SPEECH-ASSISTED VIDEO PROCESSING: INTERPOLATION AND LOW-BITRATE CODING

Tsuhan Chen, Hans Peter Graf, Kuansan Wang
AT&T Bell Laboratories, 4C528, Holmdel, New Jersey 07733
Phone: (908) 949-2708
Email: tsuhan@research.att.com

Abstract

We utilize speech information to improve the quality of audio/visual communications, such as videotelephony, videoconferencing, and multimedia. In particular, marriage of speech processing and image processing can solve problems related to lip synchronization. Two main techniques proposed in this paper are: speech-assisted interpolation and speech-assisted coding of talking head video. Audio/video sequences are presented to demonstrate our techniques.

1: Introduction

The processing and coding of talking head video can be facilitated by taking into account the associated speech information. Previous work includes speech-driven talking heads [Lippman, 1981], [Welsh et al., 1990], and integrated coding of speech and mouth images [Morishima et al., 1989], [Shah and Marshall, 1994]. In this paper, we exploit audio/visual interaction to improve lip synchronization (the synchronization between the lip movements and the acoustic speech) in talking head video. Two main techniques we propose are speech-assisted interpolation and speech-assisted video coding for low-bitrate applications.

In videotelephony, teleconferencing, and multimedia applications, due to limited bandwidth or storage space, the video coder can not encode all incoming frames. Instead, it drops some frames by subsampling the video at a fraction of the normal rate, and encodes only the low frame rate signal (which can be as low as 1–2 frames per second for some applications). This is called frame skipping. This results in jerky motion and loss of lip synchronization in talking head video. Therefore, interpolation at the decoder is required to increase the frame rate. Techniques such as linear interpolation and motion-adaptive interpolation [Netravali and Robbins, 1981], [Bergman, 1981], [Furukawa et al., 1984] were proposed. These techniques smooth out motion jerkiness, but are not able to reproduce the mouth movements. A typical speaker can pronounce more than ten sounds per second, so the positions of lips, jaw, teeth, and tongue change at very high rates. Therefore, at a frame rate as low as 1–2 frames per second, much information about mouth movements is lost. In this paper, we propose to extract information from the speech signal to render mouth movements to reproduce lip synchronization. We call this speech-assisted interpolation. Such a post-processing approach has an advantage that it is compatible with existing coding standards and it requires no side information from the encoder to the decoder. Therefore, we can combine speech-assisted interpolation with standard video coding techniques to improve both the image quality and lip synchronization. We call this speech-assisted video coding.

2: Speech-Assisted Interpolation

Fig. 1 shows an implementation of speech-assisted interpolation:

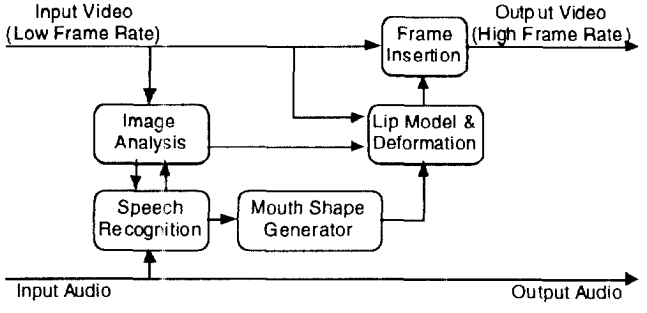

Fig. 1. Speech-assisted interpolation

The audio signal is analyzed by a speech recognizer to produce a sequence of phonemes. Continuous-speech recognition is not necessary. Instead, phoneme- or homophene[1]- level recognition is sufficient. (For non real-time applications, we can still choose to use word-level or continuous-speech recognition to improve the accuracy.). Image analysis [Graf et al., 1994] (see Section

The authors thank Dr. Barry Haskell for his inspiration, Dr. R. Civanlar for interface software, R. Schmidt for the low bitrate coding algorithm, Dr. P.-Y. Chung and Prof. Y. Wang for technical assistance, and Drs. H. Alshawi, H. Chen, W. Chou, A. Kaplan, S. Keshav, E. Petajan, A. Puri, and C.-L. Shih for fruitful discussion.

[1] Homophenes are a set of phonemes that are pronounced by the same mouth shape. In other words, a set of homophenes are visually similar.

3 for details) is applied to the input video frames to locate the mouth, i.e., the coordinates of the six white vertices in a mouth model as shown in Fig. 2. These are corners of the lips plus the outer and inner lip edges in the center of the mouth. The eight black vertices are obtained by curve fitting the white ones.

