Text-to-Audiovisual Speech Synthesizer

Udit Kumar Goyal, Ashish Kapoor and Prem Kalra

Department of Computer Science and Engineering,
Indian Institute of Technology, Delhi
pkalra@cse.iitd.ernet.in

Abstract. This paper describes a text-to-audiovisual speech synthesizer system incorporating the head and eye movements. The face is modeled using a set of images of a human subject. Visemes, that are a set of lip images of the phonemes, are extracted from a recorded video. A smooth transition between visemes is achieved by morphing along the correspondence between the visemes obtained by optical flows. This paper also describes methods for introducing nonverbal mechanisms in visual speech communication such as eye blinks and head nods. For eye movements, a simple mask based approach is used. View morphing is used to generate the head movement. A complete audiovisual sequence is constructed by concatenating the viseme transitions and synchronizing the visual stream with the audio stream. An effort has been made to integrate all these features into a single system, which takes text, head and eye movement parameters and produces the audiovisual stream.

1. Introduction

The visual channel in speech communication is of great importance, a view of a face can improve intelligibility of both natural and synthetic speech. Due to the bimodality in speech perception, audiovisual 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 audiovisual testing for bimodal perceptional quality of video conferencing systems [1]. In addition to the bimodal characteristics of speech perception, speech production is also bimodal in nature. Moreover, visual signals can express emotions, add emphasis to the speech and support the interaction in a dialogue situation. This makes the use of a face to create audiovisual speech synthesis an exciting possibility, with applications such as multimodal user-interfaces. Text-to-visual speech synthesis (TTVS) systems have conventional applications in computer animation, its use in communication is becoming important as it offers a solution to human 'face to face' communication and human communication with a computer. These TTVS systems also find applications in graphical user interfaces and virtual reality where instead of being interested in face-to-face communication, we are interested in using a human-like or 'personable' talking head as an interface. These systems can be deployed as visual desktop agents, digital actors, and virtual avatars. This system can also be used as a tool to interpret lip and facial movements to help hearing-impaired to understand speech.

This paper describes a text-to-audiovisual speech synthesizer system, which takes text as input and constructs an audiovisual sequence enunciating the text. This system introduces both eye and head movements to make the sequence more videorealistic.

J.-C. Heudin (Ed.): VW 2000, LNAI 1834, pp. 256-269, 2000.

The 3D model based facial animation techniques though may be flexible, lack video realism. In this paper, an image based approach has been used in which the facial model is constructed using a collection of images captured of the human subject. This results in a great improvement in the levels of video realism. Bregler, et al. [3] described an image based approach in which talking facial model was composed of a set of audiovisual sequences extracted from a large audiovisual corpus. Each short sequence corresponds to a triphone segment and a large database is built containing all the triphones. A new audiovisual sequence was constructed by concatenating the appropriate triphone sequences from the database. The problem with this approach was that it requires a very large number of images to cover all possible triphones context, which seems to be an overly redundant sampling of human lip configurations.

Cosatto and Graf [4] have described an approach, which attempts to reduce this redundancy by parameterizing the space of lip positions, mainly the lip width, position of the upper lip, and the position of the lower lip. This lip space was then populated using images from the recorded corpus. Synthesis was performed by traversing trajectories in that imposed lip space. The trajectories were created using Cohen-Massaro's coarticulation rules [5]. The problem with this approach was that if the lip space was not densely populated, the animations might produce jerks.

Another approach was used by Scott, et al. [6] in which facial model was composed of a set of 40-50 visemes, which were the visual manifestation of phonemes. They have used a morphing algorithm that is capable of transitioning smoothly between the various mouth shapes. However, morphing required a lot of user intervention, making the process tedious and complicated. This work was further explored by Tonny Ezzat and Tomaso Poggio [7]. They use a method for morphing developed by Beymer, et al. [8], which did not require user intervention and was capable of modeling rigid facial transformations such as pose changes and non-rigid transformations such as smiles.

