VTalk: A System for generating Text-to-Audio-Visual Speech

Prem Kalra, Ashish Kapoor and Udit Kumar Goyal

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

Abstract

This paper describes VTalk, a system for synthesizing text-to-audiovisual speech (TTAVS), where the input text is converted into an audiovisual speech stream incorporating the head and eye movements. It is an image-based system, where the face is modeled using a set of images of a human subject. A concatination of visemes –the corresponding lip shapes for phonemes— can be used for modeling visual speech. A smooth transition between visemes is achieved using morphing along the correspondence between the visemes obtained by optical flows. The phonemes and timing parameters given by the text-to-speech synthesizer determines the corresponding visemes to be used for the synthesis of the visual stream. We provide a method using polymorphing to incorporate co-articulation during the speech in our TTAVS. We also include nonverbal mechanisms in visual speech communication such as eye blinks and head nods, which make the talking head model more lifelike. For eye movement, a simple mask based approach is employed and view morphing is used to generate the intermediate images for the movement of head. All these features are integrated into a single system, which takes text, head and eye movement parameters as input and produces the complete audiovisual stream.

Keywords: Text-to-audio-visual speech, Morphing, Visemes, Co-articulation.

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-audio-visual speech (TTAVS) synthesis 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 TTAVS 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. They can make the applications more involving and engaging. Such a system can also be used as a tool to interpret lip and facial movements to help hearing-impaired to understand speech.

Computer based facial animation is now a well developed field. For achieving realistic facial movements with speech or modeling a *talking head*, different approaches have been used. Many 3D models of the human face have been developed[2]. 3D models are very flexible for generating movements in 3D and enable viewing in any orientation. However, these models still lack realism and are rendered with a synthetic look. Alternatively, an image based approach is employed which uses warping of sample images[3] [4] [5] [6]. These techniques are capable of producing photo or video realistic animations. Some have used a hybrid approach considering two and half dimensional model[7]. A web-site[8] prepared by Philip Rubin and Eric Vatikiotis-Bateson provides a collection of relevant work in this area.

In this paper we present VTalk, a system of text-to-audio-visual speech synthesis. VTalk uses image-based approach, which takes text as input and constructs an audio-visual sequence enunciating the text. The system also allows eye and head movements to make the sequence more realistic[9]. We also include co-articulation and temporal smoothing to create more visually realistic model for speech[10].

First we give an overview of VTalk. A brief description about the text-to-visual stream conversion is given in Section 3. Then, our approach to co-articulation is presented. Polymorphing[11] is used as a method for combining the effect of preceding and/or succeeding phonemes. Further, we provide our approach for incorporating the eye and head movements. A simple method for audio-visual synchronization is also presented, followed by a brief note on the integration. Finally, we conclude.

2. Overview of VTalk

An overview of VTalk 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[12]. 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.

Our primary concern is synthesis of the visual speech streams. The task of visual speech processing requires developing 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. An *audiovisual synchronization* module synchronizes the audio and visual streams. A module of *co-articulation* is added that takes visemes sequence and timing information as parameters and provides the necessary parameters to generate the final audio-visual sequence. To further enhance the animation, head and eye movement parameters are added.

