# AUDIO-VISUAL INTERACTION IN MULTIMEDIA COMMUNICATION

*Tsuhan Chen*[†] *Ram R. Rao*[‡]

[†]AT&T Labs - Research, Holmdel, New Jersey, USA [‡]Georgia Institute of Technology, Atlanta, Georgia, USA
Email: tsuhan@research.att.com Phone: (908) 949-2708 *

## ABSTRACT

To many people, the word "multimedia" simply means the combination of various forms of information: text, speech, music, images, graphics and video. What is often overlooked is the interaction among these forms. In this paper, we will present our recent results in exploiting the audio-visual interaction that is very significant in multimedia communication. The applications include lip synchronization, joint audio-video coding, and person verification. We will present the enabling technologies, including audio-to-visual mapping and facial image analysis, for these applications. Our results show that the joint processing of audio and video provides advantages that are not available when audio and video are studied separately.

## 1. INTRODUCTION

What is *Multimedia*? To many people, the word "Multimedia" simply means the presentation of a combination of various forms of information: text, speech, audio, music, images, graphics, and video. What is often overlooked, though, is the interaction among these different forms of information. For multimedia applications that involves person-to-person conversation, such as video telephony and video conferencing, the interaction between acoustic information and visual information is very significant. In this paper, we will show that joint audio-video processing often provides major improvement compared to the situation where audio and video are processed independently.

We will introduce our work in speech-assisted lip synchronization and joint audio-video coding. We will discuss the enabling technologies of these projects. Two major techniques are audio-to-visual mapping and image analysis for lip tracking.

We will also present a system for bimodal person verification. This system utilizes the multimedia capability of personal computers to provide a password-free verification process.

## 2. BIMODALITY OF HUMAN SPEECH

Human speech is bimodal both in production and perception. It is produced by the vibration of the vocal cord and the configuration of the vocal tract that is composed of articulatory organs including: the pharynx, the nasal cavity, the tongue, teeth, velum, and lips. Using these articulatory organs, together with the muscles that generate facial expressions, a speaker produces speech. To perceive speech, an observer listens to the acoustic speech and looks at visible articulatory organs and facial expressions. In fact, "McGurk Effect" [1] showed that human perception of acoustic speech can be affected by the visual cues of lip movements. For example, a video clip in which the speaker's mouth is saying /ga/ but the audio is dubbed with the sound /ba/, is often perceived as /da/. There also exists the "reverse McGurk Effect," i.e., the results of lipreading is affected by the dubbed audio speech [2]. Another example of audio-visual interaction is human lipreading that is widely used by the hearing-impaired for speech communication. In fact, people who are not hearing-impaired also utilize lipreading to some extent [3], especially when the auditory environment contains background noise.

## 3. LIP SYNCHRONIZATION AND CODING OF TALKING HEAD VIDEO

Since human speech perception is bimodal, lip synchronization becomes an important issue in videotelephony and video conferencing. One typical problem in videotelephony is that, due to limited bandwidth, the video coder often skips some frames, which results in a lower frame rate at the decoder. Frame skipping introduces artifacts such as jerky motion and loss of lip synchronization in talking-head video. To solve this, we can extract information from the speech signal and apply image processing to the mouth image to achieve lip synchronization [4, 5]. Fig. 1 shows the block diagram of this approach. In this system, image analysis [4, 6] is applied to the input video frames to find the location and shape of the mouth. Meanwhile, the audio signal is analyzed to produce a sequence of corresponding mouth shapes that are missing in the low frame rate video. Details of these components will be discussed in Sections 4. and 5. An image warping technique [7] is then applied to the input frames in order to modify the mouth shape to produce new frames that are to be inserted. Hence, lip synchronization is achieved in the high frame-rate output video.

*Example results:* We apply speech-assisted frame-rate conversion to the "Mom" sequence in which she is saying "...told me...". The result is shown in Fig. 2. The frame on the left (corresponding to "...old") and the frame on the right (corresponding to "...e") are existing frames in the low frame rate sequence. The middle one is obtained by interpolation. Note that a closed mouth shape is rendered for /m/, which would not be possible without speech information.