This paper explores further the use of viseme morphing representation for synthesis of human visual speech by introducing nonverbal mechanisms in visual speech communication such as eye blinks, eye gaze changes, eye brow movements and head nods due to which the talking facial model has became more lifelike. One approach was proposed by Tony Ezzat and Poggio [9] using a learning network, but that was found to be computationally inefficient. For eye movements, a simple cut-and-paste approach has been used. Head movements are generated using view morphing [10] in which valid intermediate views are generated by extending the existing morphing techniques.

2. Overview

An overview of the system is shown in figure 1. For converting text to speech (TTS), Festival speech synthesis system is used which was developed by Alan Black, Paul Taylor, and colleagues at the University of Edinburgh [14]. Festival system contains Natural Language Processing (NLP) unit which takes text as an input and produces

the timing and phonetic parameters. It also contains an audio speech-processing module that converts the input text into an audio stream enunciating the text

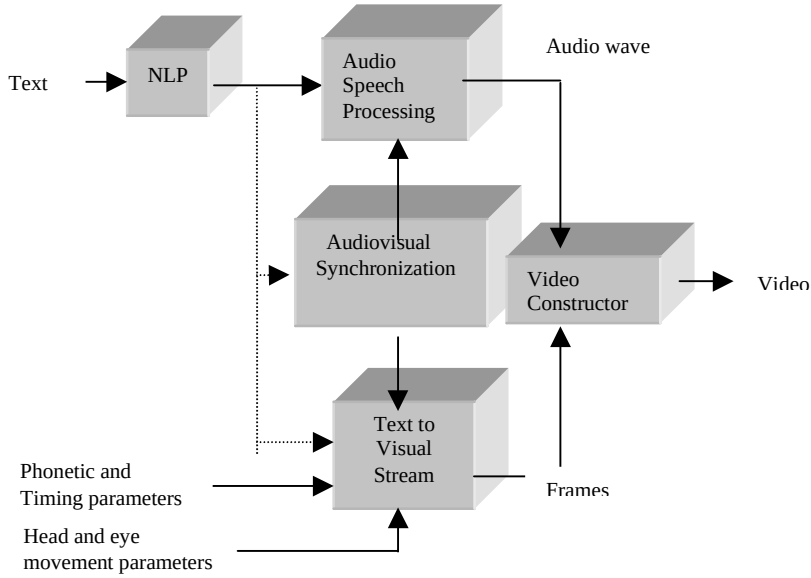

Fig. 1. Overview of the TTVS System.

One primary concern is synthesis of the visual speech streams. The entire task of visual speech processing can be divided into three sub-tasks: firstly, to develop a *text to visual stream* module that will convert the phonetic and timing output streams generated by Festival system into a visual stream of a face enunciating that text. Secondly, to develop an *audiovisual synchronization* module that will synchronize the audio and visual streams so as to produce the final audiovisual stream. Thirdly, to extend visual speech processing module to take head and eye movement parameters as input, and reflect the corresponding changes in the visual streams so as to make the facial animation more lifelike.

3. Text-to-Audiovisual Stream Conversion

The entire process of text-to-audiovisual stream conversion can be divided into four sub-tasks: viseme extraction, morphing, morph concatenation, and finally audiovisual synchronization. These sub-tasks are discussed in detail in the ensuing sections.

3.1 Viseme Extraction

The basic assumption of facial synthesis used in this approach is that the complete set of mouth shapes associated with human speech is spanned by a finite set of *visemes*. The term *viseme* is used to denote lip image extracted for the phoneme. Due to this assumption, a particular visual corpus has to be recorded which elicitates one instantiation for each viseme.

Since the Festival TTS system produces a stream of phonemes corresponding to an input text, there is a need to *map* from the set of phonemes used by the TTS to a set of visemes so as to produce the visual stream. If all the American English phonemes are covered, the one-to-one mapping between the phonemes and visemes will result in a total of 52 images, out of which 24 represent the consonants, 12 represents the monophthong, and 16 represent the diphthongs. Since current viseme literature indicates that mapping between phonemes and visemes are many-to-one, the viseme set is reduced by grouping the visemes together that are visually alike. This results in the reduction of total number of visemes, six represent the 24 consonants, seven represent the 12 monophthong phonemes and one for silence viseme.

