Audio-Visual Integration in Multimodal Communication

TSUHAN CHEN, MEMBER, IEEE, AND RAM R. RAO

Invited Paper

In this paper, we review recent research that examines audio-visual integration in multimodal communication. The topics include bimodality in human speech, human and automated lip reading, facial animation, lip synchronization, joint audio-video coding, and bimodal speaker verification. We also study the enabling technologies for these research topics, including automatic facial-feature tracking and audio-to-visual mapping. Recent progress in audio-visual research shows that joint processing of audio and video provides advantages that are not available when the audio and video are processed independently.

Keywords*—*Image analysis, multimedia communication, speech communication, speech processing, video signal processing.

I. INTRODUCTION

Multimedia is more than simply the combination of various forms of data: text, speech, audio, music, images, graphics, and video. When we discuss multimedia signal processing, it is the integration and interaction among these different media types that create challenging research topics and new opportunities. Fig. 1 illustrates research topics that deal with various types of media. In this figure, media are categorized into three major classes: the first is textual information; the second is audio, which includes speech and music; and the third is image and video. Traditional information-processing techniques usually focus on one media type. For example, language translation and natural language processing are related to the processing of text data. Likewise, compression and synthesis techniques have been developed for audio signals. Similarly, key research areas in image and video processing include compression, synthesis with computer graphics, and image indexing and retrieval.

More interesting research topics can be found, however, when we exploit the interaction among different media types. For example, using speech-recognition technology, one can analyze speech waveforms to discover the text that has been spoken. From a sentence of text, a talking-head audio-visual sequence can be generated using computer graphics to animate a facial model and text-to-speech synthesis to provide synthetic acoustic speech.

Among the possible interactions among different media types, the interaction between audio and video is the most interesting. A recent trend in multimedia research is to integrate audio and visual processing to exploit such interaction. For multimedia applications that involve person-to-person conversation, e.g., video telephony and video conferencing, such interaction is especially important. Research topics along the direction of audio-visual integration include automatic lip reading, lip synchronization, joint audio-video coding, and bimodal person authentication. The main purpose of this paper is to report recent progress in these research areas. The results described in this paper will demonstrate that major improvement can be obtained by using joint audio-video processing compared to the situation where audio and video are processed independently.

A. Outline

We begin with the introduction of the bimodality in human speech production and perception in Section II. In Section III, we describe human lip reading and review results in automated lip reading and in using lip reading to enhance speech recognition. We study speech-driven face animation in Section IV. Section V is devoted to speech-assisted lip synchronization. Following that, we discuss the enabling technologies for all audio-visual research, including facial-feature tracking in Section VI and audio-to-visual mapping in Section VII. In Section VIII, we will outline recent research in bimodal person verification. We will conclude the paper with some remarks in Section IX.

II. BIMODALITY OF HUMAN SPEECH

Audio-visual interaction is important in multimodal communication. In multimodal communication, where human

Manuscript received August 1, 1997; revised January 15, 1998. The Guest Editor coordinating the review of this paper and approving it for publication was A. M. Tekalp. The work of R. R. Rao was supported in part by the Joint Services Electronics Program under Contract DAAH-04-96-1-0161.

T. Chen was with AT&T Bell Laboratories, Holmdel, NJ 07733 USA. He is now with the Department of Electrical and Computer Engineering, Carnegie-Mellon University, Pittsburgh, PA 15213 USA (e-mail: tsuhan@ece.cmu.edu).

R. R. Rao was with AT&T Bell Laboratories, Holmdel, NJ 07733 USA. He is now with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332-1230 USA (e-mail: rr@eedsp.gatech.edu).

Publisher Item Identifier S 0018-9219(98)03484-7.

0018–9219/98$10.00 © 1998 IEEE

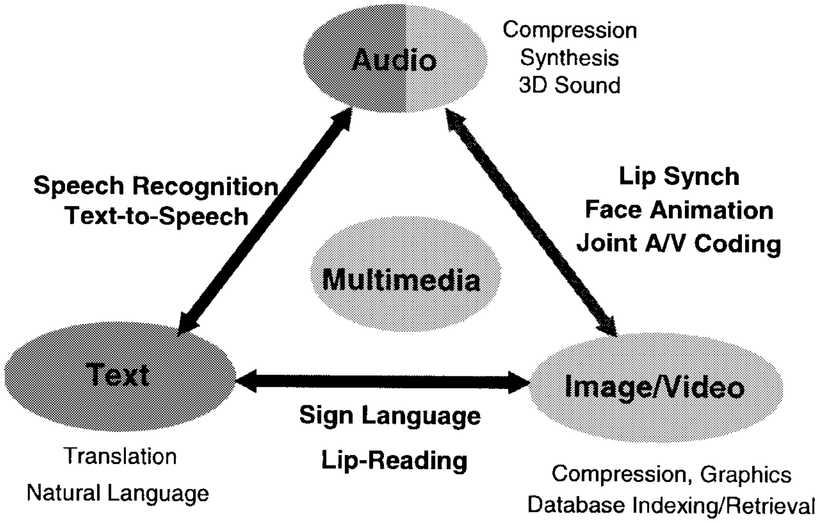


Fig. 1. Media interaction.

Table 1 Examples of McGurk Effect

Audio +	Visual →	Perceived
ba	ga	da
pa	ga	ta
ma	ga	na

speech is involved, audio-visual interaction is particularly significant because human speech is bimodal in nature. The "McGurk effect" [1] clearly demonstrates the bimodal nature of human speech perception. When humans are presented with conflicting audio and visual stimuli, the perceived sound may not exist in either modality. For example, when a person "hears" the sound /ba/ but "sees" the speaker saying /ga/, the person may not perceive either /ga/ or /ba/. Instead, what is perceived is something close to /da/. Some other examples of audio-visual combinations are shown in Table 1. This shows that the speech that is perceived by a person depends not only on acoustic cues but also on visual cues such as lip movements.

Psychologists have extensively studied the McGurk effect. They have shown that there also exists the "reverse McGurk effect," i.e., the results of visual speech perception can be affected by the dubbed audio speech [2]. In addition to artificially chosen consonant-vowel syllables such as those in Table 1, the existence of both the McGurk effect and the reverse McGurk effect in *natural* speech has also been proven by a series of experiments in [3]. The McGurk effect has been shown to occur across different languages [4] and even in infants [5]. The McGurk effect is also robust to a variety of different conditions. The same answers are given to the conflicting stimuli in cases when there are timing mismatches between the stimuli, or even when the face of a male speaker is combined with the voice of a female speaker [6].

Due to the bimodality in speech perception, audio-visual interaction becomes an important design factor for multimodal communication systems, such as video telephony and video conferencing. There has been much research that shows the importance of combined audio-visual testing for bimodal perceptual quality of video-conferencing systems [7], [8].

In addition to the bimodal characteristics of speech perception, speech production is also bimodal in nature. Human speech is produced by the vibration of the vocal cord and the configuration of the vocal tract that is composed of articulatory organs, including the nasal cavity, tongue, teeth, velum, and lips. Using these articulatory organs, together with the muscles that generate facial expressions, a speaker produces speech. Since some of these articulators are visible, there is an inherent relationship between the acoustic and visible speech.

The basic unit of acoustic speech is called a *phoneme*. In American English, the ARPABET table, consisting of 48 phonemes, is commonly used [9]. Similarly, in the visual domain, the basic unit of mouth movements is called a *viseme* [10]. A viseme therefore is the smallest visibly distinguishable unit of speech. There are many acoustic sounds that are visually ambiguous. These sounds are grouped into the same class that represents a viseme. There is therefore a many-to-one mapping between phonemes and visemes. For example, the /p/, /b/, and /m/ phonemes are all produced by a closed mouth shape and are visually indistinguishable. Therefore, they form one viseme group. Similarly, /f/ and /v/ both belong to the same viseme group that represents a mouth of which the upper teeth are touching the lower lip. Different words that *appear* the same, e.g., "pad," "bat," and "man," are said to be *homophenous*. Viseme groupings are not all based on place of articulation. For example, the /k/ and /t/ phonemes may or may not belong to the same viseme grouping. Clearly, these consonants are formed differently; the /k/ is velar and the /t/ is palatal. The difference in their formation, however, may not be visible. Some studies classify these two phonemes in the same viseme group while others do not.

Table 2 Viseme Groups for English Consonants

1	f, v
2	th, dh
3	s, z
4	sh, zh
5	p, b, m
6	w
7	r
8	g, k, n, t, d, y
9	l

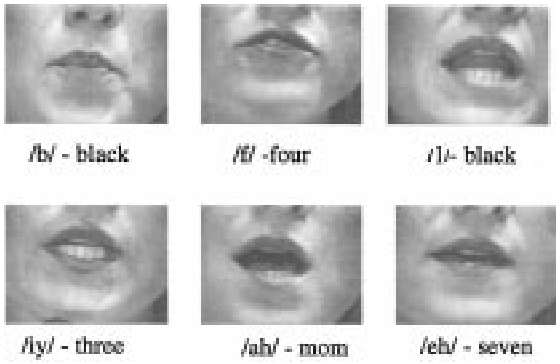


Fig. 2. Example visemes.

Most of the viseme groupings in the literature [11] are obtained by analyzing the confusions in stimulus-response matrices. In the experiments, subjects are asked to identify syllables visually in a given context such as vowel-consonant-vowel (VCV) words. The stimulus–response matrices are then tabulated. Viseme groups are identified as those clusters of phonemes in which at least 75% of all responses occur within the cluster. Fisher did one of the first studies of visemes [10]. He asked subjects visually to identify both initial and final consonants in consonant-vowel-consonant (CVC) and found that the viseme groupings for initial and final consonants differed. He also found that the confusions between consonants in a viseme class could be directional. For example, he found that the /n/ stimulus would often be confused with /t/. The /t/ stimulus, however, would rarely be confused with /n/. Last, Walden *et al.* showed that the viseme grouping could differ significantly for each speaker and with training [12].

Unlike phonemes, currently there is no standard viseme table used by all researchers [13]. It is commonly agreed, however, that the visemes for English consonants can be grouped into nine distinct groups, as in Table 2 [14]. In Fig. 2, we show a number of visemes. The top three are associated with consonants and the bottom three are associated with vowels. Strictly speaking, instead of a still image, a viseme can be a sequence of several images that capture the movements of the mouth. This is especially true for some vowels. For example, the viseme /ow/ represents the movement of the mouth from a position close to /o/ to a position close to /w/. Therefore, to illustrate some visemes, we would need to use video sequences. However, most visemes can be approximated by stationary images.