For the coding of talking head sequences, audio-coding researchers and video-coding researchers have been working independently so far. With audio-visual interaction considered, it is clear that bimodal perceptual quality tests should be examined while evaluating video or audio coding standards [8, 9]. Recently, there has been a trend of research on joint audio-video coding [4, 10, 11]. We will present an example in this section.

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 signal 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, we can explore methods that remove cross-modal redundancy. The basic premise is that there is information in the acoustic signal that can be used to help predict what the video signal should

[*] Part of the work was performed when Mr. Rao was with the University Relations Program at AT&T Bell Laboratories.

look like. If the audio indicates that a vowel is being said, one could predict that the person's mouth is open. Likewise, if the audio indicated that a /p/, /b/, or /m/ were being spoken, one could predict that the person's mouth is closed. 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. 3. In this system, an acoustic to visual predictor 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 is based on rate-distortion criteria. Therefore, this system provides a coding scheme that is scalable to a wide range of bit rates.

## 4. AUDIO TO VISUAL MAPPING

One important enabling technology for bimodal speech processing is the mapping from from speech to lip movements. This problem can be solved from two different perspectives. The first view stresses that speech is a linguistic entity. The speech is first be segmented into a string of phonemes, and then each phoneme can be mapped to the corresponding viseme. This scheme could be implemented using a speech recognizer followed by a table lookup to convert to visual parameters [5].

The other view concentrates on speech 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 [12], 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. Vector quantization [13], neural networks [14], and Gaussian mixture-based estimation [15] have been used to train this mapping.

### 4.1. Classification Based Conversion

This approach contains two stages. In the first stage, the acoustics must be classified into one of a number of groups. The second stage maps each acoustic group 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 codewords 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.

### 4.2. Neural Networks

Multilayer perceptrons [16] can also be used to convert acoustic parameters into visual parameters. In the training phase, input patterns and output patterns are presented to the network, and an algorithm called backpropogation 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.

### 4.3. Mixture Based Estimation

Our research has used a method which jointly clusters the audio-visual features. The probability distribution of the audio-visual vectors was modeled using Gaussian mixtures. With this parametric model for the joint probability distribution of audio and video, it is possible to derive the optimal estimate of the video given the audio analytically. Consider estimating a single visual parameter, $v$ given the acoustic vector $\mathbf{a}$, where a Gaussian mixture density, $f_{\mathbf{a}v}(\mathbf{a}, v)$, with $K$ mixtures governs the joint distribution of $\mathbf{a}, v$:

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

$c_i$ is the mixture weight, and $\mathcal{N}(\mu_i, \mathbf{R}_i)$ is a Gaussian density with mean, $\mu_i$, and correlation matrix, $\mathbf{R}_i$. The optimal estimate of $v$ given $\mathbf{a}$ is then given by:

$$\hat{v} = E[v|\mathbf{a}] = \int v \, \frac{f_{\mathbf{a}v}(\mathbf{a}, v)}{f_{\mathbf{a}}(\mathbf{a})} \, dv \tag{2}$$

It can be shown that the optimal estimate above can be written in a closed form [15]. Compared to the classification based approach, the advantages of the mixture based approach include the more accurate estimation and the continuity of the estimate.

### 4.4. The HMM Approach

Hidden Markov models have been used by the speech recognition community for many years. Although the majority of speech recognition systems train HMM's on acoustic parameter sets, visual speech recognition results have shown that they can be used to model the visual parameter sets also.

Consider estimating a single visual parameter, $v$, from the multidimensional acoustic parameter $\mathbf{a}$. The audiovisual parameter $\mathbf{O} = [\mathbf{a}^T \; v]^T$. The process for using HMMs for audio-to-visual conversion would proceed as follows:

*Training Phase:* Train an $N$ state, left-right, audio-visual hidden Markov model on the sequence of observations, $\mathbf{O}$ for each word in the vocabulary. This will give estimates for the state transition matrix, $A$, the Gaussian mixture densities associated 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: $b_{\mathbf{a}j}(\mathbf{a}) = \int_v b_j(\mathbf{O}) d\mathbf{O}$. This new set of observation probability density functions 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, $j$, derive the optimal estimate for the visual parameter given the acoustics, $E_j[v|\mathbf{a}]$. Since Gaussian mixtures are modeling the joint distribution of the audio-visual parameters, this quantity has a closed form solution.