For diphthongs, that are vocalic phonemes involving a quick transition between two underlying vowel nuclei, we use two images to model them visually, one to represent the first vowel nucleus and the other to represent the second. All these vowel nuclei are represented by the corresponding monophthong visemes. The only exception occurs in case of two nuclei: second nucleus of \au\ diphthong and first nucleus of the \ou\ diphthong. For them, two separate visemes are extracted. Hence, the final reduced set contains a ***total of 16 visemes*** that are to be extracted.

monophthongs		consonants	
/ i, ii /	sheep	/ p, b, m /	about
/ a, e /	bed	/ f, v /	father
/ aa, o /	father	/t, d, s, z, th, dh/	thank
/ uh, @ /	bud	/ w, r /	was
/ ir /	bird	/ch, jh, sh, jh /	sheep
/ oo /	baud	/ k, g, n, l, ng, h, y/	key
/ uu, u /	boot		
diphthongs			
/ w-au /	about		
/ o-ou /	boat		

Fig. 2. Phonemes recorded for extracting visemes. The underlined portion of each word corresponds to the target phoneme being recorded.

Initially, a video of a human subject enunciating a set of keywords is recorded. A set of keywords is chosen in such a way that it covers all the required phonemes. The recorded corpus is shown in the figure 2. After the recording of whole corpus, it is digitized and one viseme is extracted for each phoneme.

The final reduced set of extracted visemes is shown in figure 3. As the figure shows, a single viseme is extracted for many phonemes as they look alike visually.

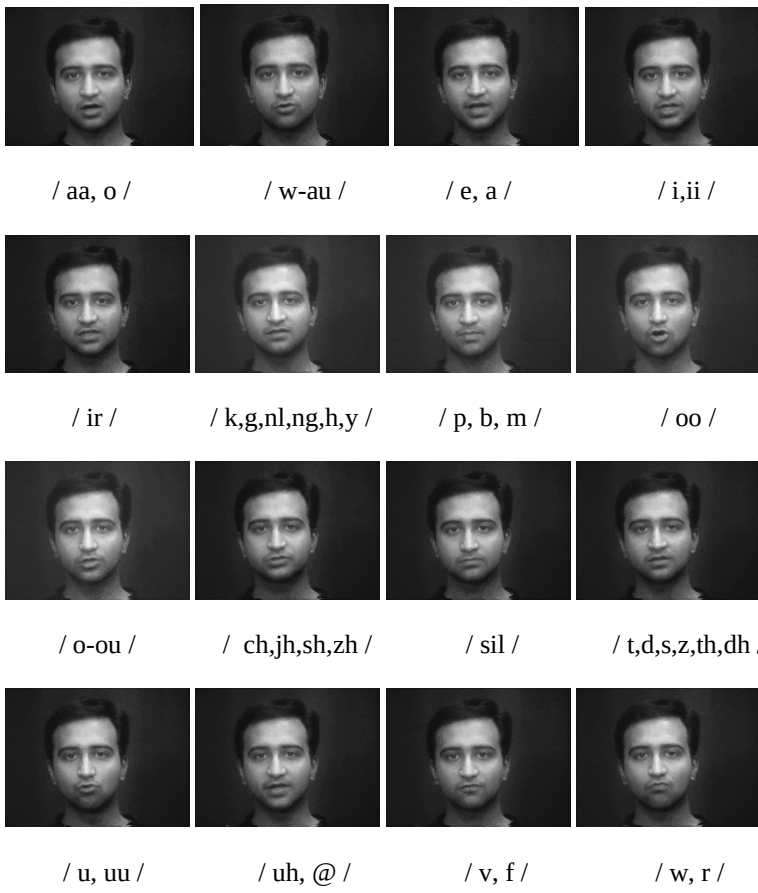

Fig. 3. Reduced set of extracted visemes

3.2 Morphing

After extracting all the visemes, a correspondence between two visemes is computed using optical flow as given by Horn and Schnuck [11]. Optical flow technique has been used since visemes belong to one single object that is undergoing motion. An

advantage of using optical flow technique is that it allows automatic determination of correspondence vectors between the source and destination images. A smooth transition between viseme images is achieved using morphing along the correspondence between the visemes. In the morphing process, first forward and reverse warping is carried out to produce intermediate warps, which are then cross-dissolved to produce the intermediate morphs.