Both in acoustic modality and in visual modality, most of the vowels are distinguishable [14]. However, the same is not true for consonants. For example, in the acoustic domain, the sounds /p/, /t/, and /k/ are very similar. When people spell words on the phone, expressions such as “B as in boy” or “D as in David,” are often used to clarify such acoustic confusion. Interestingly enough, the confusion sets in the auditory modality are usually distinguishable in the visual modality (again, illustration of bimodality in speech perception). One good example is the sounds /p/ and /k/, which can be easily distinguished by the visual cue of a closed mouth versus an open mouth.

III. Lip Reading

A good example of utilizing audio-visual interaction for human speech communication is lip reading [15, pp. 73–107], also referred to as speech reading. In this section, we introduce human lip reading and automated lip reading, i.e., lip reading done by a computer.

A. Human Lip Reading

A person skilled in lip reading is able to infer the meaning of spoken sentences by looking at the configuration and the motion of visible articulators of the speaker, such as the tongue, lips, teeth, etc., together with clues from the context. Knowledge of the positions of these articulators provides information about the content of the acoustic speech signal. Because all of the articulators are not visible, this information may sometimes be incomplete. By combining the visual content with lexical, syntactic, semantic, and pragmatic information, people can learn to understand spoken language by observation of the movements of a speaker’s face. Lip reading is widely used by hearing-impaired persons for speech understanding.

In addition to lip reading, there are many other ways in which humans can use their sight to assist in aural communication. First, the visual signal can be used to help focus attention. In addition to the binaural mechanism, one can observe which speaker is talking to help separate the speech from background noise. Second, visible speech also provides a supplemental information source that is useful when the listener has trouble comprehending the acoustic speech. Difficulties may arise for a variety of reasons, including the presence of background noise or competing speakers, such as during a cocktail party. For these situations, even people who are not hearing impaired also utilize lip reading to some extent [16]. Furthermore, listeners may also have trouble comprehending the acoustic speech in situations where they lack familiarity with the speaker, such as listening to a foreign language or an accented talker. In this case, the visual channel may provide an auxiliary channel that aids comprehension.

A study by Sumby and Pollack [17] showed that when noisy environments are encountered, visual information can lead to significant improvement in recognition by humans. They found that the recognition of audio-visual cues degrades less rapidly than acoustic cues alone. As an example, for an eight-word vocabulary, auditory recognition was near 100% at 0 dB of signal-to-noise ratio (SNR) but fell under 20% at −30-dB SNR. However, audio-visual recognition only dropped from 100 to 90% over the same range. Specifically, they showed that the presence of the visual signal is roughly equivalent to a 12-dB gain in SNR. Lip reading can also be useful when the speech is not degraded by noise. Reisberg *et al.* [18] showed that speech that is difficult to understand for reasons other than background noise can be more easily recognized with the addition of a visual signal. In his study, he showed that speech consisting of sentences from a foreign language or spoken by an accented speaker were better understood when the speaker's face was visible.

Lip-reading performance depends on a number of factors. Viewing conditions may affect the quality of the visual information. Poor lighting may make it difficult to judge the shape of the mouth or to detect the teeth or tongue. Likewise, as the speaker and listener move further apart, it becomes more difficult to view important visual cues. Factors such as viewing angle can also affect recognition. In [19], Neely found that a frontal view of the speaker led to higher recognition rates than an angled or profile view. Lip-reading performance can also be improved through training, as shown by Walden *et al.* in [12]. For example, before training, an /s/ or /z/ would be confused with a /th/ or /dh/ by a person. After training, these confusions can be eliminated.

Last, *coarticulation* can affect lip-reading performance. Coarticulation is the process by which one sound affects the production of neighboring sounds. Berger states that the place of articulation may not be fixed but may depend on context [20]. He suggests that the tongue may be in different positions for the /r/ sound in "art" and "arc." The tongue would be more forward in the word "art." This would affect the visibility of the tongue and thus recognition accuracy. In [21], Bengeruel and Pinchora-Fuller examined syllables with a VCV context. They found that consonant recognition depended on the vowels that surrounded it. For example, the middle consonant is more difficult to lip read when either vowel is a /u/ as opposed to an /ae/ or /i/. At the same time, the /u/ was the most recognizable of the vowels. This suggests that the /u/ sound is visually dominant. Its appearance is pronounced, and because of this, there is a recovery period where neighboring sounds may be masked.

Clearly, video-conferencing systems must be able to provide the full motion necessary for speech reading by the hearing impaired. Frowein *et al.* have studied the importance of frame rates with impaired listeners [22]. In these experiments, the subjects were presented with audio clips containing human speech and audio-video clips containing the same speech. The results showed that for speech understanding, a video refresh rate (frame rate) of 15 Hz was necessary. Below 15 Hz, the speech-recognition score would drop significantly. In all the cases, the presence of video, which allowed the hearing-impaired subjects to lip read, always improved the speech-recognition score, sometimes by as much as 90%, compared to the audio-only case. The experiment results also showed that the added value of audio-visual recognition over audio-only recognition was further enhanced by the presence of background noise in the audio signals. Williams *et al.* analyzed the effects of frame rates on isolated viseme recognition in human lip reading [23]. In contrast to the experiments in [22], the experiments here use subjects who have no prior experience of lip reading and who are shown video-only sequences (no audio-visual sequences) during the test phase. It was shown that at different frame rates—30, 15, 10, 5, and 2 Hz—the viseme groups derived from visual similarity among sounds could be very different. In addition, it was found that the minimum frame-rate requirement for continuous speech should be greater than 5 Hz.

Knowledge of the process by which humans extract and incorporate visual information into speech perception can be beneficial. Specifically, we would like to know what information is available in the visual channel, what types of parameters humans use to aid recognition, and what means are used to integrate the acoustic and visual information. The acoustic and visual components of the speech signal are not purely redundant; they are complementary as well. Certain speech characteristics that are visually confusable are acoustically distinct, while those characteristics that are acoustically confusable are visually distinct. For instance, the /p/ and /k/ phonemes have similar acoustic characteristics, but can easily be differentiated by the closing of the mouth. In contrast, the /p/ and /b/ phonemes have similar visual characteristics but can be acoustically differentiated by voicing. One question that remains to be answered is how the human brain takes advantage of the complementary nature of audio-visual speech and fuses the two information sources. There are two human perceptual results that a successful fusion theory must explain. First, the theory must justify the improvement in recognition that arises from viewing a person's face when background acoustic noise is present. The second perceptual result that a fusion theory should explain is the McGurk effect. There are a variety of hypotheses [24], [25] that abound regarding the fusion of audio and visual information. One supposition advocates a system where the visual information is instinctively converted to an estimate of the vocal-tract transfer function, and the two sources are combined by averaging the two transfer functions. The more widespread beliefs about how human integrate acoustic and visual information together can be classified into two different camps: early integration and late integration. In early integration, as shown in Fig. 3(a), the acoustic and visual parameter sets are combined into a larger parameter set. Recognition occurs by finding the word whose template is matched best to the audio-visual parameter sets. In late integration, as shown in Fig. 3(b), the audio is compared against an acoustic template for each word, and the video is compared against a visual template for each word. The resulting audio and visual recognition scores are then combined using a certain mechanism.

B. Automated Lip Reading

There has been much research in recent years into implementing audio-visual speech-recognition systems with

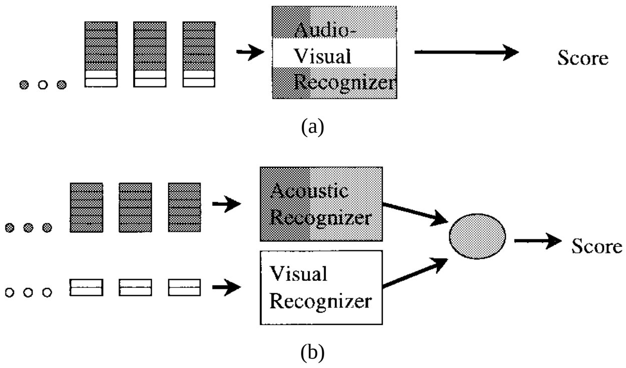


Fig. 3. (a) Early integration and (b) late integration.

computers. This is often referred to as automated lip reading. The recognition systems that have been designed vary widely and have been used to help answer a number of questions concerning audio-visual recognition. First, many systems have been designed to show that speech recognition is possible using only the visual information. Some researchers have done comparisons on a number of visual feature sets in attempts to find those features that yield the best recognition performance. Next, researchers have attempted to integrate these visual-only recognition systems with acoustic recognition systems in order to enhance the accuracy of the acoustic speech-recognition system. A number of studies have examined the competing fusion strategies of early integration, late integration, and other novel means for combining audio and video. Other studies have also done more thorough investigations into the resiliency of audio-visual recognition systems to varying levels of acoustic noise. Last, many of the systems presented in the literature have attacked different recognition tasks, implementing speaker-dependent and speaker-independent systems and examining isolated vowels, CVC syllables, isolated words, connected digits, and continuous speech.

Perhaps because of the wide variety of tasks that have been analyzed, there is no clear consensus on what an optimal audio-visual recognizer should look like. The recognizers that have been designed vary in the acoustic feature sets, visual feature sets, recognition engine, and integration strategies. The visual feature sets provide some of the greatest variation between systems. Systems range from performing relatively little processing and having a visual input that consists of a rectangular region of the image to using computer vision techniques to extract a visual feature set to be used for recognition. Because of this large variation, researchers have made attempts to compare the effectiveness of the various visual feature sets. The recognition engine also takes many forms among the various research groups. Some of the earlier recognition systems were based on dynamic time warping. Since then, there have been a number of systems that use neural-network architectures and an increasing number of systems that rely on hidden Markov models (HMM's) for recognition. Last, there have been a number of integration strategies that have been proposed in the literature. Some of the early systems used either sequential recognition or a rule-based approach. One system converted the visual signal into an acoustic power spectrum and averaged the information sources in this domain. Other systems have employed either early or late integration strategies and compared the differences between the two.