*Conversion Phase:* When presented with a sequence of acoustic vectors which correspond to a particular word, estimation can occur in one of two ways:

1. The 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 which was derived for each particular state.

2. Alternatively, the HMM can be used to find the probability of being in state $j$ at time $t$, $\gamma_t(j)$, by using the forwards-backwards algorithm. An estimate of the visual signal can then be formed as $\sum_j \gamma_t(j) E_j[v|\mathbf{a}]$.

## 5. FACIAL IMAGE ANALYSIS

Another enabling technology for bimodel speech processing is image analysis for tracking of lip movements. Existing image analysis systems can be divided into two major classes: those that classify the input image into one of several categories, and those that measure dimensions of facial features.

Vector quantization and neural networks are standard methods for classifying input images into several classes. As input, one could use intensity images, Fourier transform coefficients, binary images obtained by thresholding, and many of the other processed versions of the images which are noted above. For example, a system used in [17] for visual speech recognition experiments took input images and applied a set threshold. The binary images were then analyzed to extracted parameters such as the area of the mouth opening, the height, and the width.

Much of the recent work in facial image analysis has centered around deformable models. Both snakes [18] and deformable templates [19] fit into this category. The basic idea is that an energy function which relates a parameterized model to an image is formed. This energy function is minimized using any standard optimization technique to obtain the optimal parameter set. Snakes allow one to parameterize a closed contour, and deformable templates provide a more general parameterized model. The energy function associated with deformable templates relates the template to the image. Energy terms relating to peak potentials, valley potentials, and intensity are also common. These energy functions are often derived through both intuition and trial and error.

Now we will examine one analysis system in detail. This system, using state-embedded deformable templates, is a variant of deformable templates which exploits statistical differences in color to track the shape of the lips through successive video frames [6]. A few assumptions are made concerning the structure of the input images. First, it is assumed that the image can be divided into foreground (pixels within the outer contour of the lips) and background (pixels which are part of the face) regions. Next, it is assumed that the shape of the foreground can be modeled by two parabolas as shown in Fig. 4. The template is completely specified by the five dimensional parameter $\lambda = [x1, x2, y1, y2, y3]$. Finally, we assume that there are distinct probability density functions (pdf) which govern the distribution of pixel colors in the foreground and background.

If we have estimates for the foreground (pixels within the lips) pdf and background (pixels of the face) pdf, we can evaluate the joint probability of all pixels in the image. This joint probability is given by:

$$P[I|\lambda] = \prod_{(x,y)\epsilon fg} b_{fg}(I(x,y)) \prod_{(x,y)\epsilon bg} b_{bg}(I(x,y)) \qquad (3)$$

where $P[I|\lambda]$ is the joint probability, $I(x,y)$ is the three dimensional pixel value at location $(x,y)$, and $b_{fg}$ and $b_{bg}$ are the foreground and background pdf's, respectively. Notice the dependence on $\lambda$: if $\lambda$ is changed, different pixels become part of the foreground and background, thus changing the joint probability value. Our visual analysis system uses a maximization algorithm to find the parameter, $\lambda$ which maximizes the joint probability of the pixels in the image.

We model the foreground and background pdf's as Gaussian mixtures with two Gaussian's per mixture. Two mixtures are needed because one would expect different statistical characteristics for pixels in the lips, and for those in the mouth opening. The analysis system tracks the shape and position of the mouth through successive video frames. For each new frame, our system must compute the following quantity for every pixel:

$$P(x,y) = \log(b_{bg}(I(x,y))) - \log(b_{fg}(I(x,y))) \qquad (4)$$

Since evaluating this quantity is an intensive task, a lookup table is used to store all of the possible values of the log likelihood quantity. Once we have the log likelihood image, we can find the template parameter $\lambda$ which maximizes the quantity in (3). This is equivalent to minimizing

$$f(\lambda) = \sum_{(x,y)\epsilon fg} P(x,y) \qquad (5)$$

which can be achieved by a log search algorithm.