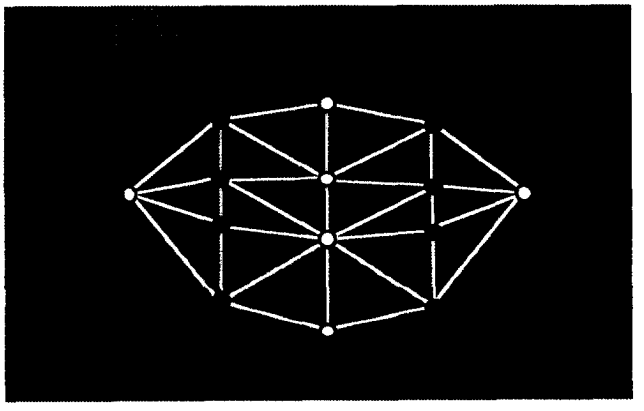

Fig. 2. The lip model

Note the interaction between image analysis and speech analysis. The results of image analysis are used to improve the accuracy of speech recognition, as done in lip-reading [Brooke and Summerfield, 1983]. On the other hand, speech information is used to improve image analysis also. For example, the rules we use to locate the six lip points depend on the phoneme corresponding to the current input frame.

The mouth shape generator is a look-up table that takes the phoneme sequence from the speech recognizer and generates corresponding mouth positions (i.e., the six lip points) that are missing in the low frame rate video due to frame skipping. In our system, a table of 12 mouth shapes performs well to cover all the mouth movements. Controlled by the output of mouth shape generator, image deformation technique [Wolberg, 1990] is applied to input frames to modify the mouth shape to produce new frames that are to be inserted. Hence, lip synchronization is reproduced.

Remarks:

1. To avoid abrupt mouth movement between phonemes, adaptive temporal smoothing is applied to these coordinates produced by the mouth shape table. The smoothing operation is adaptive so as to maintain mouth closures, e.g., explosive sounds like /p/, /b/, /m/, etc., because precise mouth closures are perceptually important for lip synchronization.
2. In the current system, we use a speaker-independent table in the position generator. To improve the performance further, we can produce speaker-dependent tables as follows. Let the system start with a default speaker-independent table at the beginning. When the speaker talks, the system keeps improving the table by extracting mouth shape parameters from existing input frames and their corresponding phonemes.
3. In our current implementation, it is assumed that the transcription is available at the phone level. As such, the Hidden Markov Model (HMM) based force alignment technique [Rabiner and Juang, 1993] is employed here for automatic speech segmentation. Each phone is associated with a three state, single Gaussian mixture HMM, and the whole utterance is modeled as a concatenation of the phone HMM's. The Viterbi algorithm is then applied iteratively to estimate the parameters for each HMM based on the 13-order LPC cepstrum coefficients computed at every 10 msec from the speech signal. The maximum likelihood HMM boundaries can be estimated in the process, and these boundaries are used as a reference of phonetic segments. When large amount of data recorded from your audio channel is available, we can actually train a speech recognizer to segment speech without knowing the transcription. The work is in progress. On the other hand, for applications where processing delay is an issue, we should choose a simple recognizer, such as an LPC or filter-bank analyzer followed by a classifier. Also note that the speech recognizer gives 100 phonemes per second, which is often higher the video frame rate. When more than one phonemes correspond to one video frame, we pick the one which is more perceptually important. For example, explosive sounds are more visually important for lip synchronization.
4. Our approach is different from the model-based coding in that, we do not match the whole face with a complicated wire-frame model, which is a very difficult task. Instead, we model only the lips, and the model is fitted to the image simply by locating the six lip points.

3: Image Analysis: Locating Feature Positions in Face

We search the whole image for the presence of a face by looking for the eyes and the mouth. Then the area around the mouth is analyzed more carefully, to measure the exact positions of the lips. For determining where a mouth is located, it is necessary to know also the positions of other features of the face, such as the eyes. Without this information it is very difficult to find reliably the mouth in a complex image. Just the mouth by itself is easily missed or other elements in the image can be mistaken for a mouth.

3.1: Finding face in image

Before the search for feature positions starts, the image is filtered, to reduce noise and to diminish the influence of variations in illumination. A bandpass filter eliminates the low spatial frequencies as well as the very high ones.

An example of an image processed in this way is shown in Fig. 3.

Fig. 3. The bandpass filtered image.

Then a morphological operation [Serra, 1982], [Pitas and Venetsanopoulos, 1990] is used to pick out areas of interest. As structuring element we use a rectangle or an ellipse with a horizontal axis that is three to four times as large as the vertical one. Typically, in a face, the areas around the eyes and the mouth show strong variations in intensity, while areas of the cheeks, the forehead and the chin are more even. Therefore, areas with strong spatial frequencies in a certain frequency range are indicative for the presence of eyes or a mouth. This assumption is very general because we do not require a specific variation of the intensity. Just the fact that the intensity varies in the area of an eye or a mouth is exploited.[2] This search strategy works well, regardless whether the eyes or mouth are closed or open.