Petajan developed one of the first audio-visual recognition systems [26]. In this system, a camera captures the mouth image and thresholds it into two levels, i.e., black and white. Binary mouth images are analyzed to derive the mouth open area, the perimeter, the height, and the width. These parameters are then used for recognition. In this system, the audio speech recognizer and the visual speech recognizer are combined in a serial fashion. In other words, the speech is processed by the acoustic recognizer first to produce the first few candidate words, and then these words are passed on to the visual recognizer for final decision. Later, Petajan *et al.* modified the system to improve the recognition performance [27]. Instead of using the height, width, area, and perimeter values, the binary mouth images themselves were used as the visual feature. A clustering scheme was used to classify these binary images. Furthermore, to compare sequence of images, dynamic time warping was used. Last, the audio-visual integration strategy was changed to a rule-based approach from the sequential integration.

The visual analysis system used by Petajan was later used by Goldschen to recognize continuous speech visually [13]. Goldschen analyzed a number of features of the binary images such as height, width, and perimeter, along with derivatives of these quantities, and used these features as the input to an HMM-based visual recognition system. He sought to find the combination of parameters that would lead to the best recognition performance. The feature set he settled upon was dominated by derivative information. This showed that the dynamics of the visual feature set is also useful for speech recognition. The same conclusion had been reached by Mase and Pentland [28] in a study that used optical flow as input for a visual speech recognizer.

Studies by Finn and Montgomery [29] have shown that the physical dimensions of the mouth can provide good recognition performance. They analyzed recognition performance of VCV syllables. They placed reflective markers on the speaker's mouth and used these to extract 14 distances, sampled at 30 frames/s. They experimented with both a mean squared error distance and an optimized weighted mean squared error distance. Equal weighting of the parameters led to a 78% viseme recognition rate. More impressively, they found that a weighted mean squared error using just five of the distances could yield a 95% recognition rate.

In [30], Yuhas *et al.* used neural networks to fuse information from the acoustic channel and information from the visual channel in a very interesting way. In this system, the pixel values of the mouth image are fed to a multilayer network directly. That is, no feature extraction for the mouth height or the mouth width was performed. The network was trained to estimate the acoustic spectrum based on the mouth image. The estimated spectrum is than combined with the true (measured) spectrum, with weighting that is determined by the noise in the acoustic channel. The combined spectrum is then fed to a recognition system. For the vowel-recognition task, it was shown that the combined system outperformed the audio-only recognizer.

In [31], Stork *et al.* used the time-delayed neural networks (TDNN's) for recognition. A TDNN can take into account the coarticulation in speech perception in that it would perform recognition based on the temporal variation of mouth parameters rather than using simply static images. Again, the reported results verified that joint audio-video recognition could outperform both the audio-only system and the video-only system. The audio-visual features were recognized using both early and late integration strategies. Stork found that late integration produced slightly better results than an early integration strategy. Furthermore, he found that with a late integration approach, he could replicate the McGurk effect with the recognition system.

Luettin [32] used active shape models to extract visual features and compared the performance of the static feature set with that of a feature set that included derivative information. He used HMM's with an early integration strategy for both speaker-independent isolated digits and speaker-independent connected digits. The isolated digit task was first tested with static versus static and dynamic features, and it was found that the dynamic features offered some improvements. Next, performance as a function of acoustic noise was examined. At high SNR, there was little gain. As the noise increased and acoustic performance degraded, the visual information was equivalent to approximately a 10-dB gain in SNR. At low SNR, when the acoustic recognition failed essentially, the audio-visual recognizer reached a performance level equivalent to using visual information alone. Silsbee developed a system by combining an acoustic speech recognizer and an automatic lip reader with a fuzzy logic [33]. It gave better performance in audio-visual speech recognition than recognition using either subsystem alone. He also pointed out that the enhancement obtained by joining audio and visual speech recognizers could be more significant at the phoneme level than at the word level. In [34], Potamianos *et al.* combined the visual features, either geometric parameters such as the mouth height and width or nongeometric parameters such as the wavelet transform of the mouth images, with the audio features to form a joint feature vector. Test data of these joint feature vectors were used to train an HMM-based speech recognizer. It was shown that joint audio-visual speech recognition outperformed both the visual-only recognition and the audio-only recognition.

Here, we present some of our recognition results from training an isolated word recognizer with audio-visual data for the digits "zero" to "nine." The audio was analyzed to provide eight Mel cepstral coefficients [9] along with the delta coefficients every 80 ms. The 30-Hz video was analyzed to extract the height and width of the outer contour of the lips. These visual parameters were upsampled and synchronized with the acoustic parameters. Four HMM-based recognition systems were trained. Each continuous HMM had ten states with two mixtures per state. The first recognition system used only the visual parameter sets. The second recognition system used only the acoustic parameter sets, and a projection-based metric [35] was used for matching to increase robustness. The final two recognition systems implemented early and late integration of the audio and visual features. The recognition systems were trained with audio at a 30-dB SNR and tested over a range between 30 and -10 dB. Results are shown in Fig. 4 as correction ratios versus the audio SNR. It can be seen that both the early and late integration strategies provide increased performance over audio-only recognition in the presence of acoustic noise. Both of these integration strategies, however, perform slightly worse than visual-only recognition in high-noise environments. This demonstrates the need for more sophisticated fusion architectures.

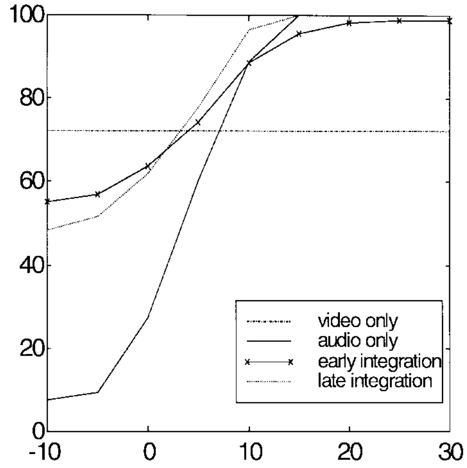


Fig. 4. Results of joint audio-visual speech recognition.

Researchers have also tried to convert mouth movements into acoustic speech directly. In [36], a system called "image-input microphone" was built to take the mouth image as input, analyze the lip features such as mouth width and height, and derive the corresponding vocal-tract transfer function. The transfer function was then used to synthesize the output acoustic speech waveform. For a test data set that contained mainly vowels, this system was shown to be useful as an input device to a computer. It was demonstrated to be particularly helpful as a speaking aid for people whose vocal cords were injured. Compared to an acoustic microphone, the image-input microphone has the advantage that it is not affected by acoustic noise and therefore is more appropriate for a noisy environment. In addition to the mouth movements that are visible to a camera, however, articulators that are invisible also determine the exact vocal-tract transfer function. Therefore, application of the image-input microphone to free speech with a large vocabulary would be limited.

Most of the lip-reading systems developed for lip-reading research can be intrusive to the users to some extent. In particular, the user has to remain relatively stationary for the visual analysis system to work well. In [37], a modular system was developed to alleviate such constraints. The visual analysis of the system is composed of an automatic face tracker followed by the lip locator. The capability of face tracking results in a speech recognizer that allows the speaker reasonable freedom of movement.

Summarizing, researchers have examined a number of different aspects of the audio-visual recognition problem. It has been shown that there are a number of types of visual parameter sets that have been useful. In particular, those parameters that quantify the shape of the mouth can be used for visual recognition. In addition, the time derivatives of these visual parameters convey useful information. The studies have shown that, in general, for small vocabularies, the visual information does not provide much benefit for clean speech, but as the acoustic speech is degraded by noise, the benefits of the visual information are evident. Also, for larger vocabulary sets, the visual signal can provide a performance gain when integrated with clean acoustic speech.

IV. Speech-Driven Face Animation

In the previous section, we have discussed research that tried to derive the acoustic domain information from visual information. Researchers have also tried to produce visual speech from auditory speech, i.e., to generate speech-driven talking heads [38]–[43]. The major applications of this technique include human–computer interfaces, computer-aided instruction, cartoon animation, video games, and multimedia telephony for the hearing impaired.

Two approaches to generating talking-head images are the flipbook method [44] and the wireframe model [two-dimensional (2-D) or three-dimensional (3-D)] approach [45]. In the flipbook approach, a number of mouth images of a person, called key frames, are captured and stored. Each image represents a particular mouth shape, e.g., a viseme. Then, according to the speech signal, the corresponding mouth images are “flipped” one by one to the display to form animation. One problem of this method is that it typically results in jerkiness during key-frame transition, especially if the number of key frames is limited. Image warping [46] can be used to create some intermediate frames to make the transition look smoother. Image warping is a process whereby one image is distorted, through the use of geometric transformations, to look like another image. It is useful for producing realistic intermediate images between video frames. To accomplish this, correspondences are made between points in two adjacent video frames. These correspondences provide a blueprint for distorting the first image into the second. Using this blueprint, the first image is warped forward by half a frame interval. The second image can be warped backward by half a frame interval by reversing this blueprint. The two intermediate images are then averaged to produce the intermediate frame that may smooth out the jerkiness between the two original frames. More than one intermediate frame can also be generated for even smoother transition.

The work in [47] uses a similar flipbook approach with some techniques to enhance the quality of the images. Instead of using a small number of mouth images representing the phonemes, they use a long video sequence of a talking person. Most likely, the sequence contains all the possible mouth shapes of the person. The sequence is analyzed to derive the phonetic transcription. When presented with input audio, matching segments from the original video are found. These segments are concatenated together with image-processing techniques to provide added realism. Time warping is also done to these segments to achieve better synchronization.

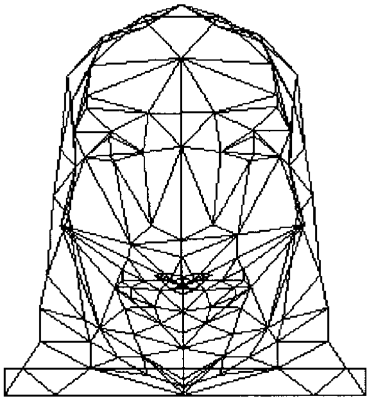


Fig. 5. The wireframe model “Candide.”

