Modeling Co-articulation in Text-to-Audio Visual Speech

Ashish Kapoor, Udit Kumar Goyal and Prem Kalra

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

Abstract

This paper provides our approach to co-articulation for a text-to-audiovisual speech synthesizer (TTAVS), a system for converting the input text to video realistic audio-visual sequence. It is an image-based system modeling the face using a set of images of a human subject. A concatenation of visemes –the corresponding lip shapes for phonemes— can be used for modeling visual speech. However, in actual speech production, there is overlap in the production of syllables and phonemes that are a sequence of discrete units of speech. Due to this overlap, vocal tract motions associated with producing one phonetic segment overlap the motions for producing surrounding segments. This overlap is called **co-articulation**. *The lack of parameterization in the image-based model makes it difficult to use the techniques employed in 3D models for co-articulation. We introduce a method using polymorphing to incorporate co-articulation in our TTAVS. Further, we add temporal smoothing for viseme transitions to avoid jerky animation.*

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 synthesis (TTAVS) 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]. It is not in the scope of this paper to list all the relevant work in this area, however, an exhaustive collection of related work can be found in [8].

In this paper we address a particular problem of co-articulation in a system of text to audio-visual speech synthesizer (TTAVS). Our TTAVS 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]. In this paper we focus our attention to the problem of co-articulation and temporal smoothing to create more visually realistic model for speech.

First we give an overview of our TTAVS. A brief description about the problem of co-articulation and some related work are given next, in Section 3. Then, our approach to model co-articulation is presented. Polymorphing [10] is used as a method for combining the effect of preceding and/or succeeding phonemes. Furthermore a temporal smoothing is added to reduce the jerkiness of the animation. A simple approach for audio-visual synchronization is also presented. Finally, some result sequences are given to demonstrate the effect of co-articulation.

2 Overview of TTAVS

An overview of our TTAVS 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 [11]. 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 entire task of visual speech processing is 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. Sixteen images corresponding to different lip shapes of sixteen different visemes are stored as a database. 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 [12]. 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. 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. A module of co-articulation is added to the earlier design of TTAVS. This module takes visemes sequence and timing information as parameters and provides the parameters to generate the final audio-visual sequence.

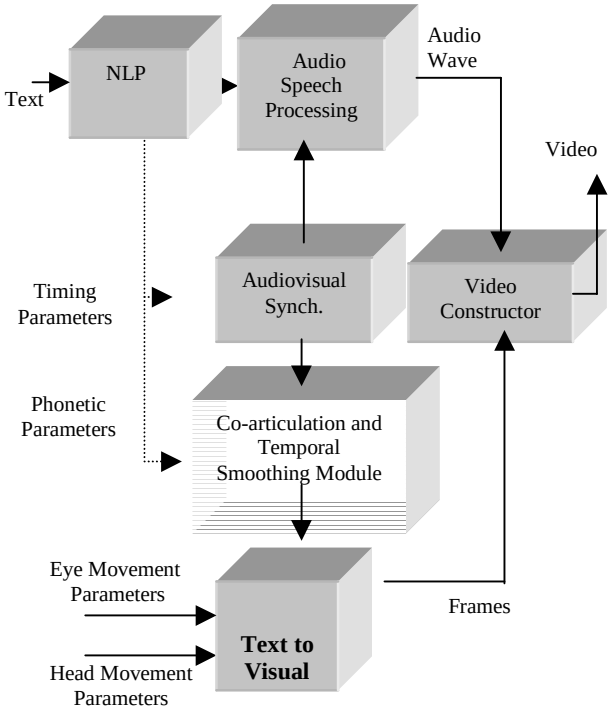

Figure 1. Overview of the TTAVS system

3 Background and Related Work

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 [13]. 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 [14] 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 [15].