3.3 Morph Concatenation

To construct a visual stream of the input text, we simply concatenate the appropriate viseme morphs together. For example, the word **"man"**, which has a phonetic transcription of \m-a-n\, is composed of two visemes morphs transitions \m-a\ and \a-n\, that are then put together and played seamlessly one right after the another. It also includes the transition from silence viseme in the start and at the end of the word.

3.4 Audiovisual Synchronization

After constructing the visual stream, next step is to synchronize the visual stream with the audio stream. To synchronize the audio speech stream and the visual stream, the total duration T of the audio stream is computed as follows.

$$T = \Sigma_i \, l(D_i) \tag{1}$$

where, $l(D_i)$ denotes the duration (in sec) of each diphone D_i as computed by Festival system.

Viseme transition streams are then created consisting of two endpoint visemes and the optical flow correspondence vectors between them. The duration of each viseme transition $l(V_i)$ is set to be equal to the duration of corresponding diphone $l(D_i)$. The start index in time of each viseme transition $s(V_i)$ is computed as

$$s(V_i) = \begin{cases} 0 & \text{if i=0} \\ s(V_{i-1}) + l(V_{i-1}) & \text{otherwise} \end{cases} \tag{2}$$

Finally, the *video stream* is constructed by a sequence of frames that sample the chosen viseme transitions. For a frame rate F, we need to create TF frames. This implies that start index in time of k^{th} frame is

$$s(F_k) = \frac{k}{F} \tag{3}$$

The frames are then synthesized by setting the morph parameter for each frame to be

$$s_k = \frac{s(F_k) - s(V_i)}{l(V_i)} \qquad \text{if } s(F_k) - s(V_i) < l(V_i) \tag{4}$$

The morph parameter is simply the ratio of the time elapsed from the start of a viseme transition to the frame, and the entire duration of the viseme transition. The condition on right hand side of the above equation ensures that correct viseme is chosen to synchronize a particular frame. Considering figure 4, it implies that frames 0,1,2,3, and 4 are synthesized from the \m-a\ viseme transition, while frames 5,6,7,8 and 9 are synthesized from the \a-n\ viseme transition.

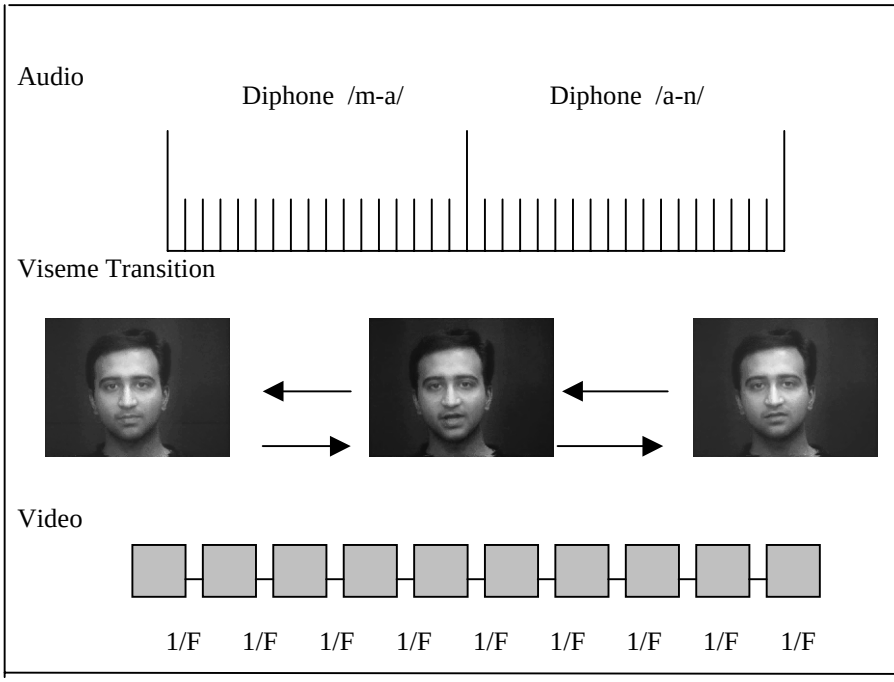

Fig. 4. Lip Synchronization

The variation of the morph parameter α for each viseme transition is shown in figure 5. This indicates that for each viseme transition, the morph parameter α varies linearly from 0 to 1 resulting in the saw-tooth type variation.