The wireframe method uses more sophisticated computer-graphics techniques to achieve better realism. A wireframe is composed of a large number of triangular patches that model the shape of a human face. One of the early facial models was developed by Parke [48]. Fig. 5 shows a more recent facial model called “Candide” developed at Linkoping University [49]. A fully 3-D model would contain vertices that correspond to points throughout the head and would allow synthesis of facial images with arbitrary orientations. However, since most animation is primarily concerned with the frontal view of the face, and not the back of the head, models are often developed only for a frontal portion of the face. The vertices in a wireframe model are manipulated to synthesize new expressions. For example, vertices in a neutral model can be repositioned to form a facial model that is smiling. To synthesize various facial expressions, the facial action coding system (FACS) [50] is often used to generate the required trajectories of the vertices for a large variety of facial expressions and movement.

A wireframe model by itself gives only a structural representation of the face. To generate natural-looking synthetic faces, a wireframe model must be combined with lighting models that specify how to map the shape and position of the wireframe into intensity when the wireframe is projected onto a 2-D image. It is possible to use simple algorithms to synthesize artificial-looking faces. For more realism, texture mapping is necessary. Texture mapping is an algorithm that maps pixel values of a 2-D image, e.g., a photo of a real person, onto patches of a wireframe. The texture from the original image helps create realistic synthetic images.

In essence, there is a duality between the two approaches we have outlined. When images are synthesized using key frames, images are known at fixed intervals, and models are used to aid in the warping process. When wireframes are used, the models are manipulated into realistic positions, then the images are texture mapped onto the models. The flipbook approach, with possible morphing, is less computationally intensive, but it requires more data (enough number of images as key frames). The wireframe approach is more computationally intensive, but it also leads to a more flexible animation system. In addition, only one image is needed for texture mapping purposes; then arbitrarily oriented images can be synthesized with the model.

So far, we have discussed only the mechanisms that can be used to create talking heads. Next, we will focus on how to produce the parameters to drive these talking heads to make them "say" certain sentences. In [39], a 3-D wireframe facial model was used to synthesize the facial expressions, particularly the lip motion. The lip parameters used therein form an eight-dimensional (8-D) vector that includes the position of the upper lip, the position of the chin, etc. These parameters are derived either by either text input or speech input. In the case of text input, a sequence of the 8-D feature vectors is manually extracted for each phoneme. The input text is then analyzed into a sequence of phonemes, and the corresponding lip-feature vectors are used to drive the facial model to produce the lip motion. In the case of speech input, a classifier that derives lip parameters from the input acoustic speech drives the lip-feature points. The input speech is first converted into linear perdition coding (LPC) cepstrum [9]. During the training of the classifier, the acoustic features, i.e., LPC cepstra, in the training data are classified into groups using vector quantization. For each acoustic group, the corresponding lip-feature vectors are grouped together and the centroid is computed. In the classifying phase, the LPC cepstrum of the input speech is classified, and the corresponding lip-feature centroid is produced by the classifier which in turns drives the facial wireframe model. In this work, facial features other than lip movements, such as eyebrows and eyelids, were controlled randomly to make the facial motion look more realistic.

In [41], a 3-D wireframe similar to that in [39] was used to synthesize talking faces to be used as human-compute interfaces. To derive the lip movements from the speech signal, HMM techniques [9] that are widely used for speech recognition are extended to find the speech-to-viseme mapping. The acoustic speech is LPC analyzed and vector quantization (VQ) classified into 64 groups. The mouth shape is parameterized using height and width and classified into 16 groups using VQ. The 16×16 transition probabilities (transition from one mouth shape to another) and the 16×64 observation probabilities for a given mouth shape's producing a particular sound are estimated from the training data to form the HMM. During the recognition phase, the speech waveform is first LPC analyzed and vector quantized. Given the HMM, the Viterbi algorithm [9] is then used to find the most probable sequence of mouth shapes. Last, the sequence of mouth shapes is fed to the facial model to create the mouth movements.

A recent work with focus on a multimedia telephone for hearing-impaired people is presented in [42] and [43]. In this work, the conversion from speech to lip movements is performed by a number of TDNN's. A major advantage of this approach is that TDNN's operate on not only the current speech frame but also its neighbors. Therefore, the estimated lip features can incorporate information from neighboring sounds. This helps model coarticulation effects from speech production.

V. Lip Synchronization

Because human speech perception is bimodal, lip synchronization is one of the most important issues in video telephony and video conferencing. What can one do if the frame rate is not adequate for lip-synchronization perception? This is a typical situation in video-conferencing equipment with bandwidth constraints. One solution is to extract information from the acoustic signal, which determines the corresponding mouth movements, and then process the speaker's mouth image accordingly to achieve lip synchronization. We will discuss this approach in more detail later in this section.

On the other hand, it is also possible to warp the acoustic signal to make it sound synchronized with the person's mouth movement [51]. This approach is very useful in non-realtime applications, such as dubbing in a studio. In movie production, the dialogue is usually recorded in a studio to replace the dialogue recorded while filming a scene because the latter has poor quality due to background noises. To ensure lip synchronization, a system known as Wordfit [51] was designed to time warp the studio dialogue to match the original recording. First, a spectrum analyzer analyzes both the studio audio and the original recording. The results are then input to a processor to find the best "time-warping path" that is required to modify the time scale of the studio audio to align the original recording. According to the time-warping path, the studio audio is edited pitch synchronously, i.e., a period of sound segment being cut out or repeated. Thus, the studio dialogue can be made in synchronization with the original lip movement.

Instead of warping the audio, one can time warp the video. In fact, one can warp the image of the speaker to make the lip movement fit the studio dialog. We now discuss this approach using the following application as the general framework. Consider a typical problem in video coding. A video codec often skips some frames to meet the bandwidth requirement, which results in a lower frame rate at the decoder. Frame skipping introduces artifacts such as jerky motion and loss of lip synchronization. To solve this, we can extract information from the speech signal and process the mouth image accordingly to achieve lip synchronization [52]. Fig. 6 shows the block diagram of this approach. In this system, the feature tracking [52], [62] module analyzes the input video frames to find the location and shape of the mouth. Meanwhile, the audio-

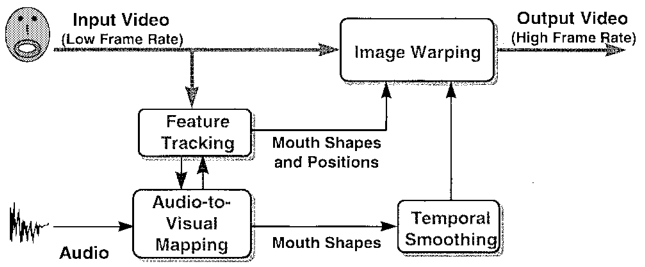


Fig. 6. Speech-assisted interpolation.

to-visual mapping module analyzes the audio signal and produces a sequence of the corresponding mouth shapes that are missing in the low-frame-rate video. Image warping is then applied to the input frames to modify the mouth shape to produce new frames that are to be inserted to the original video. Hence, lip synchronization is achieved in the high-frame-rate output video.

Here, note the interaction between image analysis and speech analysis. The results of image analysis can be used to improve the accuracy of speech recognition, as done in automatic lip reading. On the other hand, speech information can be used to improve the result of image analysis. For example, we can decide whether the mouth is open or closed by speech analysis and then apply different algorithms to locate the lip points. Mouth closures, e.g., during /p/, /b/, /m/, and silence, are important perceptual cues for lip synchronization. Therefore, the lip synchronization is good as long as speech analysis and image analysis together detect these closures correctly and image synthesis renders mouth closures precisely.

In addition to the low frame rate, another issue that causes loss of lip synchronization in video conferencing is the transmission. Typically, the transmission delay for video is longer than the audio delay. We can always delay the audio to match the video. Doing so would increase the overall delay, however, and hence is not desirable in interactive two-way communication. The speech-assisted video-processing technique can solve this problem by warping the mouth image of the speaker to be synchronized with the audio. Therefore, effectively, we can actually decrease the overall delay.

Furthermore, the speech-assisted interpolation scheme as described above can be embedded into a video codec. In a typical video codec, such as H.263, there is usually some constraint on the number of frames that can be skipped between two coded frames. This is needed in order to prevent too much jerkiness in the motion rendition and also to prevent too much loss of lip synchronization. In the situation where a good interpolation scheme, such as speech-assisted interpolation, is in place at the decoder, the encoder can skip more frames than usual. Therefore, more bits can be assigned to each frame that is coded, so the image quality can be improved, in addition to the improvement in lip synchronization. In a typical video codec, the decision on which frames to skip is based mostly on the available bit rate. For better lip synchronization, frame skipping can be controlled also by the perceptual importance of mouth shapes and the ease of speech-assisted interpolation at the decoder. For example, the encoder can avoid skipping frames that are crucial for lip synchronization, e.g., frames that contain mouth closures.

Simple modification of the above-mentioned technique can lead to many other applications, including dubbing of foreign movies, human–computer interfaces, and cartoon animation. One main issue in dubbed foreign films is the loss of lip synchronization. Also, in cartoon animation, one main task is to animate the lip synchronization of the cartoon characters. To facilitate these tasks, we can analyze the speech or the image of the dubber and then modify the dubbee's mouth accordingly. The same technique can also be combined with language translation to create an automatic audio-visual dubbing system, in which the mouth motion is made synchronous with the translated speech. In some applications where the transcript is available, e.g., closed captioning, text-based language translation can also be used.

For many years, audio and video coding have been studied independently. Recently, there has been a trend of research on joint audio-video coding [52]–[54]. One example is to exploit the correlation between the audio and video signals in a predictive coding manner. Predictive coding of video has traditionally used information from previous video frames to help construct an estimate of the current frame. The difference between the original and estimated signals can then be transmitted to allow the receiver to reconstruct the original video frame. This method has proven extremely useful for removing the temporal redundancy in video. Similarly, the prediction can be done in a cross-modal manner to explore cross-modal redundancy. The basic idea is that there is information in the acoustic signal that can be used to help predict what the video signal should look like. Since the acoustic data is also transmitted, the receiver is able to reconstruct the video with very little side information.

This process is shown in Fig. 7. In this system, an acoustic-to-visual mapping module estimates a visual parameter set, such as mouth height and width, given the acoustic data. The image-analysis module measures the actual parameter set from the video. The measured parameter set is compared with the parameter set estimated from the acoustics, and the encoder decides what information must be sent. If the acoustic data lead to a good prediction, no data have to be sent. If the prediction is slightly off, an error signal can be sent. If the prediction is completely wrong, the measured parameter set can be sent directly. The decision of what information needs to be sent can be based on rate-distortion criteria. Hence, this system provides a coding scheme that is scalable to a wide range of bit rates.

