detection of face appearance (**upper**) and shape (**lower**) using active appearance models [29].

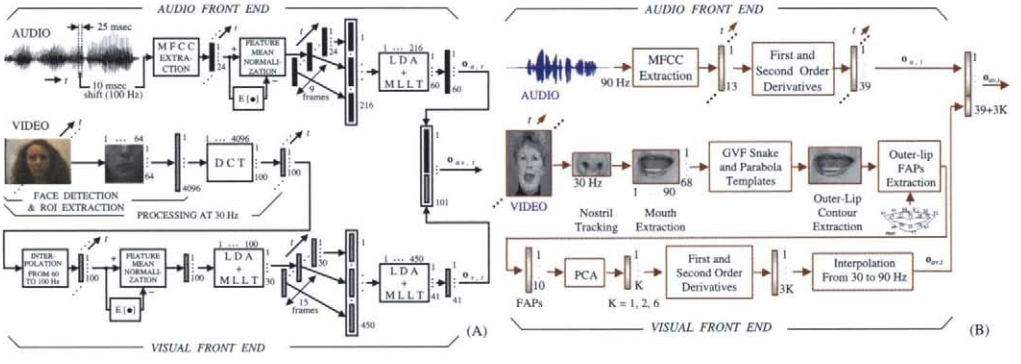


FIGURE 10.8.5 Two implementations of visual feature extraction, depicted schematically in parallel with the audio front end, as used for audiovisual automatic speech recognition experiments in this chapter: **A**: The appearance-based visual front end system of IBM Research, also employed for bimodal speaker recognition [72] and audio enhancement [19]; **B**: the shape-based system of Northwestern University [40], also used for speech-to-video synthesis [52] and audiovisual speaker recognition [73].