Pelachaud [15] 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 upon the speech rate. This can be seen from the fact that a person talking slowly moves her lips much more than a person speaking rapidly. The first step in the algorithm is to apply co-articulation rules to those highly deformable clusters that are context dependent. These rules involve looking ahead to the next highly visible vowel. Then, current viseme is adjusted to make it consistent with the previous and next vowel shapes. Although this set of rules suffices, but no complete set of rules currently exists. Next, the relaxation and contraction times of mouth shape muscles are considered. It checks if each action has time to contract after the previous phoneme or to relax before the next phoneme. 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. Finally, geometric relationships between successive actions are considered. For example, closure of the lips is easier from a slightly parted position than from a puckered position. These actions are obtained by designing a table of weights corresponding to how similar two actions are.

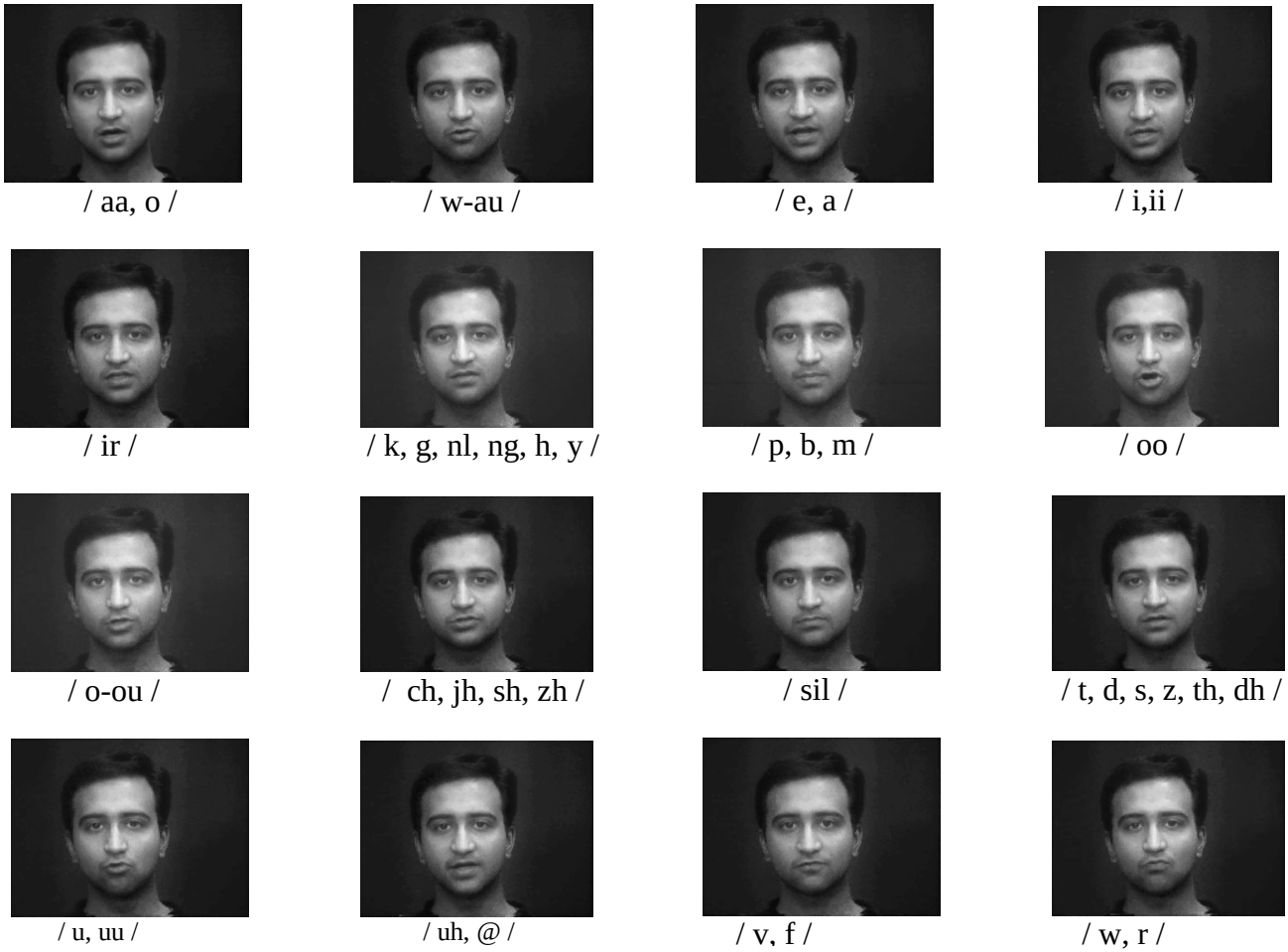

Figure 2. Reduced set of extracted visemes.