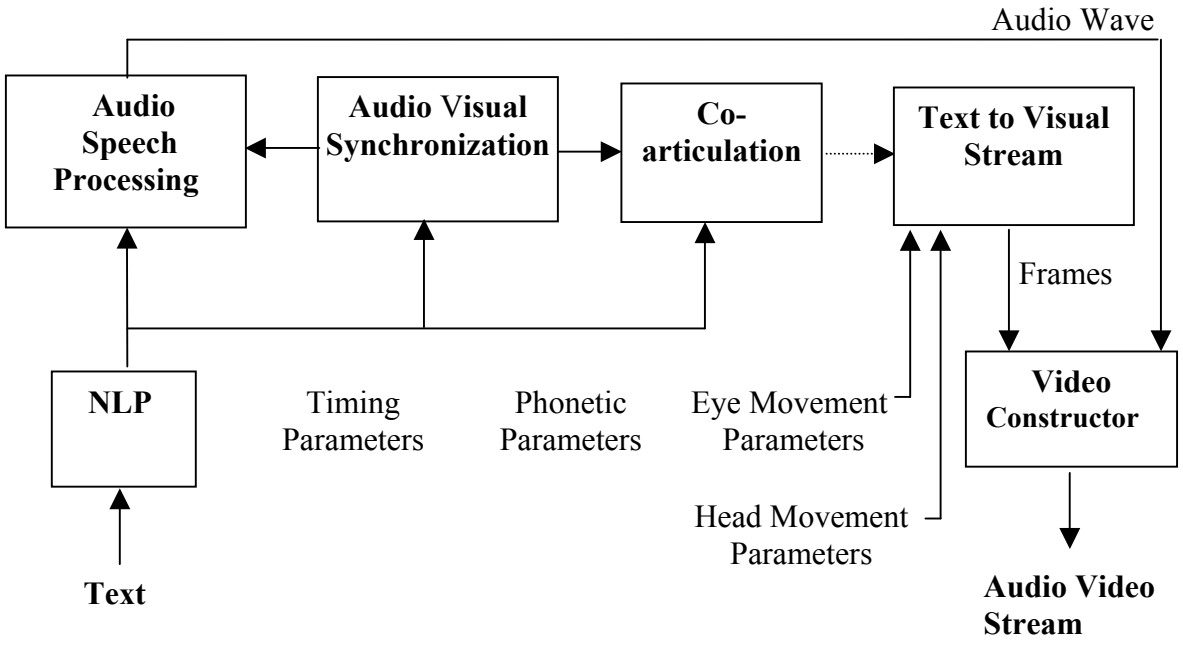

Figure 1: Overview of VTalk

3. Text-to-Visual Speech

Sixteen images corresponding to different lip shapes of sixteen different visemes are stored as a database (see Figure 2). Morphing along the sequence of phonemes spoken generates audiovisual output. After extracting all the visemes, a correspondence between two visemes is computed using optical flow as given by Horn and Schnuck[13]. 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 then reverse warping is carried out to produce intermediate warps, which are then cross-dissolved to produce the intermediate morphs. 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 other. It also includes the transition from silence viseme in the start and at the end of the word.