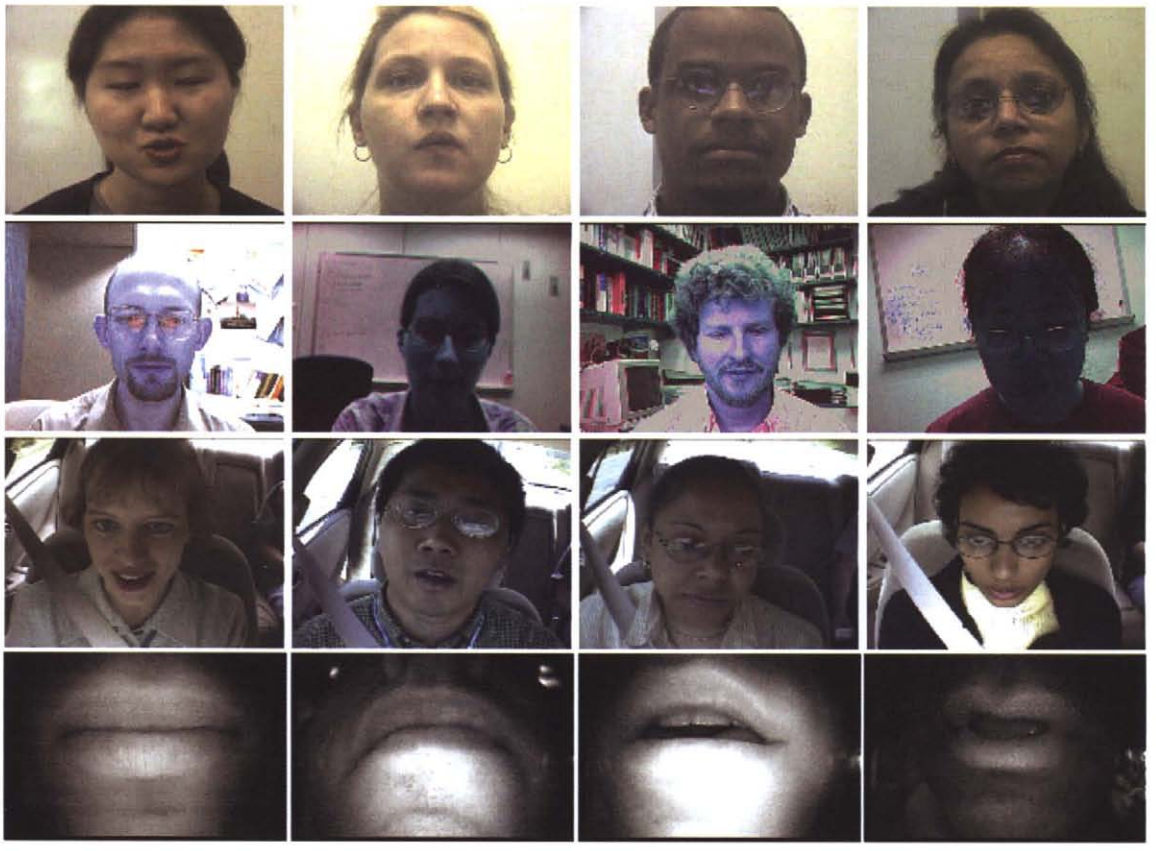


FIGURE 10.8.9 Example frames from the four IBM audiovisual automatic speech recognition corpora discussed in Sections 4.1 and 4.2. Top-to-bottom: Full-face data collected in the studiolike, office, and car environments; bottom line: Region of interest–only data captured by a specially designed headset [46].

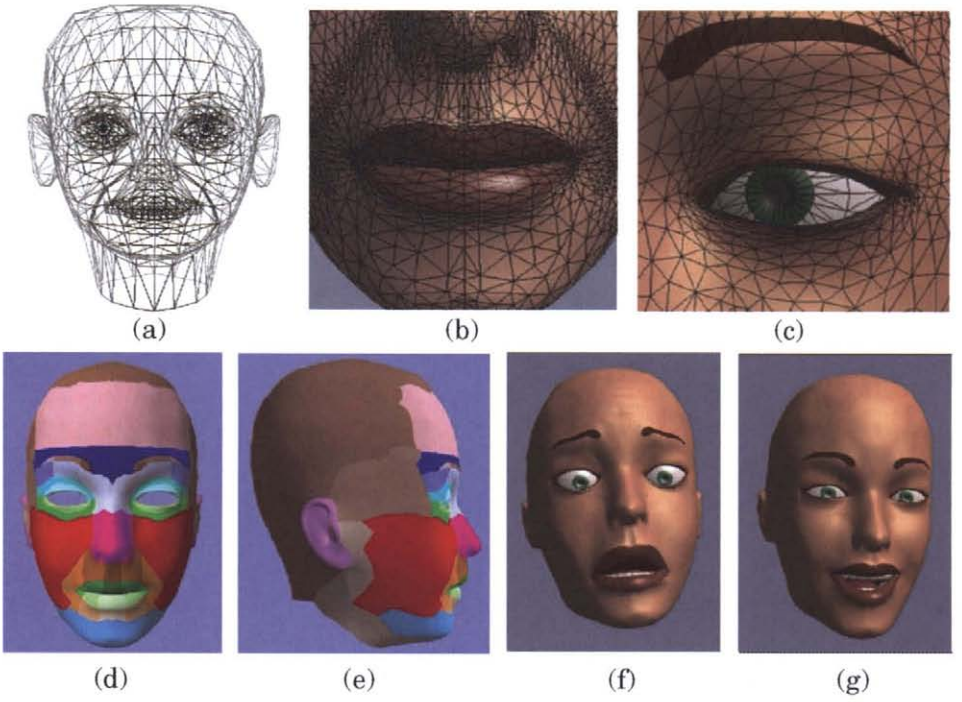


FIGURE 10.8.11 MPEG-4 compliant facial animation. (A): A polygonal mesh [93]; (B,C) detailed structure of the most expressive face regions; (D,E) three-dimensional surface is divided into areas corresponding to feature points affected by facial animation parameters and (F,G) synthesized expressions of fear and joy. (B–G) Correspond to model *Greta* (reproduced with permission from [59]).