Our work in many ways is similar to the idea proposed by Pelachaud [15]. The main difference comes from the fact that we use image-based animation instead of model based approach. In model based approach, FACS [16] 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.

4 Our Approach

Co-articulation can occur in two ways: *forward*, and *backward*. Forward co-articulation occurs when articulation of a speech segment depends on upcoming segments. The speech posture for one phonetic segment is anticipated in the formation of an earlier segment in the phonetic string. For example, while speaking 'to' the lips get curled (as in /u/) when /t/ is being enunciated. Backward co-articulation occurs when a speech segment depends on preceding segments, for example, in 'put', *t* is influenced by *u*. The speech posture for one segment is carried over to a later segment in the phonetic string.

The process of converting text-to-audiovisual stream can be divided into four sub-tasks: viseme extraction, morphing, morphing cancatenation, and audio-visual syncronization [9]. Due to many-to-one mapping between phonemes and visemes, the final reduced set of extracted visemes contains total of 16 visemes as shown in Figure 2. It may be observed that the co-articulation cannot be modeled by simple image morphing between two visemes. As preceding/succeeding viseme affects the vocal tracts, the transition between two visemes also gets affected by other neighboring visemes. Instead of modeling a simple transition between two visemes we need to look into a more complex affair, where this transition is also getting affected by other co-articulating visemes. This is modeled by polymorphing [10] , which 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, hence 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 [17] [18]. 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 [10]. The

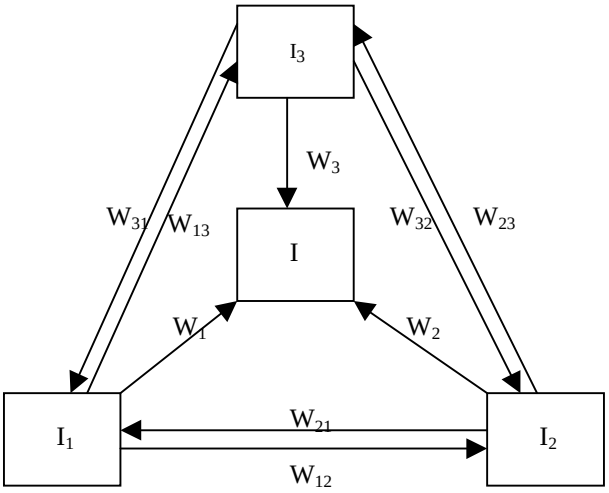

Figure 3. Polymorphing among three different images.

schematic diagram of the process is shown in Figure 3. 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 **b** = (b_1, b_2, …, 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 **b** = (b_1, b_2,…, 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

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, @/, /i, r/
10. /sil/

Figure 4. Ranking of phonemes in increasing order of deformability.

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 [15]. The visemes 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 4. This ranking is based on Pelechaud's work [15] and our practical experience, and we have observed that it holds true for the cases, we have considered.

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=(alpha, beta, gamma). Figure 5 shows an example of polymorphing among 3 different visemes. Alpha, beta and gamma correspond to components of I_1, I_2 and I_v respectively. Gamma[V_1] denotes the barycentric parameter gamma that would be used at the start of transition from V_1 . Further alpha[V_1] and beta[V_1] also denote the values of barycentric parameters alpha and beta at the beginning of transition from V_1. Obviously alpha[V_1] = 1-gamma[V_1] and beta[V_1] = 0.