Finally, each frame is synthesized using the morph algorithm discussed earlier and hence, the final visual sequence is constructed by concatenating the viseme transitions, played in synchrony with the audio speech signal generated by the TTS system. It has been found that lip-sync module produces very good quality synchronization between the audio and the video.

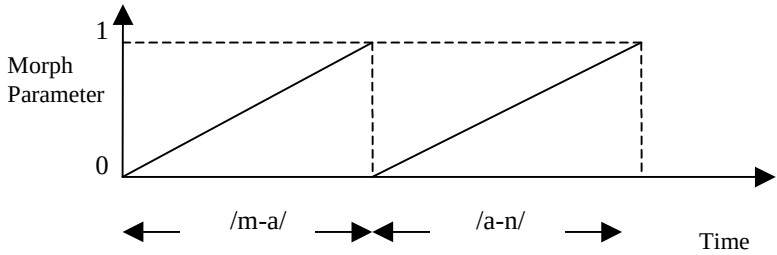

Fig. 5. Morph parameter variation

4. Eye Movement

Although conversion of text to audiovisual speech stream gives good animation results, yet the video does not look much video realistic since only the lips of the face are moving. As a step towards making it more video realistic, eye movement has been incorporated in the facial model.

A simple ***mask-based*** approach has been used to incorporate eye movement in the facial model. Since eyes affect only upper portion of the face and do not overlap with the lip movement, mask-based approach can be used. First, images are extracted for the various eye movements like opening and closing of eyes, raised eyebrows, eyeball movement, etc. While taking the sample images, it has been assumed that head remains still in all the sample images. A *base* image of the face is taken which in our case is taken to be the same as /*sil*/ viseme. The next step is to define a mask that consists of all the pixels contained in the portion covered by left and the right eye.

After defining the mask, depending on the parameters that control the position of the eye, morphing is carried out between the source and destination images. Source image is taken to be the base image and destination image can be closed eye image, raised eyebrow image, or left eyeball movement image, etc. The intermediate image is determined using the morph eye parameter, the mask is then applied to the intermediate image to find the intermediate eye position, which is then pasted on an image of the face giving the resulting intermediate image. Since, the eye movement is performed in parallel with the text-to-audiovisual conversion, the image on which the eye mask is pasted, is taken to be the intermediate image generated during the text-to-audiovisual stream conversion process. In this way, the effect of eye movements is achieved in parallel with the text-to-audiovisual conversion, thus resulting in an increase in video-realism. This is shown in figure 6.

We have associated separate parameters with the eye movement. These parameters will be the start time and the duration of eye movement. The start time can also be specified as a percentage of the entire duration of the audiovisual stream. From the start time and the duration of the eye movement, end time of the eye movement can be determined.

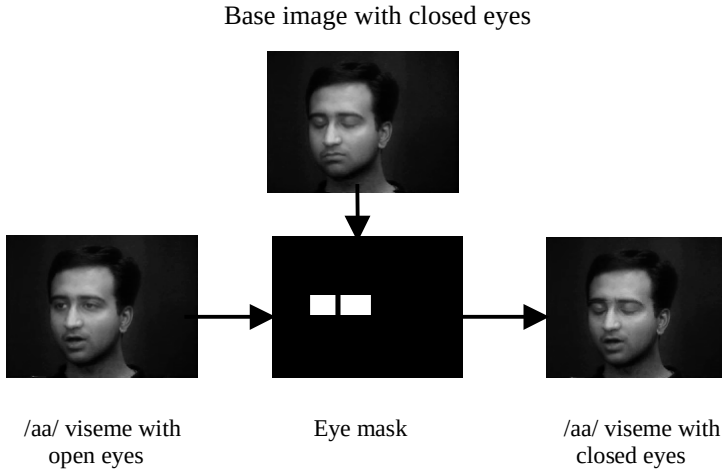

Fig. 6. Eye Masking

During this duration, the eye movement is carried out such that eyes will first make transition from open eyes to closed eyes for half of the duration, and closed to open eyes for the remaining half of duration of eye movement.