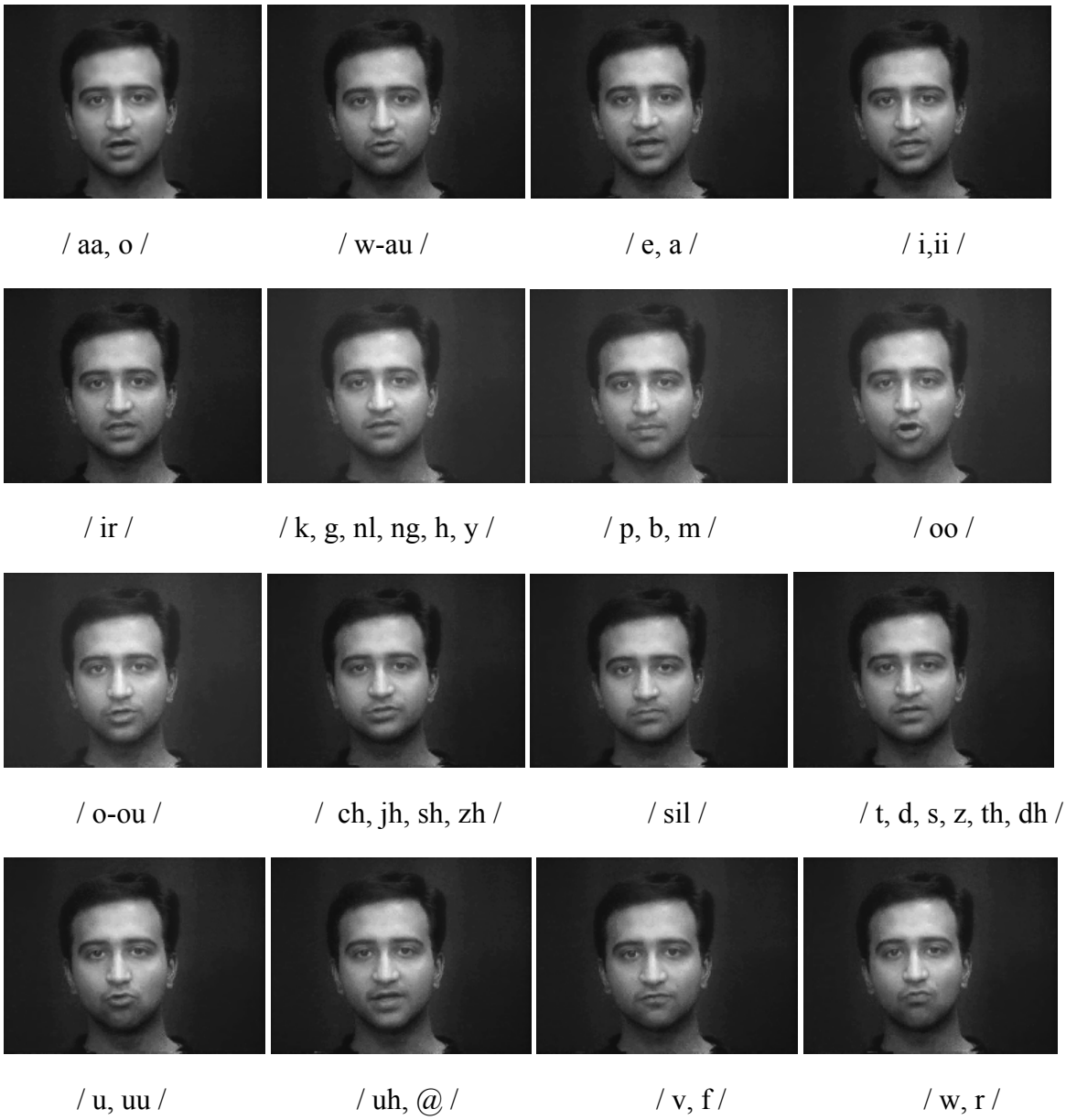

Figure 2: Reduced set of extracted visemes

4. Co-articulation

In actual speech production, there is an overlap in the production of syllables and phonemes that are a sequence of discrete units of speech. Due to this overlap, boundaries between these discrete speech units are blurred, i.e., vocal tract motions associated with producing one phonetic segment overlap the motions for producing surrounding phonetic segments. This overlap is called as *co-articulation*. Co-articulation is the consequence of the physical dynamics of the vocal tract and the vocal tract postures required for various sounds[14]. As there are physical limits to how quickly the speech postures change, rapid sequences of speech sounds require that the posture for one sound anticipate the posture for the next sound or the posture for the current sound is modified by the previous sound. For example, while speaking 'to' the lips get curled (as in /uu/) when /t/ is being enunciated.

An interesting question concerning the perception of visual speech is to what degree co-articulation is important. Benguerel and Pichora-Fuller[15] have examined co-articulation influences on lip reading by hearing impaired and normal hearing individuals. They demonstrated that the degree of recognition dropped in absence of co-articulatory effect. A simple approach to co-articulation problem is to look at the previous, the present, and the next phonemes to determine the current mouth position. However, this may give incorrect results since the current mouth position depend on phonemes up to five positions before or after the current phoneme[16].

Pelachaud[16] has proposed a three-step algorithm for determining the effects of co-articulation. This algorithm depends on the notion of clustering and ranking phoneme lip shapes based on their deformability. In this context, deformability refers to the extent that the lip shape for a phoneme cluster can be modified by corresponding phonemes. Ranking is from the least deformable, such as *f*, *v* cluster to most deformable clusters, such as *s* and *m*. This deformability also depends on the speech rate. Our work in many ways is similar to the idea proposed by Pelachaud[16]. The main difference comes from the fact that we use image-based animation instead of model based approach. In model based approach, FACS[17] or similar scheme can be used for parameterization of basic movements based on 3D muscle actions. However, image based approach requires a different way to parameterize particularly when the sample size is not very large. Since our underlying approach uses morphing methods, we have devised a scheme of parameterization based on polymorphing. Polymorphing allows us to generate an image that is a blend of more than two images. In our approach we have assumed that any viseme transition is affected by at most one more viseme, thus we only consider polymorphing among three images.

