AUDIO DRIVEN FACIAL ANIMATION FOR AUDIO-VISUAL REALITY

T. A. Faruquie, A. Kapoor, R. Kate*, N. Rajput, L V. Subramaniam*

IBM India Research Lab
Block I, Indian Institute of Technology
New Delhi 110016, India

ABSTRACT

In this paper, we demonstrate a morphing based automated audio driven facial animation system. Based on an incoming audio stream, a face image is animated with full lip synchronization and expression. An animation sequence using optical flow between visemes is constructed, given an incoming audio stream and still pictures of a face speaking different visemes. Rules are formulated based on coarticulation and the duration of a viseme to control the continuity in terms of shape and extent of lip opening. In addition to this new viseme-expression combinations are synthesized to be able to generate animations with new facial expressions. Finally various applications of this system are discussed in the context of creating audio-visual reality.

1. INTRODUCTION

Humans communicate verbally using words and sentences. Humans also communicate non-verbally using expressions, gestures and prosody. The design and implementation of computer systems that cover the whole range of human-human like interaction by using faces and voices is one of the challenging objectives of Human Computer Interaction research. In this work we look at the conversion of speech into visual information to create audio-visual reality. Given an incoming audio stream and pictures of a face representing different visemes, which are different, distinguishable lip shapes [7, pp. 394-395], an animation sequence is constructed. We have taken 12 face images corresponding to the 12 visemes. These viseme images are aligned and optical flows for transition in-between these visemes are computed (12x11 total) and stored. At run time, for an incoming audio stream, using a phoneme to viseme mapping, the corresponding video frame is identified and transition frames between visemes are generated using the optical flows.

Morphing based animation has been considered in the past [5]. In this paper we seek to extend this to include animation with expression. For a richer scope of animation it is necessary to be able to animate the face with appropriate expressions. In [3] it is assumed that there exists a video database of the head to be synthesized, wherein, the subject is present in the expression to be synthesized at least once. In our system, given visemes with two or more different facial expressions, a method is presented that can generate the remaining visemes with these facial expressions. The Seven basic expressions considered are neutral, surprise, fear, disgust, anger, happiness and sadness. In section 2 we present the audio-driven facial animation system model. In Section 3 the method of generating the animation is discussed. In Section 4 we discuss some applications and present an evaluation of the system in the context of creating audio-visual reality. Finally conclusions are presented in Section 5.

* on summer training from Indian Inst. of Tech., Delhi.

2. SYSTEM MODEL

The audio-driven facial animation system consists of the extraction module, the synthesis module and the background processing module.

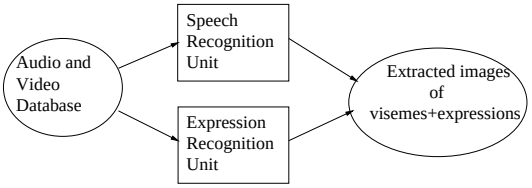

Figure 1: Extraction Module

Figure 1 shows the extraction module. For an incoming stream of synchronized audio+video we first recognize the phoneme and then map this phoneme to its corresponding viseme and take the corresponding video frame to represent this viseme. The expression recognition unit can be either audio based [2][9] or video based [4]. A short sentence like "The sharp quick brown fox jumped over the lazy dog." captures all the 12 visemes.

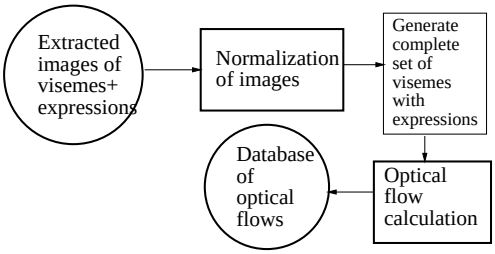

Figure 2: Background Processing Module

In the background processing module, shown in Figure 2, the extracted images are corrected for small pose differences. Then it may be possible that all visemes in all expressions may not have been extracted. This module generates the complete set of viseme+expression combinations (e_n, v_m), where, $n = 1, \ldots, 7$, and, $m = 1, 2, \ldots, 12$. Finally optical flows between different visemes within an expression and between the expressions are computed and stored.