All the areas marked as candidates for an eye or a mouth are analyzed, to determine their sizes and shapes, applying connected component analysis. Potential eye-pairs are then identified by searching through the list of eye candidates. Each pair is scored, based on position, size, and orientation. This selection of eye pairs is further refined by combining eye pairs with candidates for mouth areas. Fig. 4 shows an image where all the eye and mouth locations found by the algorithm are marked. The circles mark where eyes might be present, and the lines mark possible mouth positions.

[2] Color information, which helps to differentiate the lips and the face, can also be used.

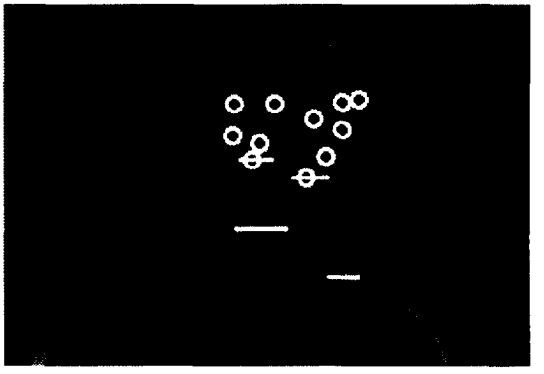

Fig. 4. Candidates for eyes and mouth.

In Fig. 4, there are a total of twelve eye-mouth combinations that have to be analyzed further. Each of these combinations is scored with a classifier that compares size-ratios with those of models and takes orientation and position into account. Constraints can be entered to limit the areas where candidates are being considered. In this way, information from previous frames can be taken into account. In Fig. 5 the eye-mouth combination with the highest score is drawn.

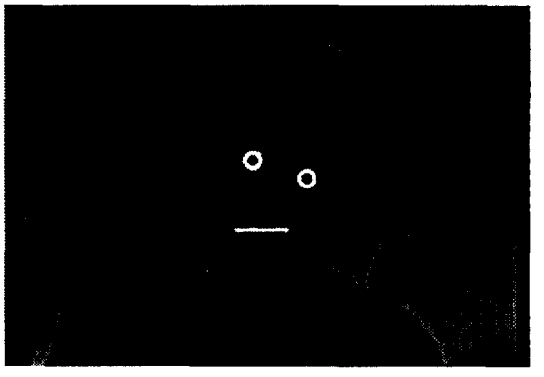

Fig. 5. The eye-mouth locations with the highest score.

3.2: Finding outline of lips

After the approximate position of the mouth has been found, a finer analysis is done, to determine precisely the outlines of the lips. A morphological operation provides an estimate of the inner area of the mouth. This information is used as a starting point for a search for the lip edges. Analyzing several vertical cross sections through the lips provides an idea what types of intensity variations are present across the lips. The measured values are compared with a list of reference patterns, to guide the strategy for finding the edges.

Sometimes a vertical cross section through the mouth provides a good measure for the outer edges of the lips. Sometimes thresholding the image leads to a better result. Which strategy is preferable depends on the illumination and the contrast between the lips and the surrounding

skin. Fig. 6 shows the result of the algorithm for finding the lip outline.

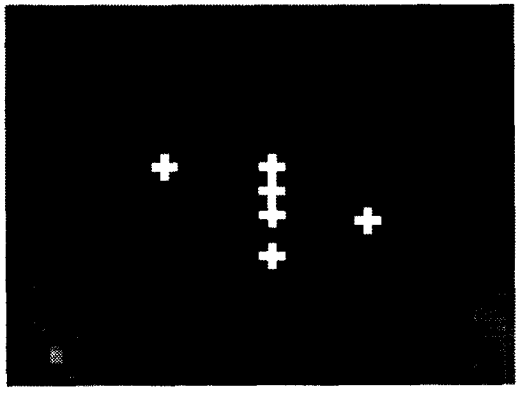

Fig. 6. The six lip points found by the algorithm.

Remark: In our current system, image analysis is done independently in each frame. Using temporal correlation (so-called tracking) will improve the temporal consistency of the feature points.

4. Example Results

We apply speech-assisted interpolation to a talking-head sequence in which the speaker is saying "...demo..." The result is shown below. The frame on the left (corresponding to the sound /e/) and the frame on the right (corresponding to the sound /o/) are existing frames in the low frame rate sequence. The middle one (corresponding to the sound /m/) is obtained by interpolation. Note that a closed mouth shape is rendered for the sound /m/, which would not be possible without speech information.