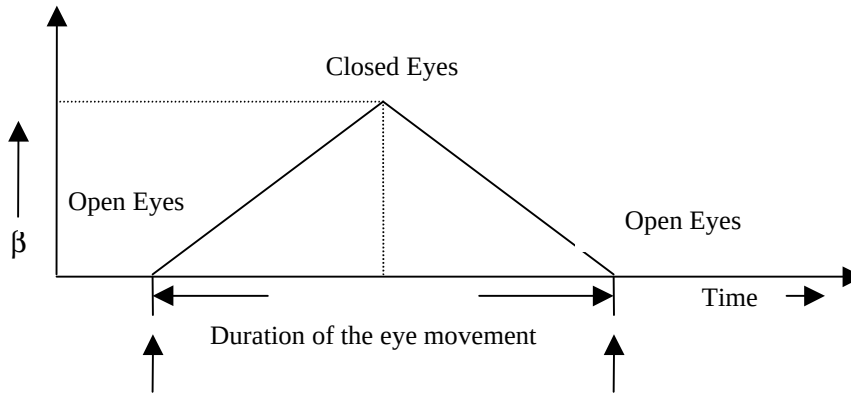

Fig. 7. Transitions for eye movement

This transition from open to closed eyes and vice versa is controlled by a separate morph parameter β. This parameter will control the extent of morph that is to be done to determine the intermediate position of eyes in the image. As figure 7 indicates, the morph parameter β varies linearly from 0 to 1 to produce intermediate eye images for open to close eyes transition for half of the duration and then varies linearly from 1 to 0 for close to open eyes transition for the rest of the duration.

5. Head Movement

The head being stable for a long time makes an impression of a dummy. The head movement is introduced to make the animation more realistic. We use view morphing approach as proposed by Seitz and Dyer [10] to interpolate human face in different poses. View morphing is a simple extension of the normal morphing technique that allows current morphing techniques to synthesize changes in viewpoint and other 3D effects. This technique is *shape-preserving* i.e., from two images of a particular object, it produces a new image representing a view of the same object.