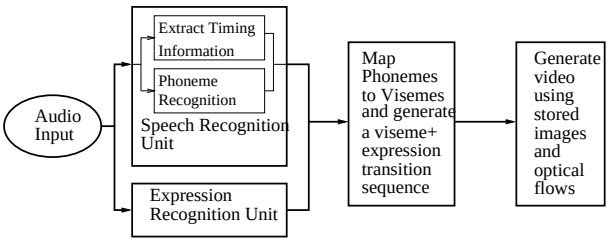

Figure 3: Synthesis Module

Phoneme	Viseme No	Phoneme	Viseme No
a, h	$Viseme1$	g, k, d, n, t, y	$Viseme7$
e, i	$Viseme2$	f, v, w	$Viseme8$
l	$Viseme3$	h, j, s, z	$Viseme9$
r	$Viseme4$	sh, ch	$Viseme10$
o, u	$Viseme5$	th	$Viseme11$
p, b, m	$Viseme6$	$silence$	$Viseme12$

Table 1. Phoneme to Viseme Mapping Rule

Figure 3 shows the synthesis module. From an incoming audio stream timing information, phoneme transitions and expressions are extracted. The phonemes are then mapped to the corresponding visemes. This mapping is shown in Table 1. The timing information and phoneme transition can also be extracted for a novel language whose speech recognition engine is not available [6]. The expression recognition unit based on audio gives the correct expression. However in our case the expression maps have been explicitly provided. Together the viseme+expression combination determines the frame to be used from the database, the timing information tells how long this viseme+expression lasts and the phoneme transitions in turn give the viseme transitions. These viseme transitions are brought about using precomputed optical flows.

3. AUDIO-VISUAL ANIMATION

In this section we show how the animation system proposed in this paper achieves a realistic interaction using faces and voices. The voice of the speaker is left unaltered.

3.1. Normalization of images

The system waits for the first occurrence of a viseme+ expression combination and extracts all possible combinations from the audio+video footage. The images so obtained may not be aligned. If these images are used for animation then the resulting sequence will have disturbing and unintended head motions. We therefore need to align the images. We use a method similar to [1] to normalize the images. There are two components of motion between the images, 3-D rigid body motion and non rigid motion. The rigid component is due to the head rotation, translation etc. and the non rigid component is due to changes in expression and lip shape. The face can be approximated as a single plane viewed under a perspective projection [8]. As a result it is possible to describe the optical flows by the following eight-parameter model:

$$u(x,y) = a_0 + a_1x + a_2y + p_0x^2 + p_1xy \quad (1)$$

$$v(x,y) = a_3 + a_4x + a_5y + p_0xy + p_1y^2 \quad (2)$$

Since non rigid motions of facial features are not captured well by this model we can use this model to extract the 3D rigid body component of motion and to align the images. To estimate the parameters we use the approach suggested by Tsai and Huang [10] with modifications. Tsai and Huang's method is based on perspective displacement field model which is different from the kind of model we are using. This method is basically a least square fit over the image gradients and we use Singular Value Decomposition to calculate the above parameters.

Given facial images I_1 and I_2, we first estimate the 3D rigid body motion component from I_2 to I_1. Next, we warp image I_2 using this model to align with I_1 and having viseme shape/expression of I_2. Some images may have slight facial deformation due to the assumed planar model for the face under perspective projection. Given a set of images we can align them with respect to a single image and repeat the whole process iteratively.

3.2. Lip Synchronization with Audio

The timing information is extracted from the incoming audio stream using the speech recognition unit. The lip movement synchronization and the extent of morph is governed by this timing information. Given two normalized viseme images intermediate frames are generated using optical flow based morphing techniques similar to [5]. Suppose the viseme transition between v_1 and v_2 occurs in time T. To generate a frame at time $0 < t < T$ we use image warping using the optical flows. We calculate the optical flows from v_1 to v_2 (say OF_1) and from v_2 to v_1 (say OF_2). The viseme v_1 is warped along OF_1 and viseme v_2 along OF_2. The two obtained images are cross dissolved in a weighted sense to obtain a final image which is the generated frame.