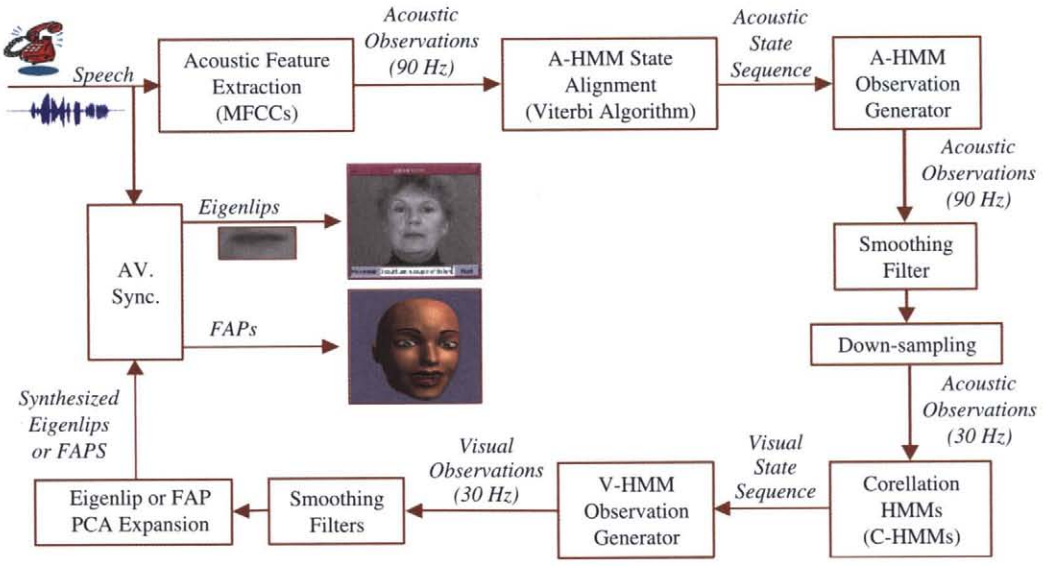


FIGURE 10.8.13 The speech-to-video synthesis systems developed in [50, 52] utilize narrowband speech to generate two possible visual representations: eigenlips that can be superimposed on frontal face videos for animation, or facial animation parameters that can be used to drive an MPEG-4–compliant facial animation model.