4.1. Polymorphing

Traditional image morphing considers only two input images at a time, the source and the target images[18 19]. This limits any morphed image to the features and colors blended from just two input images. Morphing along multiple images involves a blend of several images and this process is called polymorphing[11]. We can consider *n* input images as a point in *(n-1)* dimensional simplex. An in-between image is considered a point in the simplex. All points are given in barycentric coordinates by $\mathbf{b} = (b_1, b_2, \ldots, b_n)$, subject to the constraints $b_i>=0$ and $b_1+b_2+\ldots+b_n=1$. Suppose that we want to generate an in-between image I at a point $\mathbf{b} = (b_1, b_2,\ldots, b_n)$ from input images I_1, I_2,.. I_n. Let W_{ij} be the warp function from image I_i to image I_j. When applied to I_i, W_{ij} generates a warped image, where the features of I_i coincide with their corresponding features in I_j. For each image I_i we define a warp function $W_i = b_1W_{i1} + b_2W_{i2} + \ldots.+ b_nW_{in.}$ This warp function when applied to image I_i results in new image NI_i . The final image is given by $b_1NI_1 + b_2NI_2 + \ldots.+ b_nNI_n$.

4.2. Co-articulation using Polymorphing

In addition to the source and target visemes, we also have a component of a third viseme that affects the shape because of the co-articulation. While generating an intermediate morph (image), we use a set of rules that decides the contribution of each viseme. The sixteen visemes are ranked according to the deformability. The ones that are ranked lower do not affect the visemes ranked above them. For example, viseme /f/ does not get affected by any preceding or succeeding viseme hence it is ranked one. The complete ranking is shown in figure 3. We consider /sil/ as a special in the sense that while looking forward and backward in the co-articulation phase we stop looking beyond the occurrence of /sil/, further it does not affect any other viseme, whatever its rank may be.

1. /v, f/
2. /o-ou /, /oo/, /u, uu/, /w-au/
3. /w, r/
4. /p, b, m/
5. /t, d, s, z, th, dh/
6. /ch, jh, sh, zh/
7. /aa, o/, /i, ii/, /e, a/
8. /k, g, nl, ng, h, y/
9. /uh, @/, /ir/

Special case: /sil/

Figure 3: Ranking of phonemes in increasing order of deformability

We represent each generated image in transition from image I_1 to image I_2 getting affected by a third co-articulatory viseme I_v in barycentric coordinates b=(α, β, γ). Figure 4 shows an example of polymorphing among 3 different visemes. α, β, γ correspond to components of I_1, I_2 and I_v respectively. γ [V_1] denotes the barycentric parameter γ that would be used at the start of transition from V_1 . Further α [V_1] and β [V_1] also denote the values of barycentric parameters α and β at the beginning of transition from V_1.