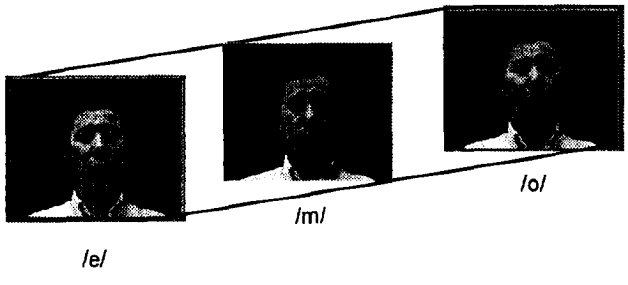

Fig. 7. Result of speech-assisted interpolation

5: Speech-Assisted Video Coding

We combine speech-assisted interpolation with existing video coding algorithms (e.g., px64 [Liou, 1991]) to improve the image quality. These coders typically drop a number of frames without coding them (frame skipping) to satisfy the bit rate constraint, at the expense of loss of lip synchronization. To minimize motion jerkiness and maintain lip synchronization, they normally put a constraint on the number of frames that can be dropped. With speech-assisted interpolation at the decoder, such a constraint can be alleviated, so the encoder is allowed to drop more frames, and encode the remaining frames with better quality. Hence, we improve both the image quality and lip synchronization. We test this technique with a px64-based codec and produce two coded sequences at 64 kbit/s, one coded with the standard frame skipping, and the other coded with doubled frame skipping with speech-assisted interpolation at the decoder end. The latter shows better image quality and lip synchronization.

6: Extensions and Applications

Speech-assisted video processing can be exploited in a number of other ways:

1. **Speech-assisted motion-adaptive interpolation**: We can combine our technique with motion-adaptive interpolation. That is, we use speech information to improved the lip synchronization and use motion-adaptive interpolation to smooth the head motion, etc.

2. **Speech-assisted motion compensation**: For motion-compensated interframe coding techniques, e.g., [Netravali and Robbins, 1979], the prediction error is typically very large around the mouth area (and blinking eyes, too), because the mouth is the most active area in a taking head video sequence. With speech information, we can obtain better prediction of the mouth and reduce the prediction error.

3. **Talking agents**: The same lip model we use for interpolation can be used to generate human-like talking agents for human-machine interface, computer-aided instruction, video games, and animation. For example, we can take a still picture of a real person, fit the lip model on the mouth, and deform the mouth to produce lip-synchronized talking head. Moreover, if we create a table for gestures used in sign language, and a model for arms and hands, we can generate video of sign language.

4. **Dubbing for foreign films and language translation**: One main issue in dubbed foreign films is the loss of lip synchronization. To solve this problem, we can analyze the speech of the dubber, and then modify the dubbee's mouth accordingly. The same technique can also be used in language translation in audio/visual media, in which the mouth image can be made synchronous with the translated speech.

5. **Perception-based frame-skipping**: A traditional video coder skips frames based on the rate controller. For lip synchronization consideration, frame-skipping should be controlled also by the perceptual

importance of mouth shapes and the ease of speech-assisted interpolation at the decoder end.

References

Bergman, H. C., "Motion-Adaptive Interpolation of eliminated TV-fields," Picture Coding Symp., 12.2, June 1981.

Brooke, N. M., and Summerfield, Q., "Analysis, synthesis, and perception of visible articulatory movements," Journal of Phonetics, vol. 11, pp. 63–76, 1983

Furukawa, A, Koga, T., and Iinuma, K, "Motion-Adaptive interpolation for videoconference pictures," IEEE Intl. Conf. on Communications, vol. 2., pp. 707–710, Amsterdam, Netherlands, May 1984.

Graf, H. P., Chen, T., Brown, K, and Jackel, L. D., "Animated images of speakers for image compression," Proc. of Lifelike Computer Characters, p. 52, Snowbird, Utah, October 1994.

Liou, M., "Overview of the px64 kbit/s video coding standard," *Communications of the ACM*, April 1991.

Lippman, A., "Semantic bandwidth compression: speechmaker," PCS, June 1981.

Morishima, S, Aizawa, K., and Harashima, H., "An intelligent facial image coding driven by speech and phoneme," ICASSP '89, Glasgow, UK, 1989.

Netravali, A. N., and Robbins, J. D., "Motion-Compensated television coding: Part I," *Bell System Technical Journal*, vol. 58, no. 3, March 1979.

Netravali, A. N., and Robbins, J. D., "Motion-Adaptive interpolation of television frames," Picture Coding Symposium, 12.1, June 1981.

Pitas, I., and Venetsanopoulos, A. N., Nonlinear Digital Filters, Kluwer Academic Publishers, Boston, 1990.

Rabiner, L. R., and Juang, B. H., *Fundamentals of speech recognition*, Printice Hall, Eaglewood Cliffs, NJ, 1993.

Serra, J., Image Analysis and Mathematical Morphology, Academic Press, London, 1982.

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

Welsh, W. J., Simons, A. D., Hutchinson, R. A., and Searby, S., "Synthetic face generation for enhancing a user interface," Proc. Image'Com Conf., pp. 177–182, Bordeaux, November 1990.

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