We restrict the extent of the morph depending upon the viseme and the duration of viseme transition. Figure 4 shows the rules used by our system. Consider a viseme transition between v_a and v_b in duration Tc. Now, if $Tc < Th$, where Th is a threshold that is heuristically set, we generate the morph until $t = Tc/Th$. But there is a catch, consider a transition from viseme v_b to v_c in duration Tn. If $Tn > Th$ then viseme v_b needs to be emphasized and hence the morph to v_b should be complete. In this case we extend the duration of transition $v_a - v_b$ and reduce the duration of transition $v_b - v_c$ by Q, where $Q = Min(Th - Tc, Tn - Th)$. If the transition $v_b - v_c$ was long enough then viseme v_b would be morphed from v_a. Further, visemes that represent p, b, m and v, f have to be morphed completely because these visemes involve lip closure or near closure. So if transition occurs to any of these visemes, then the morph is completed irrespective of the duration.

Suppose v_b was not completely morphed then to generate the morph to viseme v_c we cannot use the optical flows between v_b and v_c computed using the images in our database. We need to know the optical flow between the generated (and incomplete) viseme v_b and v_c. Since the optical flow computations are too costly and almost impossible in real time, we use the transitivity between the optical flows $v_a - v_b$ and $v_b - v_c$ to calculate an approximate optical flow, which is used to generate the morph. Our system uses a threshold $Th = 100ms$ at $30fps$.

3.3. Facial Expression Synthesis

In the background processing module we complete the set of vimseme+expression combinations. The central problem we solve is that given visemes v_1 and v_2 with facial expression e_1 and

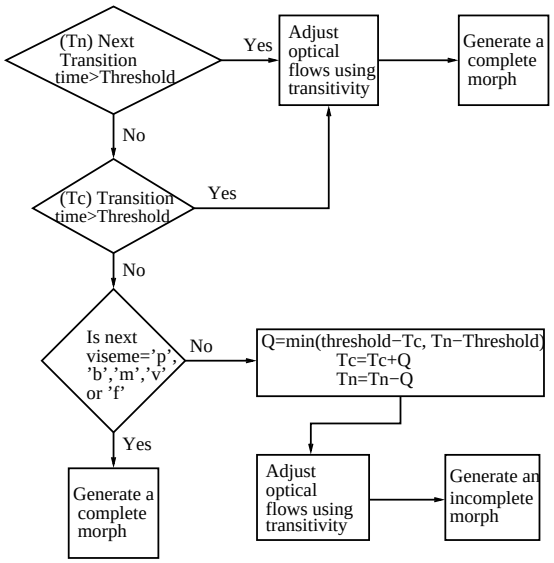

Figure 4: Audio Synchronization

viseme v_1 with facial expression e_2, how to generate viseme v_2 with facial expression e_2 i.e. given (e_1, v_1), (e_1, v_2) and (e_2, v_1) we want to generate (e_2, v_2). We exploit the similarity that is found in transitions between visemes for every facial expression. Here an important task is to appropriately insert the new facial features of viseme v_2 (not present in v_1) and to delete the facial features not present in viseme v_2 (but present in v_1). We employ optical flow techniques to accomplish all these tasks.

We accomplish this as follows (see Figure 5). Find the correspondence of pixels in (e_1, v_1) going to (e_1, v_2), call it $flow_1$ and from (e_1, v_1) to (e_2, v_1), call it $flow_2$. Now put the velocity of every pixel in (e_1, v_1) given by $flow_1$ on the corresponding pixel of (e_2, v_1) (found according to $flow_2$). Call the optical flow of (e_2, v_1) thus obtained as $flow_{new}$. Generate (e_2, v_2) from (e_2, v_1) using $flow_{new}$.