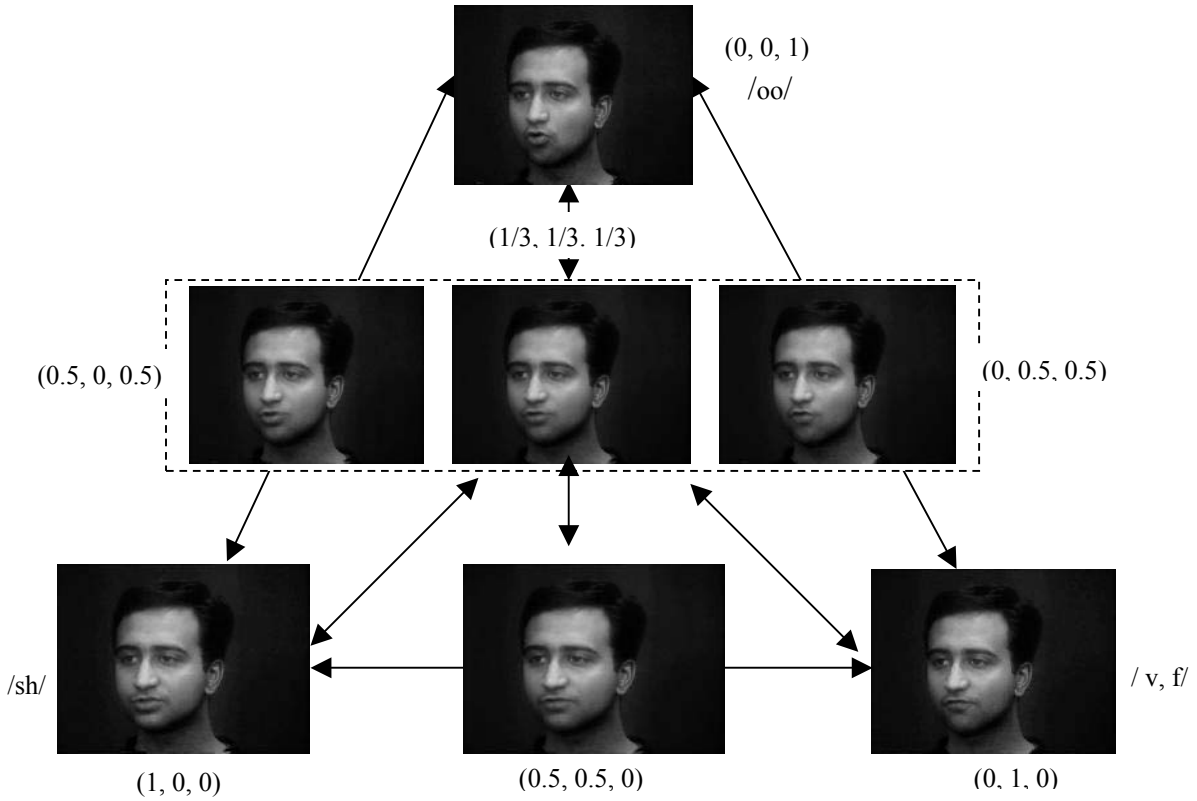

Figure 4: Polymorphing among visemes /sh/, /v/ and /oo/

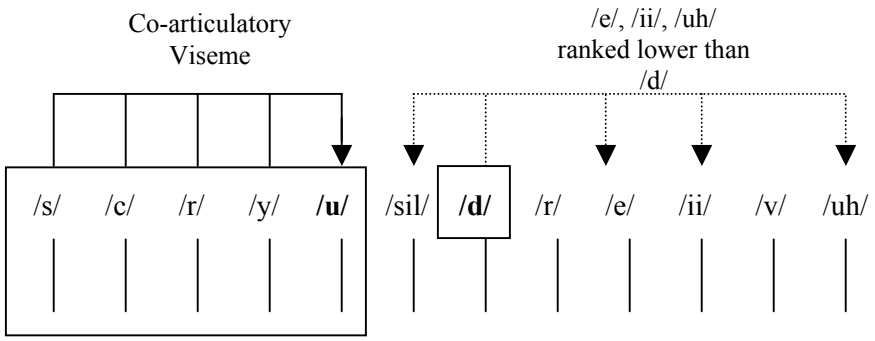

Figure 5: Co-articulation rules applied while enunciating 'screw driver'.

Figure 5 shows the application of co-articulation rule to enunciate 'screw driver'. The first few visemes corresponding to /s/, /c/, /r/, /y/ are affected by the viseme /u/ as /u/ is ranked higher. Further due to presence of /sil/, /d/ and /r/ are not affected by /u/. /e/ and /ii/ do not affect either /d/ or /r/ as /e/ and /ii/ are ranked lower than /d/ and /r/. Again there is no effect on /e/ and /ii/ as both are vowels. Finally /v/ is also not affected since it is least deformable i.e., has the highest rank among all the visemes. Note that the barycentric parameter γ reduces exponentially with increase in distance from the co-articulating viseme[14].