VI. Lip Tracking

A major task in the study of audio-visual interaction is the analysis of both the audio and the video signals. The acoustic analysis has been well studied for many decades. Researchers have developed a number of ways to param-

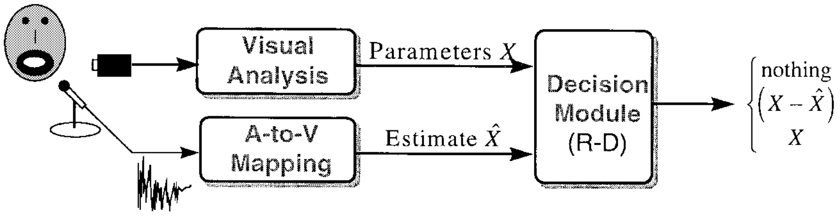


Fig. 7. Cross-modal predictive coding.

eterize speech waveforms. For example, linear-predictive coefficients and line-spectral pairs are often used for coding purposes, and filter-bank outputs or cepstral coefficients are used for recognition purposes. The question then arises as to how to analyze the visual signal, i.e., the lip movement. Unlike the speech signal, which is essentially one dimensional (1-D), the visual input is a 3-D video signal, with two spatial dimensions and one temporal dimension. A visual analysis system must convert this sequence of images into a meaningful parameter. In this section, we will review image-processing algorithms that extract visual parameters to be used by a recognition system.

Visual analysis systems can be divided into two major classes. The first classifies the mouth image into one of several categories, e.g., into visemes. The other measures parameters or dimensions from the input image, e.g., the mouth height and the mouth width. For the first class, vector quantization and neural networks are standard methods for classifying input images into several categories. For these systems, intensity images, Fourier transform coefficients, and thresholded binary images are often used as the input.

For the second class of image-analysis systems, the task is to obtain parameters or dimensions that have some physical significance. For instance, we might wish to measure the height between the lips and the width between the corners of the mouth. An even more ambitious task would be to construct a model for the lips and find the parameters of the model that provides the closest match between the model and the image.

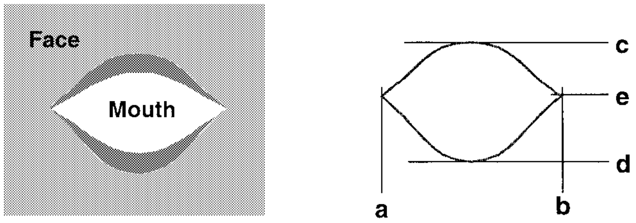
The system used in [26] for visual speech recognition took input images and applied a threshold. The resulting binary images were then analyzed, and parameters such as the area of the mouth opening, height, width, and perimeter length were extracted to provide an adequate representation of the shape of the mouth. They were successfully used in the speech-recognition experiments therein. In another system designed by Prasad *et al.* [55], the vertical and horizontal projections of both intensity and edge images were used to locate points of interest on the mouth. The distances between these points were successfully used in speech-recognition applications.

Much of the recent work in visual analysis has centered on deformable models. Both deformable templates and snakes fit into this category. The basic idea is that an energy function that relates a parameterized model to an image is formed. This energy function is minimized using any standard optimization technique, and the resulting parameter set is obtained. Snakes [56] allow one to parameterize a closed contour. The simplest snakes have energy functions that are sums of internal and external energy forces. The internal energy acts to keep the contour smooth while the external energy acts to attract the snake to key features in the image such as edges. When the total energy is minimized, the optimal snake location can be found. Researchers have also added more complexity into the system by adding surface learning [57] and flexible appearance models [58]. These techniques seek to constrain the position of snake nodes to a smaller subspace that is constructed through eigenvector decomposition. At the end of each minimization stage, the resulting snake parameters are projected onto the smaller subspace of "eigen-lips" to constrain the shape of the snake to those that are acceptable.

Deformable templates [59] provide both a parameterized model and an energy function. The model may be simple, such as modeling the outer contour of the mouth by two intersecting parabolas. More complex models can also be built. For example, the mouth-open template given in [59] consists of five parabolas. These models have different numbers of parameters. The simplest two-parabola model may be specified by four parameters, while a complex mouth model may contain more than ten parameters.

The energy function associated with deformable templates relates the template to the image. For instance, part of the energy function may accumulate the negative edge strength along the contour of the template. Therefore, this term will attract the contours of the template to the edges in the image. Energy terms relating to peak potentials, valley potentials, and intensity are also common. The energy function may also have terms that bias the template to keep parameters within reasonable limits. For instance, an energy term might be added to keep the top edge of the lower lip below the lower edge of the upper lip. These energy functions are often derived through both intuition and trial and error. Once an acceptable energy function is defined, an optimization algorithm is used to find the best-fit template for each image. Many researchers have used deformable templates to achieve excellent results in speech recognition, and several extensions have been studied [60]. Some have suggested incorporating Kalman filtering techniques into deformable templates [61]. These extensions attempt to exploit correlation between adjacent video frames and add continuity to the extracted parameter sets.

The system in [62] utilizes state-embedded deformable templates, a variant of deformable templates that exploits statistical differences in color to track the shape of the lips through successive video frames. Assume that based



(a) (b)

Fig. 8. Lip tracking. (a) The face image. (b) The template.

on pixel colors, the image can be divided into foreground (pixels within the outer contour of the lips) and background (pixels that are part of the face) regions. The shape of the foreground is modeled by a template composed of two parabolas, as shown in Fig. 8. This template is completely specified by the five parameters a–e. When the parameters change, the shape and position of the template changes. This template divides the image into foreground and background regions. Last, we assume that there are distinct probability density functions (pdf's) that govern the distribution of pixel colors in the foreground and background. Since the lips and face have different colors, the assumption is valid.

The pdf for the foreground pixels—the lips and interior of the mouth—and the pdf for the background pixels—the face—are first estimated. Based on these two pdf's, the joint probability of all pixels in the image can be calculated. The visual analysis system then uses a maximization algorithm to find the template that maximizes the joint probability. Sample results from this tracking algorithm are shown in Fig. 9. Images in the top row show the derived template overlaid on the original image. The bottom image shows pixels that are more likely to be part of the foreground.

It has also been suggested that geometric features like mouth shapes could be combined with other image-based features like the Karhunen–Loeve transform (KLT) of the mouth image. The work in [57] used the result of KLT of grayscale mouth images to assist the tracking of mouth shapes. Chiou and Hwang [63] directly combined mouth shapes with KLT for lip reading with color video.

VII. Audio-to-Visual Mapping

One key issue in bimodal speech analysis and synthesis is the establishment of the mapping between acoustic parameters and mouth-shape parameters. In other words, given the acoustic parameters, such as the cepstral coefficients or filter-bank coefficients, one needs to estimate the corresponding mouth shape and vice versa.

A number of approaches have be proposed for the task of converting acoustic speech to mouth-shape parameters. The problem can be solved from two different perspectives. The first view stresses that speech is a linguistic entity. The speech is first segmented into a sequence of phonemes, and then each phoneme is mapped to the corresponding viseme. This scheme could be implemented using a complete speech recognizer followed by a lookup table [52]. The advantage of this approach is that the acoustic speech signal is explored to the full extent, so that all the context information is utilized and coarticulations are completely incorporated. Therefore, this approach provides the most precise speech analysis. However, this approach has a certain amount of computation overhead because, to achieve audio-to-visual mapping, one does not really need to recognize the spoken words or sentences. In addition, the construction of the lookup table that maps phonemes to visemes is nontrivial.

The other view concentrates on speech's being a physical phenomenon. Since there is a physical relationship between the shape of the vocal tract and the sound that is produced, there may exist a functional relationship between the speech parameters, typically LPC cepstrum, and the visual parameters set. The conversion problem becomes one of finding the best functional approximation given sets of training data. There are many algorithms that can be modified to perform this task. These approaches include vector quantization [39], neural networks [42], and HMM's with Gaussian mixtures [64].

A. Classification-Based Conversion

This approach contains two stages, as shown in Fig. 10. In the first stage, the acoustics must be classified into one of a number of classes. The second stage maps each acoustic class into a corresponding visual output. In the first stage, vector quantization can be used to divide the acoustic training data into a number of classes. For each acoustic class, the corresponding visual code words are then averaged to produce a visual centroid. Therefore, each input acoustic vector would be classified using the optimal acoustic vector quantizer, then mapped to the corresponding visual centroid. One problem with this approach is the error that results from averaging visual feature vectors together to form the visual centroids. Another shortcoming of the classification-based method is that it does not produce a continuous mapping but rather produces a distinct number of output levels. This often leads to a staircase-like reproduction of the output.

B. Neural Networks

Neural networks can also be used to convert acoustic parameters into visual parameters. In the training phase, input and output patterns are presented to the network, and an algorithm called back propagation can be used to train the network weights. The design choice lies in selecting a suitable topology for the network. The number of hidden layers and the number of nodes per layer may be experimentally determined. Furthermore, a single network can be trained to reproduce all the visual parameters, or many networks can be trained with each network estimating a single visual parameter.

C. Direct Estimation

In this approach, the best estimate of the visual parameters is derived directly from the joint statistics of the audio and visual parameters. Consider the case where the visual parameter is 1-D, and let $f_{\mathbf{a}v}(\mathbf{a}, v)$ denote the joint distribution of the feature vector $[\mathbf{a}^T, v]^T$ composed of the

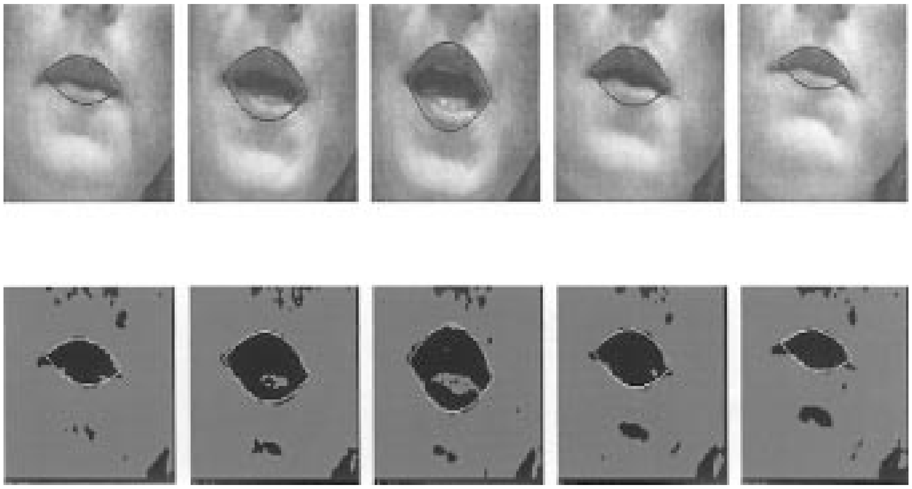


Fig. 9. Example result of lip tracking.