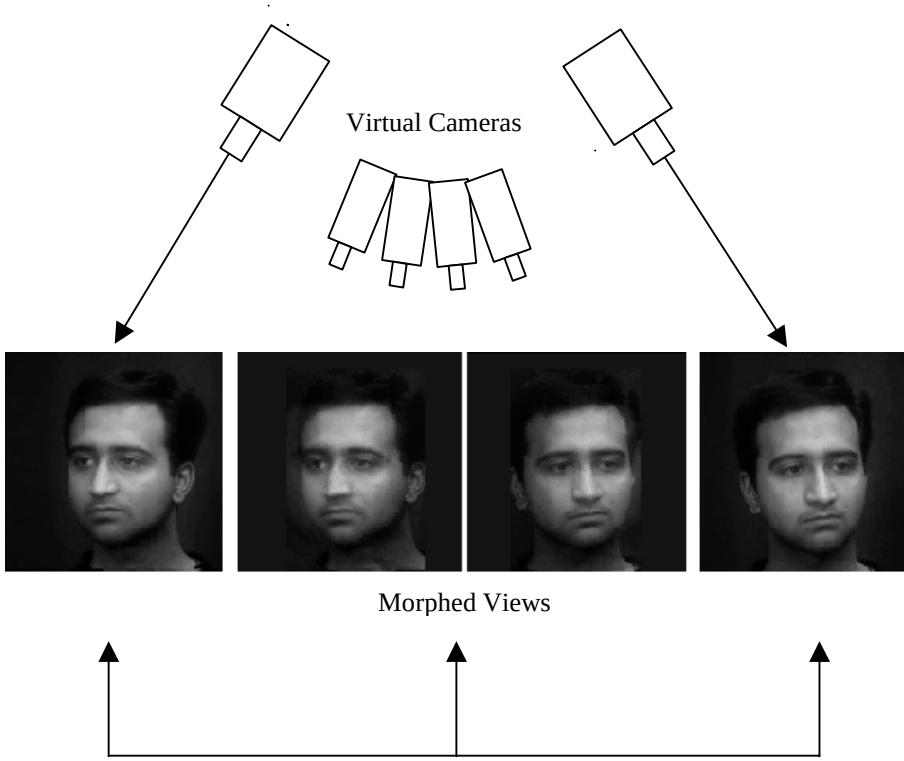

Fig. 8. View morphing of an object taken from the different viewpoints produces intermediate images that preserve the shape.

If the different views of the same object are parallel, then normal morphing techniques produce valid intermediate views. The term *valid* means that they preserves the shape. However, for non-parallel views, the intermediate views are not valid, i.e. they do not preserve the shape. To generate the valid intermediate views I_α between two images I_0 and I_1, where α lies between 0 and 1, Seitz [10] described an approach which requires following steps:

a) Prewarping the images I_0 and I_1.
b) Generate intermediate image $I_\alpha^{\sim}$ from the prewarped images using morphing techniques.
c) Postwarp image $I_\alpha^{\sim}$ to produce final intermediate view I_α

The reader is referred to [10] for details of this approach. As shown in figure 8, it appears that the intermediate images are the head image at the intermediate positions (α=0.7, α=0.3) while moving from left to right. In our case, during the construction of visual stream, a pause is inserted in the visual stream. During the pause interval, head is moved from left to right or vice-versa to give a feel of realistic animation with head turning. Similarly, effects such as head nod and head roll can be produced using this approach.

6. Integration

This system conceives speech – affecting a part of the mouth, expressions – consisting of eye movements, and head movements as three channels or streams. The integration of these channels involves superposition or overlaying of associated actions to each channel. This requires temporal specification of each action constituting a particular channel. This contains the start and the duration of the action. A scripting language may be designed to incorporate the action schedule of actions in all three channels. Figure 9 depicts temporal integration of channels in a form of action chart.

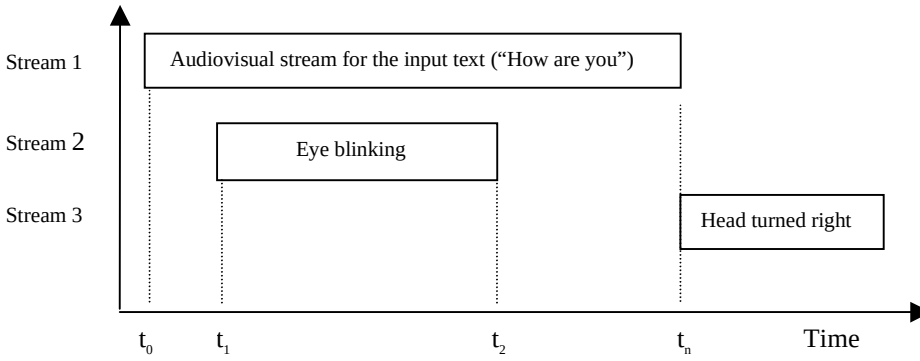

Fig. 9. Integration of the system

7. Results

We have developed a text-to-audiovisual speech synthesizer system. Several audiovisual sentences are synthesized to test the process of visual speech synthesis, audio synchronization, eye and head movements. Some of the generated audiovisual sequences can be accessed at http://www.cse.iitd.ernet.in/~pkalra/VW2000/results.

These include sequences where the actor speaks sentences like ‘Let me tell where you have gone’ and ‘Good Morning Sir, thank you for coming here’. Some sentences with simultaneous non-verbal clues like eye blinks have also been generated. Finally, an example is included where all three channels --speech, eye movements and head movement -- have been integrated. Figure 10 shows a sequence where the model speaks a sentence while blinking its eyes and then turns its head from left to right.