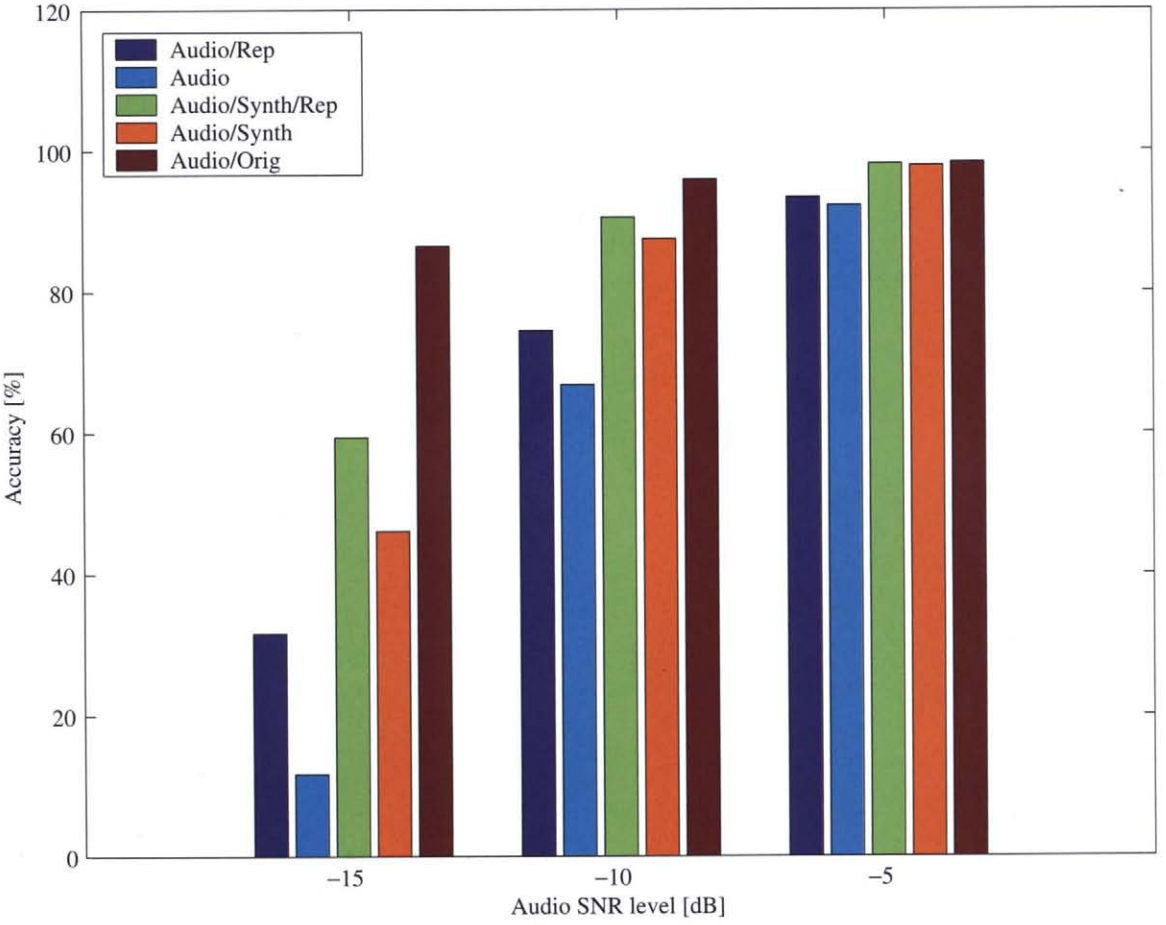


FIGURE 10.8.14 Intelligibility-based subjective evaluation of the speech-to-video synthesis system developed at Northwestern University [48, 50]. Human speech perception is compared using audio-only vs. audio with synthesized video and vs. audio with natural video of the lip region. For the first two conditions, results for repetitive presentation of the stimuli to the subjects are also given ("Rep"). Experiments are reported over three acoustic noise conditions. SNR, signal-to-noise ratio.

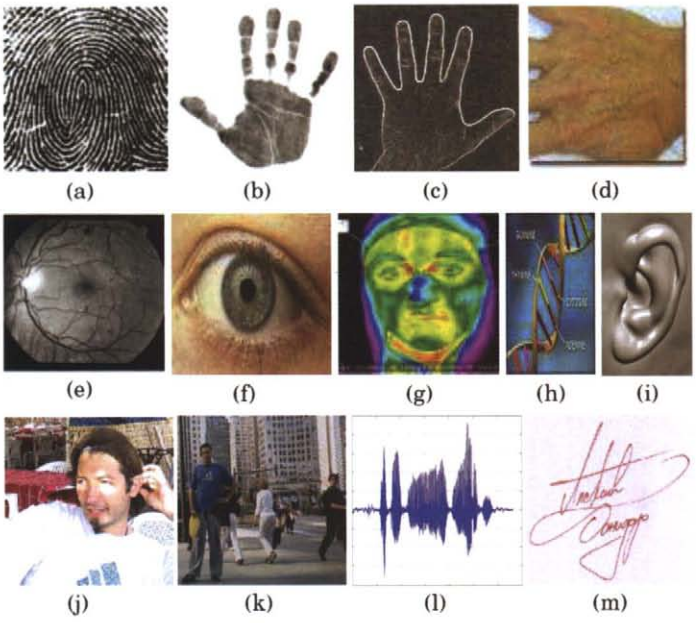


FIGURE 10.8.15 Biometric characteristics: (A) fingerprints; (B) palm print; (C) hand and finger geometry; (D) hand veins; (E) retinal scan; (F) iris; (G) infrared thermogram; (H) DNA; (I) ears; (J) face; (K) gait; (L) speech; (M) signature.