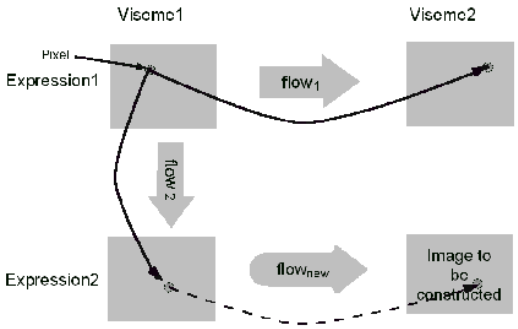

Figure 5: New Viseme-Expression Pair Generation

To introduce the new features that appear in viseme v_2 (see Figure 6), detect the facial features that appear in (e_1, v_2) which were not there in (e_1, v_1) using $flow_1$. The pixels in (e_1, v_2) which do not correspond to any pixel in (e_1, v_1) stand for the new features. Find the correspondence of pixels in (e_1, v_2) going to (e_1, v_1), call this $flow_3$. Carry the pixels (new features) found using $flow_1$ to (e_2, v_2) in the same way as the nearby corresponding pixels in (e_1, v_1) go to (e_2, v_1) according to $flow_2$. These nearby corresponding pixels in (e_1, v_1) are determined by the correspondence of pixels given by $flow_3$ on the nearby pixels in (e_1, v_2).

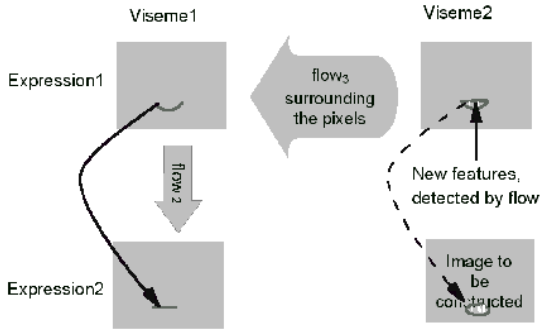

Figure 6: Introducing New Features

To suppress the facial features disappearing in viseme v_2 (see Figure 7), detect the features that are present in (e_1, v_1) but which disappear in (e_1, v_2) using $flow_3$. The pixels in (e_1, v_1) which do not correspond to any pixel in (e_1, v_2) stand for the disappearing features. Find where these pixels go in (e_2, v_1) using $flow_2$. While constructing the new image from (e_2, v_1) suppress these pixels. This way these features won't appear in the new image. Figure 8 and Figure 9 are examples of new viseme+expression combinations generated from the existing ones.

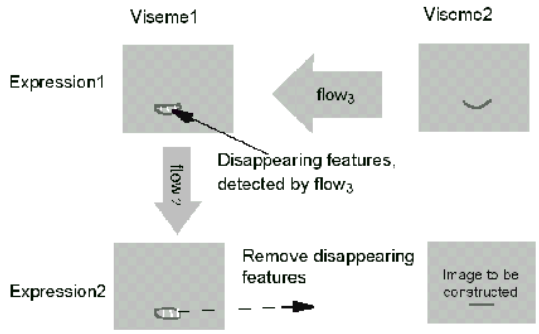

Figure 7: Suppressing Disappearing Features

4. SYSTEM EVALUATION

Various application scenarios motivate audio driven facial animation. These include bandwidth reduction for video teleconferencing, movie dubbing, user-interface agents and avatars, and multimedia telephones for hard of hearing people. Simple experiments have shown the value of the visual channel in speech comprehension [7], for example the McGurk effect. In many scenarios it is possible that the listener is in a crowded and noisy environment. Vision adds redundancy to the signal and provides evidence of those cues that would be irreversibly masked by noise or hearing impairments [7]. The system was tested over one person with hearing impairment over different sentences. The audio was left