Fig. 10. A generated audiovisual sequence where the model first speaks, blinks its eye and then turns its head.

There are some constraints imposed in the process of sequence generation. While recording the video sequence, the subject should refrain from moving the head. If there is a relative head movement among the images, it may cause discontinuity or jerkiness in the animation. Similarly, artifacts may be observed while masking for eye movements, when images get miss-aligned due to the head movement. Methods for stablizing are being explored to account for small head movements.

8. Conclusion

In this paper we present, a text-to-audiovisual speech synthesis system capable of carrying out text to audiovisual conversion. The efforts have been mainly focused on making the system more video-realistic. This system also takes care of nonverbal mechanisms for visual speech communication like eye blinking, eye ball movement, eyebrow raising, etc. In addition, the system includes head movement, which has been incorporated during the pause or at the end of the sentence.

The work is being extended to introduce co-articulation in the facial model [15]. The lack of parameterization in the image-based model makes it difficult to use the techniques used in 3D facial animation models for introducing co-articulation. The introduction of co-articulation in the synthesis would further improve the audiovisual realism of the system. Introducing composition of speech with facial expressions that affect the mouth region can further enhance the system. The Festival system supports intonation parameters, we plan to incorporate them to change the emotion accordingly. Further there is a need to incorporate the head movement while enunciating the text.

Acknowledgments

Authors would like to extend their thanks to Vineet Rajosi Sharma, who helped in getting the samples made.

References

1. Tsuhan Chen and Ram R. Rao, *Audio-Visual Integration in Multimodal Communication,* Proc. IEEE, Vol 86, No. 5, pages 837-852.

2. G. Wolberg. Digital Image Warping, IEEE Computer Society Press, Los Alamitos, *C.A., 1990.*

3. C.Bergler, M.Covell and M.Slaney. *Video Rewrite. Driving visual speech with audio.* In SIGGRAPH'97 Proceedings, Los Angeles, CA, August 1997.

4. E.Cosatto and H.Graf. *Sample based synthesis of photorealistic talking heads*. In Proceedings of Computer Animation'98, pages 103-110, Philadelphia, Pennsylvania, 1998.

5. M.M.Cohen and D.W.Massaro, *Modeling coarticulation in synthetic visual speech*. In N.M.Thalmann and D.Thalmann, editors, Models and Techniques in Computer Animation, pages 138-156, Springer-Verley, Tokyo, 1993.

6. S.H.Watson, J.P.Wright, K.C.Scott, D.S.Kagels, D.Freda and K.J.Hussey. *An advanced morphing algorithm for interpolating phoneme images to simulate* speech. Jet Propulsion Laboratory, California Institute of Technology, 1997.

7. Tony Ezzat and Tomaso Poggio. *Visual Speech Synthesis by Morphing Visemes (MikeTalk)*. A.I Memo No: 1658, MIT AI Lab, May 1999.

8. D.Beymer, A. Shashua and T. Poggio. *Example based image analysis and synthesis.* Technical Report 1431, MIT AI Lab, 1993.

9. Tony Ezzat and Tomaso Poggio. *Facial Analysis and Synthesis using Image-Based Models*. In Proceedings of the Workshop on the Algorithmic Foundations of Robotics, Toulouse, France, August 1996.

10. Steven M. Seitz and Charles R. Dyer. *View Morphing*. In Proceedings of SIGGRAPH'96, pages 21-30, 1996.

11. B.K.P Horn and B.G.Schnuck. *Determining Optical flow.* Artificial Intelligence, 17:185-203, 1981.

12. F. Parke and K. Waters. *Computer Facial Animation.* A. K. Peters, Wellesley, Massachusetts, 1996.

13. Alan Watt and Fabio Policarpo. *The Computer Image.* ACM Press, New York, SIGGRAPH Series, New York.

14. Black and P.Taylor. *The Festival Speech Synthesis System.* University of Edinburgh, 1997.

15. Catherine Pelachaud. Communication and Coarticulation in Facial Animation. Ph.D. Thesis, Department of Computer and Information Science, Univ. of Pennsylvania, Philadelphia, 1991.