## 6. PERSON VERIFICATION

Recently, there have been a number of techniques [20, 21, 22] that use lip movement and speech to identify or verify a person. Modern personal computers with multimedia capabilities, i.e., cameras and microphones, provide the basis for these techniques. In this section, we outline our implementation of such a system.

Existing methods for user verification are based on finger prints, irises, face images, or voice. Using still images along, however, can be ineffective because it is easy to store and use pre-recorded images. Use of voice only is not reliable either because it is possible to rearrange phonemes from a pre-recorded speech of a person to generate new phrases. Another problem with voice-only systems is that they fail under acoustic noise or echo. Joint use of voice and video can solve these problems.

During the registration phase, the voice and the lip movements of a user are recorded while the user is saying a certain phrase. During the verification phase, the user is then asked to read the displayed phrase. To verify the user, the chosen features of the user's voice and video data are compared with that of the stored data. We use the time variation of the mouth height and width, i.e., $|y3 - y1|$ and $|x2 - x1|$ in Fig. 4, as video features. The LPC cepstrum coefficients are selected as the audio features. We combine the video features and the audio features to form a single feature vector. To match the sequence of extracted features of the user to the database, we perform dynamic time warping [12] (Type II is chosen for simplicity) which is commonly used in speech recognition. If the distance between two waveforms after dynamic time warping is below a prescribed threshold, we declare there is a match and the user is verified.

*Example results:* Fig. 5 shows the time variations of the mouth height of each subject while saying "Hello! How are you?" twice. It can be seen that the lip movements while saying the same phrase vary a great deal from individual to individual; but they stay consistently the same for the same subject. The result of dynamic time warping shows that the the scores of "match" and "no match" differ by a factor of more than 30.

## 7. CONCLUDING REMARK

Although we live in a world where we have audio-visual media and audio-visual transmission in everyday life, so far speech researchers 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, we can invent a large number of new techniques and applications.

## REFERENCES

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

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

[3] Summerfield, Q., "Lipreading and audio-visual speech perception," *Phil.Trans. R.Soc.Lond.*, pp.71-78, 1992.

[4] Chen, T., Graf, H. P., and Wang, K., "Speech-assisted video processing: Interpolation and low-bitrate coding," 28th Annual Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, CA, pp. 975–979, October 1994.

[5] Chen, T., Graf, H. P., and Wang, K., "Lip-synchronization using speech-assisted video processing," *IEEE Signal Processing Letters*, vol. 2, no. 4, pp. 57–59, April 1995.

[6] Rao, R. and Mersereau, R., "On merging hidden Markov models with deformable templates," ICIP 95, D.C., 1995.

[7] Wolberg, G., *Digital Image Warping*, IEEE Computer Society Press, 1990.

[8] Voran, S., and Wolf, S., "Proposed framework for subjective audiovisual testing," ANSI Working Group T1A1.5, T1A1.5/93-151, November 1993.

[9] Bellcore, "Experimental combined audio/video subjective test method," ITU-T SGC/12-01, February 1994.

[10] Shah, D., and Marshall, S., "Multi-modality coding system for videophone application," WIASIC '94, Berlin, Germany, October 1994.

[11] Rao, R., and Chen, T., "Cross-modal predictive coding," Symp. on Multimedia Comm. and Video Coding, New York, October 1995.

[12] Rabiner, L. R., and Juang, B. H., *Fundamentals of speech recognition*, Prentice Hall, Englewood Cliffs, NJ, 1993.

[13] Morishima, S., Aizawa, K., and Harashima, H., "An intelligent facial image coding driven by speech and phoneme," ICASSP, p. 1795, Glasgow, UK, 1989.

[14] Lavagetto, F., "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Trans. on bilitation Engineering*, pp. 1–14, March 1995.

[15] Rao, R. and Chen, T., "Exploiting audio-visual correlation in coding of talking head sequences," PCS, Melbourne, Australia, March 1996.

[16] Lippmann, R., "An introduction to computing with neural nets," IEEE ASSP Magazine, pp. 4–22, April 1987.