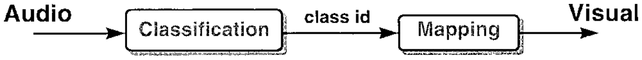


Fig. 10. Classification-based approach.

acoustic features and the visual parameter. If we know the joint pdf $f_{\mathbf{a}v}(\mathbf{a}, v)$, then the optimal estimate of v given $\mathbf{a}$ is clearly given by

$$\hat{v} = E\langle v|\mathbf{a}\rangle = \int v \frac{f_{\mathbf{a}v}(\mathbf{a}\, v)}{f_{\mathbf{a}}(\mathbf{a})}\, dv.$$

Next, suppose we can model the joint probability distribution of the audio-visual vectors as a Gaussian mixture. With this parametric model, it is possible to derive the optimal estimate of the video given the audio analytically. Suppose we use a Gaussian mixture with K Gaussian functions to represent $f_{\mathbf{a}v}(\mathbf{a}, v)$, i.e.,

$$f_{\mathbf{a}v}(\mathbf{a}, v) = \sum_{i=1}^{K} c_i \aleph(\mu_i, \mathbf{R}_i)$$

where c_i is the mixture weight and $\aleph(\mu_i, \mathbf{R}_i)$ is a Gaussian function with mean μ_i and correlation matrix $\mathbf{R}_i$. Note that each μ and $\mathbf{R}$ can be partitioned as

$$\mu = \begin{bmatrix} \mu_{\mathbf{a}} \\ \mu_v \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \mathbf{R}_{\mathbf{a}} & \mathbf{R}_{\mathbf{a}v} \\ \mathbf{R}_{\mathbf{a}v}^T & \sigma_v^2 \end{bmatrix}$$

where μ_a and μ_v are the means of the acoustic features and the visual parameter, respectively, $\mathbf{R}_{\mathbf{a}}$ is the autocorrelation matrix of the acoustic features, $\mathbf{R}_{\mathbf{a}v}$ is the covariance matrix of the acoustic features and the visual parameter, and σ_v^2 is the variance of the visual parameter. The optimal estimate of v given $\mathbf{a}$ is given by $\hat{v} = E\langle v|\mathbf{a}\rangle = \int v \frac{f_{\mathbf{a}v}(\mathbf{a}, v}{f_{\mathbf{a}}(\mathbf{a})}\, dv$ and can be written in a closed form as follows:

$$\hat{v} = \sum_{i=1}^{K} \frac{c_i \aleph(\mu_{\mathbf{i}, j}, \mathbf{R}_{\mathbf{a}, i})|_{\mathbf{a}}}{f_{\mathbf{a}}(\mathbf{a})} \mathbf{b}_i^T \begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix}$$

where each $\mathbf{b}$ equals

$$\mathbf{b} = \begin{bmatrix} 1 & \mu_{\mathbf{a}}^T \\ \mu_{\mathbf{a}} & \mathbf{R}_{\mathbf{a}} \end{bmatrix}^{-1} \begin{bmatrix} \mu_v \\ \mathbf{R}_{\mathbf{a}v} \end{bmatrix}.$$

Analyzing the above equation, one can see that each Gaussian mixture component actually yields an optimal linear estimate for v given $\mathbf{a}$

$$\mathbf{b}_i^T \begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix}.$$

These estimates are then *nonlinearly* weighted by $c_i \aleph(\mu_{\mathbf{a}, i}, \mathbf{R}_{\mathbf{a}, i})|_{\mathbf{a}} / f_{\mathbf{a}}(\mathbf{a})$ to produce the final estimate. Compared to the classification-based approach, the advantages of the mixture-based approach include the more accurate estimation and the continuity of the estimate. Consider a simple example of 1-D $\mathbf{a}$ and 1-D v, as shown in Fig. 11. Suppose the joint pdf can be modeled as one mixture of two Gaussian functions, of which the contour plots are shown in Fig. 11. The solid line shows the best estimate of v given $\mathbf{a}$ based on the Gaussian mixture model. Note that the estimate is approximately linear when it gets close to each of the two Gaussian functions and that the nonlinearity shows up when it is in between the two Gaussian functions. The dotted line shows the prediction based on a classification approach. It is clearly seen that the Gaussian mixture approach gives a smoother estimate, while the classification-based method gives a staircase response, in which a small noise in the input could result in a large variation in the output.

D. HMM Approach

HMM's have been used by the speech-recognition community for many years. Fig. 12 shows a four-state HMM together with the defining parameters. Although the majority of speech-recognition systems train HMM's on acoustic parameter sets, we will show that they can be used to model the visual parameter sets also.

Let $\mathbf{O} = [\mathbf{a}^T, v]^T$ denote the joint audio-visual parameter. The process for using an HMM for audio-to-visual conversion is as follows.

1) Training Phase: Train an N-state, left-right, audio-visual HMM on the sequence of observations $\mathbf{O}$ for each word in the vocabulary. This will give estimates for the state transition matrix, the Gaussian mixture densities associated

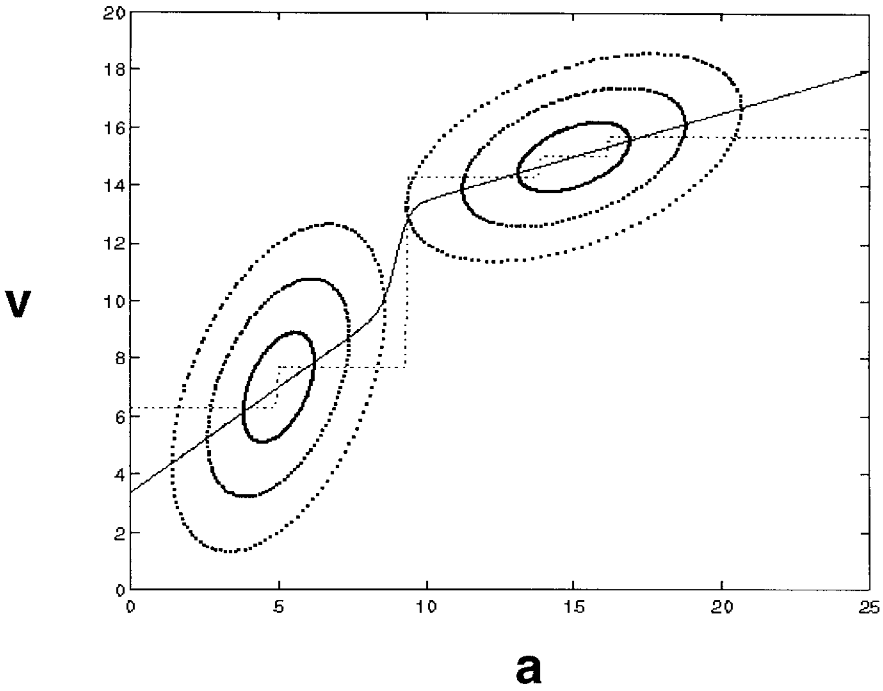


Fig. 11. A simple example for Gaussian mixture-based prediction.

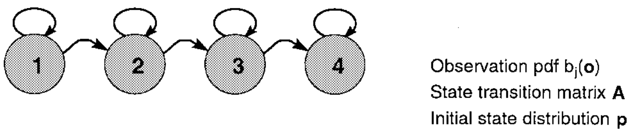


Fig. 12. An example HMM.

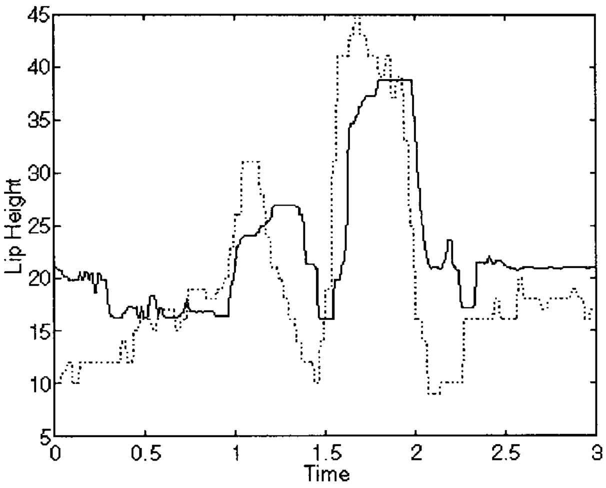


Fig. 13. The result using HMM.

with each state, $b_j(\mathbf{o})$, and the initial state distribution. Then extract an acoustic HMM from this set of parameters by integrating over the visual parameter $b_{\mathbf{a}j}(\mathbf{a}) = \int b_j(\mathbf{o})dv$. This new set of observation pdf's, along with the previously measured state transition matrix and initial state distribution, will serve as a model for the evolution of the acoustic parameters. For each state j, one can derive the optimal estimate for the visual parameter given the acoustics $E_j\langle v|\mathbf{a}\rangle$. As discussed earlier, this quantity has a closed-form solution, since a Gaussian mixture is used to model the joint distribution of the audio-visual parameters.

2) Conversion Phase: When presented with a sequence of acoustic vectors that correspond to a particular word, the audio HMM can be used to segment the sequence of acoustic parameters into the optimal state sequence using the Viterbi algorithm. Next, the optimal estimate for the visual vector can be found using the estimation function that was derived for each particular state. Alternatively, the probability for each state at each time instant can be evaluated using the forward-backward recursion [9], and the estimates from all states can be weighted together to form the final estimate.

E. Simulation Results

Here, we compare the results of the HMM-based method and the neural-network-based method. Fig. 13 shows the result of the HMM-based method. The dotted line represents the height variation of the mouth when speaking a particular phrase. The solid line represents the estimation. Fig. 14 shows the result of the same phrase using the neural-network approach. It can be seen that the HMM-based approach gives better approximation to the original waveform.