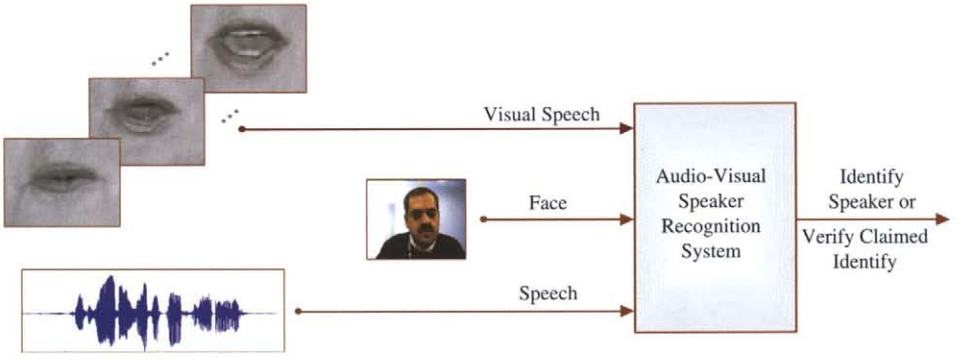


FIGURE 10.8.16 Block diagram of an audiovisual speaker recognition system that utilizes static (face image) and dynamic (visual speech) visual information together with acoustic information.

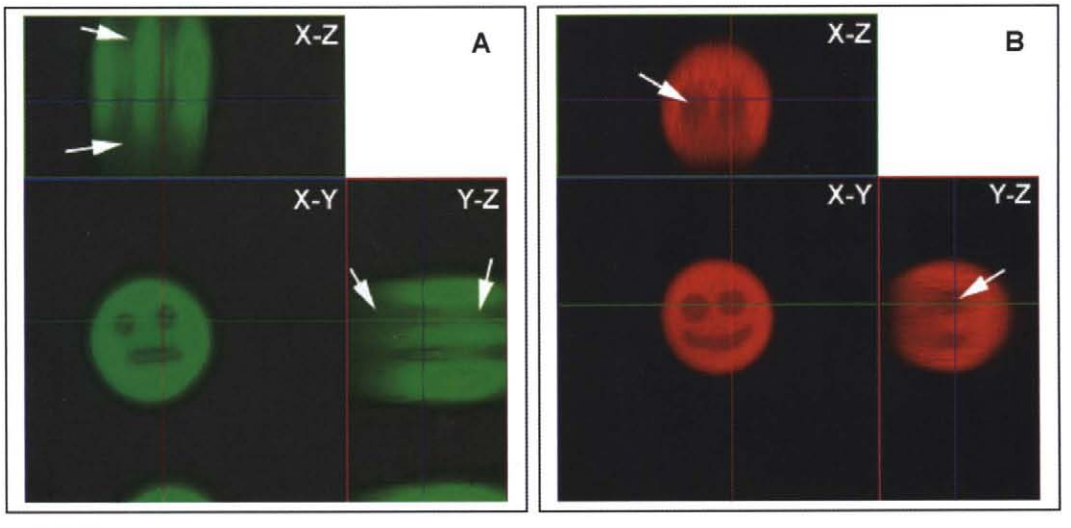


FIGURE 10.9.7 Fluorescence excitation in (a) confocal microscopy, and (b) two-photon microscopy. The bleach pattern can be viewed as a footprint of the excitation region. While in confocal microscopy (a), illumination results in excitation throughout the whole thickness of the specimen, the two-photon excitation results in illumination being restricted to the focal plane (b).