4.3. Temporal Smoothing

Temporal smoothing deals with the timing constraints of the speech. Considering the relaxation and contraction times of mouth shape muscles: if the time between two consecutive phonemes is smaller than the contraction time of the muscles, previous phoneme is influenced by the contraction of the current phoneme. Similarly, for the case when time between consecutive phonemes is smaller than relaxation time, current phoneme will influence the next one.

For the temporal smoothing, our approach is quite straightforward. A threshold time is defined which, in our implementation, is considered as 3 frames. The temporal smoothing occurs only if duration of the viseme transition is less than this threshold, i.e., three frames long (hence considered jerky). So in our scheme, if the duration is less than the threshold, first we look for the duration of the next transition, which if large (2*Threshold), the duration of the current viseme is extended to the threshold and the duration of the following viseme is reduced accordingly. Else, the extent of morph i.e., the contribution of the current viseme transition is reduced linearly. When the duration of current viseme transition is less than the threshold and the duration of next viseme is less than 2*Threshold we limit the extent of morph.

5. Audio Visual 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.

$$T = \Sigma_i \, l(V_{i \, to \, i+1})$$

Where, $l(V_{i\ to\ i+1})$ denotes the duration (in sec) of each viseme transition from V_i to V_{i+1} as computed by Festival and preprocessed by the co-articulation and temporal smoothing module.

Viseme transition streams are then created consisting of two endpoint visemes, the co-articulating viseme and the optical flow correspondence vectors between them. The start index in time of each viseme transition $s(V_{i\ to\ i+1})$ is computed as

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

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

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

The frames between a transition from V_i to V_{i+1} are then synthesized by setting the barycentric parameters for each frame as:

$$\gamma_k = \gamma[V_i] + \frac{s(F_k) - s(V_{i\ \mathrm{to}\ i+1})}{l(V_{i\ \mathrm{to}\ i+1})} \times extent\_of\_morph \times (\gamma[V_{i+1}] - \gamma[V_i])$$

$$\beta_k = \frac{s(F_k) - s(V_{i\ \mathrm{to}\ i+1})}{l(V_{i\ \mathrm{to}\ i+1})} \times extent\_of\_morph \times (1 - \gamma_k)$$

$$\alpha_k = 1 - \beta_k - \gamma_k$$

The morph parameters α_k, β_k and γ_k correspond to the weights given to the start image I_i the final image I_{i+1} and the image I_v corresponding to the co-articulating viseme respectively.

Finally, each frame is generated from the polymorphing between I_i, I_{i+1} and I_v using optical flows between these as warp functions and using α_k, β_k and γ_k as barycentric co-ordinates. The final visual sequence is constructed by concatenating the viseme transitions, played in synchrony with the audio speech signal generated by the TTAVS system. It has been found that lip-sync module produces very good quality synchronization between the audio and the video.

6. Eye and Head 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 is moving. As a step towards making it more video realistic, eye movement and head movement have been incorporated in the facial model.

6.1. Eye Movement

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 eye. 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.