VIII. Bimodal Person Verification

Existing methods for person verification are mainly based on either face images or voice. Using each single modality, however, has certainly limitations in both security and robustness. Using still images alone can be ineffective because it is easy to store and use prerecorded images. Image-only person verification can also suffer from image-

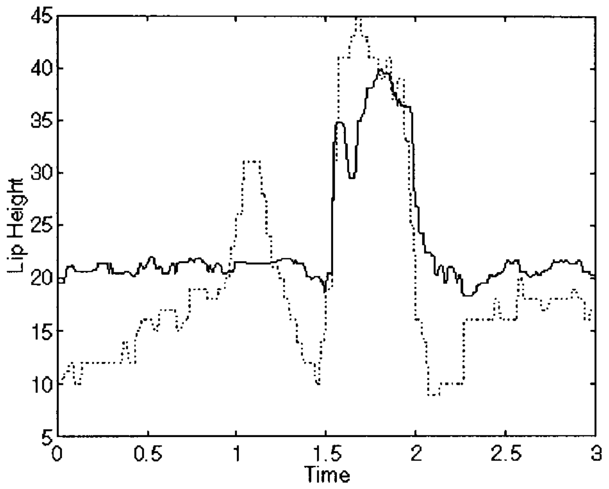


Fig. 14. The result using neural network.

coding artifacts and variation in lighting conditions. On the other hand, use of voice only for verification is not reliable either because it is possible to rearrange phonemes from a prerecorded speech of a person to synthesize different phrases. In addition, voice-only systems may fail when the acoustic environment is noisy or contains echo, such as in a typical office environment. Joint use of voice and video can solve these problems. By combining these two modalities, we can obtain more secure and more robust person-verification systems.

Recently, there have been a number of techniques [65]–[67] that use lip movement, together with acoustic speech, to identify or verify a person. Modern personal computers with multimedia capabilities, i.e., cameras and microphones, make these techniques particularly attractive. During the registration phase, the user says a chosen phrase, and the voice and lip movements of the user are recorded into a data base. During the verification phase, the user is then asked to read the displayed phrase. The user's voice and video data are then compared with those in the data base to verify the user.

Until recently, lip movement has been used mainly for speech recognition and not for speaker verification. In [68], it was demonstrated that lip movement also contained information about a person's identity. For illustration, we present a simple example here. Fig. 15 shows the time variations of the mouth height of two persons, each saying "Hello! How are you?" two times. Note that the lip movements while saying the same phrase vary a great deal from individual to individual, but they stay relatively consistent for the same person. With dynamic time warping [9], a technique commonly used in acoustic-based speaker verification, to match the features, the scores of "match" and "no match" could differ by a factor of more than 40.

The implementation given in [67] used the time variation of the mouth height and width as video features and the LPC coefficients as the audio features. The video features and the audio features were combined to form a single feature vector, with weighting that depends on the acoustic background noise. The weighting was chosen so that the weight assigned to the audio features would be small when the acoustic background noise level was high. To match the sequence of extracted features of the user to the data base, dynamic time was performed. If the distance between the captured features and the prestored features after dynamic time warping was below a prescribed threshold, a match was declared and the user was verified. The threshold was chosen so that the false acceptance rate and the false rejection rate were approximately equal. Under such conditions, it was shown that the error rate for voice-only speaker verification could be as high as 40% for noisy speech, compared to 5% for clean speech. With joint audio-visual verification, the error rate for noisy speech could be reduced to 10%. For clean speech, adding the visual information did not affect the error rate by any significant amount.

In [65], acoustic features composed of cepstral coefficients were combined with visual features, including the motion of lip contours, to achieve speaker identification. The combination, or the so-called multisensor data fusion, was done using principle component analysis or linear discriminant analysis. It was demonstrated that the joint audio-visual approach could achieve an error rate of 1.25% for speaker identification, compared to 6.25% for audio-only speaker identification and 12.5% for the image-based approach.

Both systems reported in [65] and [67] took an early integration approach, i.e., the acoustic features and the visual features were integrated before they were sent to the matching algorithm. In [66], a late-integration approach was taken, in which the visual and acoustic features were used for acoustic matching and visual matching, respectively, and then the scores of the two matching modules were combined together to form the final decision for person verification. A novel method for normalizing both modalities to a common confidence interval was also proposed. The results showed that the integrated verification system outperformed the acoustic subsystem by reducing the false-acceptance rate from 2.3 to 0.5%.

IX. Concluding Remarks

Although we live in a world where we have audio-visual media and transmission in everyday life, so far, speech and image researchers have been working independently. In this paper, we have shown that once we break down the boundary between speech research and image research, a large number of new techniques and applications could be invented [69]. All these results clearly demonstrate that the joint processing of audio and video provides additional capabilities that are not possible when audio and video are studied separately.

Audio-visual interaction can be exploited in many other ways. It certainly goes beyond the issue of lip synchronization. For example, the bit-rate allocation between audio and video remains an open issue in audio-visual communication. Recently, joint use of audio and video has been applied to multimedia content classification and segmenta-

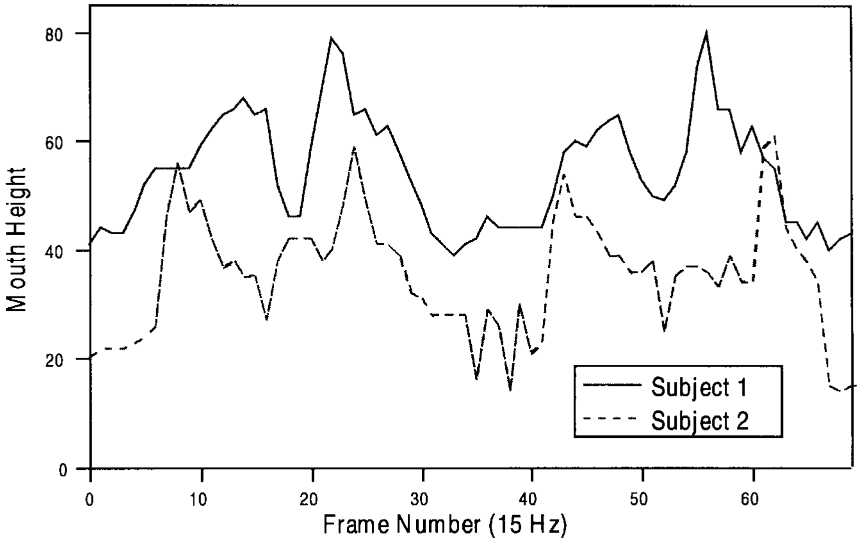


Fig. 15. Time variation of mouth height.

tion content [70]–[72]. Although preliminary, it has been shown that joint audio and video processing provides more reliable classification than using each modality. This is an area in joint audio and video research that will become very popular.

Acknowledgment

The authors wish to thank Prof. A. M. Tekalp for handling the review process of this paper. They appreciate very much the anonymous reviewers' suggestions and comments that helped enhance the paper significantly.

References

[1] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature,* pp. 746–748, Dec. 1976.

[2] R. D. Easton and M. Basala, "Perceptual dominance during lipreading," *Perception Psychophys.,* vol. 32, pp. 562–570, 1982.

[3] D. J. Dekle, C. A. Fowler, and M. G. Funnell, "Audiovisual integration in perception of real words," *Perception Psychophys.,* vol. 51, no. 4, pp. 355–362, 1992.

[4] A. Fuster-Duran, "Perception of conflicting audio-visual speech: An examination across Spanish and German," in *Speechreading by Humans and Machines,* D. Stork and M. Hennecke, Eds. Berlin, Germany: Springer-Verlag, 1996, pp. 135–143.

[5] D. Burnham and B. Dodd, "Auditory–visual speech perception as a direct process: The McGurk effect in infants and across languages," in *Speechreading by Humans and Machines,* D. Stork and M. Hennecke, Eds. Berlin, Germany: Springer-Verlag, 1996, pp. 103–114.

[6] K. Green, "The use of auditory and visual information in phonetic perception," in *Speechreading by Humans and Machines,* D. Stork and M. Hennecke, Eds. Berlin, Germany: Springer-Verlag, 1996, pp. 55–77.

[7] S. Voran and S. Wolf, "Proposed framework for subjective audiovisual testing," *ANSI Working Group T1A1.5,* vol. T1A1.5, pp. 93–151, Nov. 1993.

[8] Bellcore, "Experimental combined audio/video subjective test method," ITU-T Study Group 12, SGC/12-01, Feb. 1994.

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

[10] C. Fisher, "Confusions among visually perceived consonants," *J. Speech Hearing Res.,* vol. 11, pp. 796–804, 1968.

[11] E. Owens and B. Blazek, "Visemes observed by hearing-impaired and normal-hearing adult viewers," *J. Speech Hearing Res.,* vol. 28, pp. 381–393, Sept. 1985.

[12] B. Walden, R. Prosek, A. Montgomery, C. Scherr, and C. Jones, "Effects of training on the visual recognition of consonants," *J. Speech Hearing Res.,* vol. 20, pp. 130–145, 1977.

[13] A. J. Goldschen, "Continuous automatic speech recognition by lipreading," Ph.D. dissertation, George Washington University, Washington, DC, Sept. 1993.

[14] B. Dodd and R. Campbell, Eds., *Hearing by Eye: The Psychology of Lipreading.* London, England: Lawrence Erlbaum, 1987.

[15] K. W. Berger, *Speechreading: Principles and Methods.* Baltimore, MD: National Educational Press, 1972.

[16] Q. Summerfield, "Lipreading and audio-visual speech perception," *Phil. Trans. Royal Soc. London,* pp. 71–78, 1992.

[17] W. Sumby and I. Pollack, "Visual contributions to speech intelligibility in noise," *J. Acoust. Soc. Amer.,* vol. 26, no. 2, pp. 212–215, Mar. 1954.

[18] D. Reisberg, J. McLean, and A. Goldfield, "Easy to hear, but hard to understand: A lipreading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lipreading,* B. Dodd and R. Campbell, Eds. London, England: Lawrence Erlbaum, 1987, pp. 97–113.

[19] K. Neely, "Effect of visual factors on the intelligibility of speech," *J. Acoust. Soc. Amer.,* vol. 28, no. 6, pp. 1275–1277, Nov. 1956.

[20] K. Berger, *Speechreading: Principles and Methods.* National Educational Press, 1972.

[21] A.-P. Benguerel and M. Pichora-Fuller, "Coarticulation effects in lipreading," *J. Speech Hearing Res.,* vol. 25, pp. 600–607, 1982.