[17] Petajan, E. D., "Automatic lipreading to enhance speech recognition," IEEE Global Telecommunications Conf., pp. 265–272, Atlanta, GA, November 1984.

[18] Kass, M., Witkin, A., and Terzopoulos, D., "Snakes: active contour models," Proc. Int'l Conf. on Computer Vision, pp. 259–68, London, 1987.

[19] Yuille, A., Hallinan, P., and Cohen, D., "Feature extraction from faces using deformable templates," Int'l Journal of Computer Vision, pp. 99–111, 1992.

[20] C. C. Chibelushi, J. S. Mason, and F. Deravi, "Integration of acoustic and visual speech for speaker recognition," 3rd Euro. Conf. on Speech Comm. and Tech., Berlin, Sept. 1993.

[21] J. Luettin, N. A. Thacker, S. W. Beet, "Speaker Identification by lipreading," ICSLP, October 1996.

[22] Civanlar, M. R., and Chen, T., "Password-free network security through joint use of audio and video," SPIE Photonic East, November 1996.

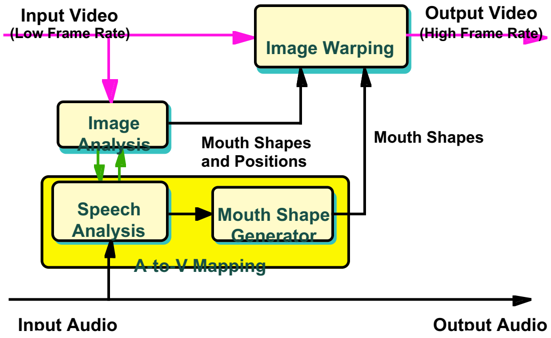

**Figure 1. Lip synchronization.**

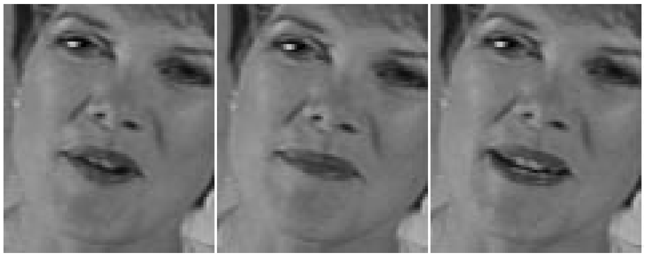

**Figure 2. Result of lip synchronization.**

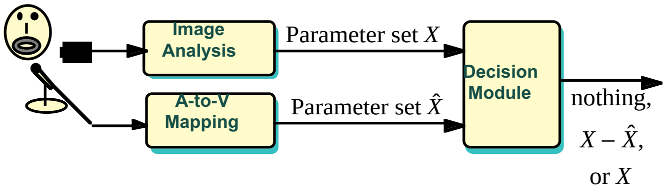

**Figure 3. Cross-modal predictive coding.**

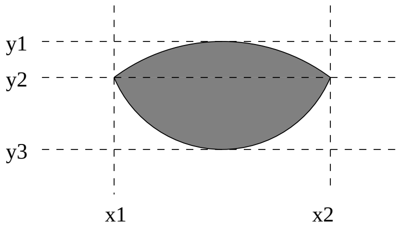

**Figure 4. Template.** $\lambda = [x1, x2, y1, y2, y3]$.

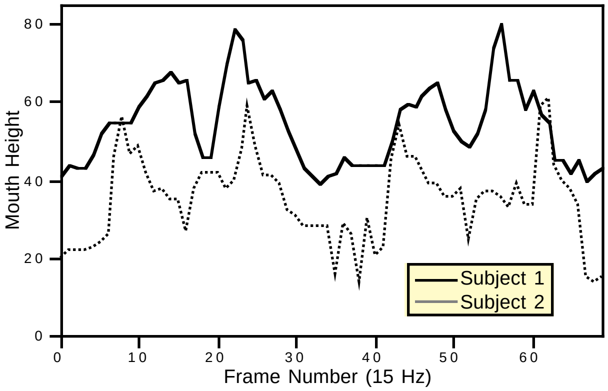

**Figure 5. Time variations of the mouth height.**