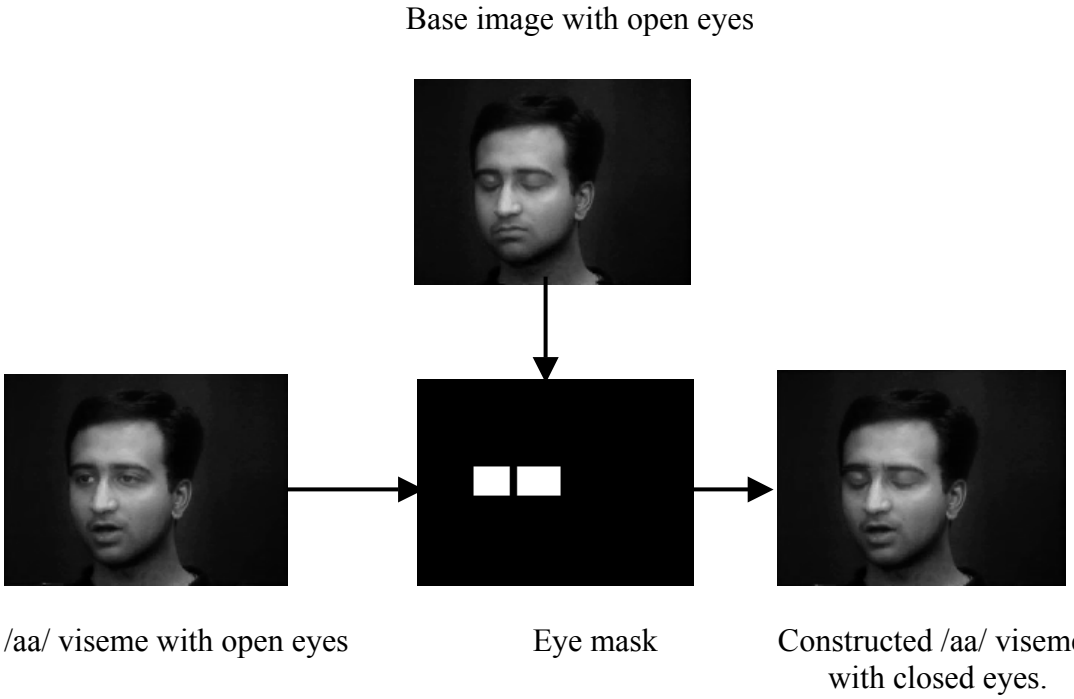

Figure 6: Image Masking

6.2. 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[20] 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.

Figure 7: View morphing between two images of an object taken from the different viewpoints produces intermediate images that preserves 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 and Dyer[20] 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

As shown in figure 7, 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.

7. Integration and Results

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.

Various other audio-visual streams corresponding to sentences like ‘Twenty Two’, ‘Temporary Food Stew’ were generated both with and without co-articulation and temporal smoothing. The audio-visual streams generated with co-articulation and temporal smoothing are much more smooth and realistic than the streams generated using simple morphing.

Some result sequences can be accessed at http://www.cse.iitd.ernet.in/~pkalra/VTalk/

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 and head movement. Efforts have also been made on adding the co-articulation and introducing timing constraint for temporal smoothing.

Next, we plan to work on introducing composition of speech with facial expressions that affect the mouth region to 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.

Acknowledgements

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

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

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

[5] Mathew Brand, *Voice Puppetry*, Proc. SIGGRAPH '99, pp. 21-28, 1999.

[6] C Bregler, M Covell and M Slaney, *Video Rewrite: Driving Visual Speech with Audio*, Proc. SIGGRAPH'97, pp. 353-360, 1997.

[7] Tzong-Jer Yang, I-Chen Lin, Cheng-Sheng Hung, Jian-Feng Huang and Ming Ouhyoung, *Speech Driven Facial Animation*, Proc. of Computer Animation and Simulation Workshop '99, pp. 99-108.

[8] *Talking Heads*, http://www.haskins.yale.edu/haskins/HEADS/contents.html.

[9] Udit Kumar Goyal, Ashish Kapoor and Prem Kalra. *Text-to-Audio Visual Speech Synthesizer*, in Proceedings of Virtual-Worlds 2000, Paris July 5-7, 2000.

[10] Ashish Kapoor , Udit Kumar Goyal, and Prem Kalra. *Modeling Co-articulation for Text-to-Audio Visual Speech Synthesizer*, in Proceedings of ICVGIP 2000, Bangalore.

[11] Seungyong Lee, George Wolberg, Sung Yong Shin. *Polymorph: Morphing along Multiple Images.* In IEEE Computer Graphics and Applications, Jan 1998.

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

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

[14] 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.

[15] A. P. Benguerel, and M.K. Pichora-Fuller, *Coarticulation effects in lipreading*, Journal of Speech and Hearing Research, 25, pp. 600-607, 1982.

[16] Catherine Pelachaud, *Communication and Coarticulation in Facial Animation*, Ph.D. thesis, Department of Computer and Information Science, Univ. of Pennsylvania, Philadelphia, 1991.

[17] P Ekman and W Fresen, *Facial Action coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, CA, 1978.

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

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

[20] Steven M. Seitz and Charles R. Dyer. *View Morphing*, University of Wisconsin, in Proceedings of SIGGRAPH'96, pages 21-30, 1996.