[22] H. W. Frowein, G. F. Smoorenburg, L. Pyters, and D. Schinkel, "Improved speech recognition through videotelephony: Experiments with the hard of hearing," *IEEE J. Select. Areas Commun.,* vol. 9, pp. 611–616, May 1991.

[23] J. Williams, J. Rutledge, D. Garstecki, and A. Katsaggelos, "Frame rate and viseme analysis for multimedia applications," in *Proc. IEEE Multimedia Signal Processing Conf.,* Princeton, NJ, June 1997, pp. 13–18.

[24] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lipreading,* B. Dodd and R. Campbell, Eds. London, England: Lawrence Erlbaum, 1987, pp. 3–51.

[25] D. Massaro, "Bimodal speech perception: A progress report," in *Speechreading by Humans and Machines,* D. Stork and M. Hennecke, Eds. Berlin, Germany: Springer-Verlag, 1996, pp. 79–101.

[26] E. D. Petajan, "Automatic lipreading to enhance speech recognition," in *Proc. IEEE Global Telecommunications Conf.,* Atlanta, GA, Nov. 1984, pp. 265–272.

[27] E. D. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *Proc. CHI'88,* pp. 19–25.

[28] K. Mase and A. Pentland, “Automatic lipreading by optical flow analysis,” *Syst. Comput. Jpn.*, vol. 22, no. 6, pp. 67–75, 1991.

[29] K. E. Finn and A. A. Montgomery, “Automatic optically-based recognition of speech,” *Pattern Recognit. Lett.*, vol. 8, no. 3, pp. 159–164, 1988.

[30] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, “Integration of acoustic and visual speech signals using neural networks,” *IEEE Commun. Mag.*, pp. 65–71, Nov. 1989.

[31] D. G. Stork, G. Wolff, and E. Levine, “Neural network lipreading system for improved speech recognition,” in *Proc. Int. Joint Conf. Neural Networks,* 1992, pp. 285–295.

[32] J. Luettin, N. A. Thacker, and S. W. Beet, “Speechreading using shape and intensity information,” in *Int. Conf. Spoken Language Processing,* Philadelphia, PA, Oct. 1996, pp. 58–61.

[33] P. L. Silsbee and A. C. Bovik, “Medium-vocabulary audio-visual speech recognition,” in *Proc. NATO-ASI BUBION,* 1993.

[34] G. Potamianos, E. Cosatto, H. P. Graf, and D. B. Roe, “Speaker independent audio-visual database for bimodal ASR,” in *Proc. Eur. Tutorial Workshop Audio-Visual Speech Processing,* Rhodes, Greece, Sept. 1997.

[35] B. Carlson and M. Clements, “A projection-based likelihood measure for speech recognition in noise,” *IEEE Trans. Speech Audio Processing,* vol. 2, pp. 97–102, Jan. 1994.

[36] K. Otani and T. Hasegawa, “The image input microphone—A new nonacoustic speech communication system by media conversion from oral motion images to speech,” *IEEE J. Select. Areas Commun.*, vol. 13, pp. 42–48, Jan. 1995.

[37] P. Duchnowski, M. Hunke, D. Büsching, U. Meier, and A. Waibel, “Toward movement-invariant automatic lipreading and speech recognition,” in *Proc. ICASSP,* 1995, pp. 109–122.

[38] A. Lippman, “Semantic bandwidth compression: Speechmaker,” in *Proc. Picture Coding Symp.*, June 1981.

[39] S. Morishima, K. Aizawa, and H. Harashima, “An intelligent facial image coding driven by speech and phoneme,” in *Proc. IEEE ICASSP,* Glasgow, U.K., 1989, p. 1795.

[40] W. J. Welsh, A. D. Simons, R. A. Hutchinson, and S. Searby, “A speech-driven ‘talking-head’ in real time,” in *Proc. Picture Coding Symp.*, 1990, pp. 7.6-1–7.6-2.

[41] ——, “Synthetic face generation for enhancing a user interface,” in *Proc. Image’Com Conf.,* Bordeaux, France, Nov. 1990, pp. 177–182.

[42] F. Lavagetto, “Converting speech into lip movements: A multimedia telephone for hard of hearing people,” *IEEE Trans. Rehab. Eng.*, vol. 3, pp. 1–14, Mar. 1995.

[43] ——, “Time-delay neural networks for estimating lip movements from speech analysis: A useful tool in audio-video,” *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 7, pp. 786–800, Oct. 1997.

[44] P. Griffin and H. Noot, “The FERSA project for lip-sync animation,” in *Lecture Notes Comput. Sci.*, vol. 1024, pp. 528–529, 1995.

[45] K. Waters and T. M. Levergood, “DECface: An automatic lip-synchronization algorithm for synthetic faces,” DEC Cambridge Research Lab, Tech. Rep., Sept. 1993.

[46] G. Wolberg, *Digital Image Warping*. New York: IEEE Computer Society Press, 1990.

[47] C. Bregler, M. Covell, and M. Slaney, “Video rewrite: Driving visual speech with audio,” in *Proc. ACM SIGGRAPH’97,* pp. 353–360.

[48] F. I. Parke, “Parameterized models for facial animation,” *IEEE Comput. Graph. Applicat. Mag.*, vol. 12, pp. 61–68, Nov. 1982.

[49] M. Rydfalk, “CANDIDE: A parameterized face,” Linkoping Univ., Sweden, Rep. LiTH-ISY-I-0866, Oct. 1987.

[50] P. Ekman and W. Friesen, *The Facial Action Coding System.* San Francisco, CA: Consulting Psychologists Press, 1978.

[51] P. J. Bloom, “High-quality digital audio in the entertainment industry: An overview of achievements and challenges,” *IEEE Acoust., Speech, Signal Processing Mag.*, pp. 2–25, Oct. 1985.

[52] T. Chen, H. P. Graf, and K. Wang, “Lip-synchronization using speech-assisted video processing,” *IEEE Signal Processing Lett.*, vol. 2, pp. 57–59, Apr. 1995.

[53] D. Shah and S. Marshall, “Multi-modality coding system for videophone application,” in *Proc. WIASIC’94,* Berlin, Germany, Oct. 1994.

[54] R. Rao and T. Chen, “Cross-modal predictive coding,” in *Proc. Symp. Multimedia Communication and Video Coding,* New York, Oct. 1995, pp. 301–308.

[55] K. Prasad, D. Stork, and G. Wolff, “Preprocessing video images for neural learning of lipreading,” Ricoh California Research Center, Tech. Rep. CRC-TR-9326, Sept. 1993.

[56] M. Kass, A. Witkin, and D. Terzopoulos, “Snakes: Active contour models,” in *Proc. Int. Conf. Computer Vision,* London, England, 1987, pp. 259–268.

[57] C. Bregler and Y. Konig, “‘Eigenlips’ for robust speech recognition,” in *Proc. ICASSP’94,* Adelaide, Australia, 1994, pp. 669–672.

[58] A. Lanitis, C. Taylor, and T. Cootes, “Automatic tracking, coding and reconstruction of human faces, using flexible appearance models,” *Electron. Lett.*, vol. 30, no. 19, pp. 1587–1588, Sept. 1994.

[59] A. Yuille, P. Hallinan, and D. Cohen, “Feature extraction from faces using deformable templates,” *Int. J. Comput. Vision,* vol. 8, no. 2, pp. 99–111, 1992.

[60] P. Silsbee, “Motion in deformable templates,” in *Proc. ICIP’94,* Austin, TX, 1994, pp. 323–327.

[61] C. Kervrann and F. Heitz, “Robust tracking of stochastic deformable models in long image sequences,” in *Proc. ICIP’94,* Austin, TX, 1994, pp. 88–92.

[62] R. Rao and R. Mersereau, “On merging hidden Markov models with deformable templates,” in *Proc. ICIP’95,* Washington, DC, pp. 556–559.

[63] G. I. Chiou and J.-N. Hwang, “Lipreading from color video,” *IEEE Trans. Image Processing,* vol. 6, pp. 1192–1195, Aug. 1997.

[64] T. Chen and R. Rao, “Audio-visual interaction in multimedia communication,” in *Proc. ICASSP,* Munich, Germany, Apr. 1997, vol. 1, pp. 179–182.

[65] C. C. Chibelushi, J. S. Mason, and F. Deravi, “Integration of acoustic and visual speech for speaker recognition,” in *Proc. 3rd Euro. Conf. Speech Communication and Technology,* Berlin, Germany, Sept. 1993.

[66] P. Jourlin, J. Luettin, D. Genoud, and H. Wassner, “Acoustic-labial speaker verification,” *Pattern Recognit. Lett.,* to be published.

[67] M. R. Civanlar and T. Chen, “Password-free network security through joint use of audio and video,” in *Proc. SPIE Photonic East,* Nov. 1996, pp. 120–125.

[68] J. Luettin, N. A. Thacker, and S. W. Beet, “Speaker identification by lipreading,” in *Proc. Int. Conf. Spoken Language Processing,* Philadelphia, PA, Oct. 1996, pp. 62–65.

[69] R. Chellappa, T. Chen, and A. Katsaggleos, “Audio-visual interaction in multimodal communication,” *IEEE Signal Processing Mag.*, pp. 37–38, July 1997.

[70] J. Nam and A. H. Tewfik, “Combined audio and visual streams analysis for video sequence segmentation,” in *Proc. IEEE ICASSP,* Munich, Germany, Apr. 1997, pp. 2665–2668.

[71] C. Saraceno and R. Leonardi, “Audio as a support to scene change detection and characterization of video sequences,” in *Proc. IEEE ICASSP,* Munich, Germany, Apr. 1997, pp. 2597–2600.

[72] Y. Wang, J. Huang, Z. Liu, and T. Chen, “Multimedia content classification using motion and audio information,” in *Proc. IEEE Int. Symp. Circuits and Systems,* Hong Kong, June 1997, pp. 1488–1491.

Tsuhan Chen (Member, IEEE), for a photograph and biography, see this issue, p. 753.

Ram R. Rao received the B.S. degree in electrical engineering from Rutgers—The State University, New Brunswick, NJ, in 1992 and the M.S. degree in electrical engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1993. He currently is pursuing the doctoral degree there

Since 1993, he has been a Research Assistant at Georgia Tech. His research interests lie in digital signal processing, specifically in the areas of video coding, image analysis, and speech recognition.