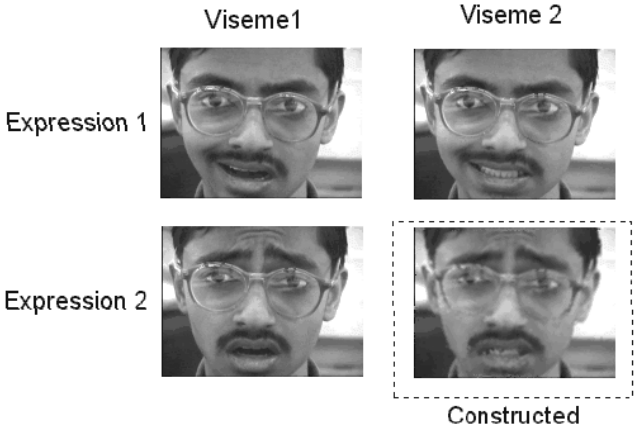

Figure 8: Existing Images and the Constructed Image with New Features Appearing

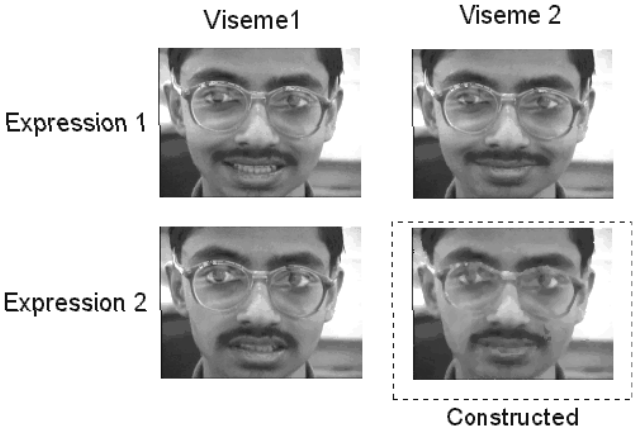

Figure 9: Existing Images and the Constructed Image with Disappearing Features

clean in all cases. It was found that the addition of video improved speech understanding by at least 50%.

This system is valuable where video has to be generated. Examples of such scenarios include:

Visual e-mail: At the receiving end the email is ”read out” by the sender. The receiver mailbox activates the correct person, to read out the mail, by matching the address.

Newscast: In many cases involving a field reporter, the audio is available but due to various reasons, the corresponding video is not available. Usually a photograph of the person is shown on the TV screen along with the audio. Using the system presented here, a video of the person speaking can be generated and shown along with the audio. Vision directs the listener's attention and sustains interest.

Entertainment: Making people say things they normally would not. For example popular actors are made to say different things and ”interact” with people.

Many other uses of this system can be thought of. A talking face has the advantage of directing the listener's attention and sustaining interest. An audio-visual reality is created if the animated face is able to hold human attention and successfully engage the person in useful conversation or task. To obtain feedback on the quality of the animation, clips were made and shown to a number of people. The feedback was very positive and in many cases, unless specifically mentioned, the animated clip passed off as an original. However, when many synthesized expression visemes are used in the animation, noticeable artifacts at the teeth and lips start appearing.

5. CONCLUSIONS

An automated system for creating an additional channel for communication is presented. From audio and a few images of a person, a facial animation with lip sync and appropriate expressions is generated. The animation looks realistic and individual variability is preserved. It is also possible to generate new lip shapes in expressions previously not seen by the system. For the future it would be worthwhile to consider including other features like correct gaze following, controlled pose variation, eyebrow movement and eye blinking in the animation system.

6. REFERENCES

[1] M. J. Black and Y. Yacoob, “Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion,” *Proceedings of the Fifth International Conference on Computer Vision*, pp. 374-381, 1995.

[2] J. F. Cohn and G. S. Katz, “Bimodal expression of emotion by face and voice,” *Proceedings of the Sixth ACM International Multimedia Conference on Face/Gesture Recognition and Their Applications*, ACM Press, pp. 41-44, 1998.

[3] E. Cosatto and H. P. Graf, “Photo-realistic talking-heads from image samples,” *IEEE trans. Multimedia*, Vol. 2, No. 3, pp. 152-163, September 2000.

[4] A. A. Essa and A. P. Pentland, “Coding, analysis, interpretation and recognition of facial expressions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 7, pp. 757-763, 1997.

[5] T. Ezzat and T. Poggio, “Miketalk: A talking facial display based on morphing visemes,” *Proceedings of IEEE Computer Animation ’98*, pp. 96-102, 1998.

[6] T. A. Faruquie, C. Neti, N. Rajput, L. V. Subramaniam and A. Verma, “Translingual visual speech synthesis,” *IEEE International conference on Multimedia and Exposition*, 30 July - 02 Aug, 2000.

[7] D. W. Massaro, *Perceiving talking faces: From speech perception to behavioural principles*, MIT Press, 1998.

[8] F. I. Parke and K. Waters, *Computer Facial Animation*, Wellesley MA: A K Peters, 1996.

[9] V. C. Tartter and D. Braun,“Hearing smiles and frowns in normal and whisper register,” *Journal of the Acoustical Society of America*, 96, pp. 2101-2107, 1998.

[10] R. Y. Tsai and T. S. Huang, “Estimating three- dimensional motion parameters of a rigid planar patch,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 29, No. 6, pp. 1147-1152, 1981.