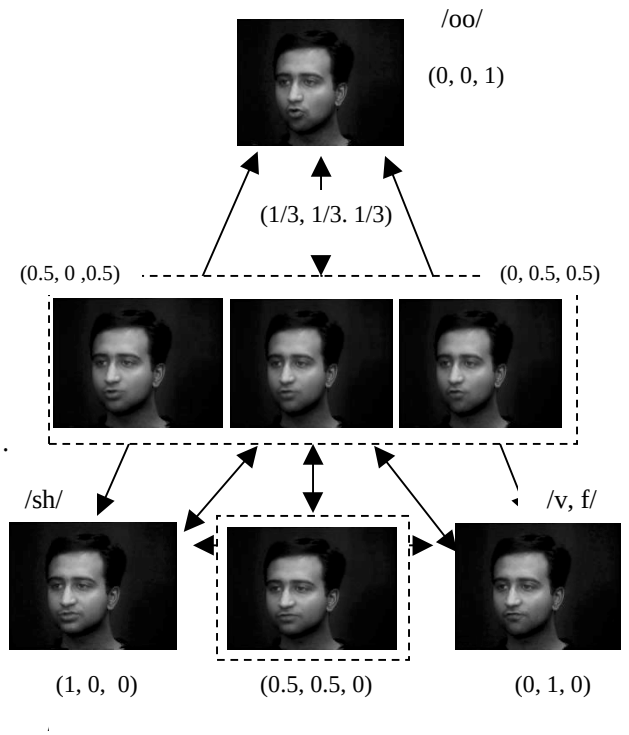

Figure 5. Polymorphing among visemes /sh/, /v/ and /oo/

To generate the co-articulating parameter gamma[V_i] for a viseme V_i in the audio-visual transition sequence we proceed as follows:

For every Viseme V_1 (with rank R_{v1})
If (*V_1 is not a vowel*)
{
Look forward of the viseme (within 5 phoneme transitions) till a vowel/silence is found. Let R_f be the rank of this vowel found at position p_f relative to the current position.
Look backward of the viseme (within 5 phoneme transitions) till a vowel/silence is found. Let R_b be the rank of this vowel found at position p_b relative to the current position.
If ($R_f > R_b$) then
{
If ($R_f > R_{v1}$)
{
gamma[V_1]= K*exp(-p_f/5)
Here V_1 is affected by the forward viseme of higher rank.
}
}
else
{
If ($R_b > R_{v1}$)
{
gamma[V_1] = K*exp(-p_b/5)
Here V_1 is affected by the backward viseme of higher rank.
}
}
}
End For

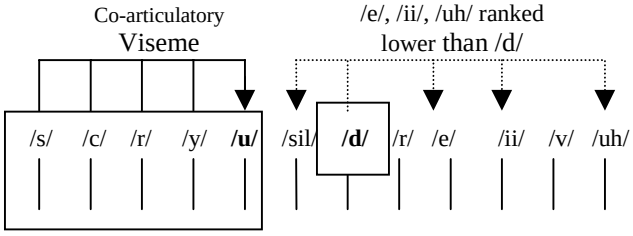

Figure 6. Co-articulation rules applied while enunciating 'screw driver'.

Figure 6. 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 gamma reduces exponentially with increase in distance from the co-articulating viseme [13]. The constant K is found experimentally and in our example K=0.5.

5 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 of the current viseme transition is reduced linearly. The whole procedure can be described as follows:

For every Viseme Transition from V_1 to V_2 do

```
If (V2= 'p' or 'b') then
        extent_of_morph_v1-v2=1.0;
else
        if (duration of  transition<THRESHOLD TIME)
        {
           if (duration of next transition > =
                                    2*THRESHOLD TIME)
                    {
                      Duration of current transition is
                      increased to the threshold and
                      duration of the next transition is
                      reduced accordingly;
                     }
                else
                   {
                     extent_of_morph_v1-v2=
                         (duration/THRESHOLD TIME)
                   }
        }
End For
```

Note that if second viseme V_2 in transition is 'p' or 'b' then transition is forced to completion, as while speaking /p/ or /b/ the lips should be finally closed hence we can't do away with reducing the extent of morph. Further the parameter extent_of_morph$_{v1\text{-}v2}$ is used for determining the barycentric parameters alpha and beta as explained in the next section.

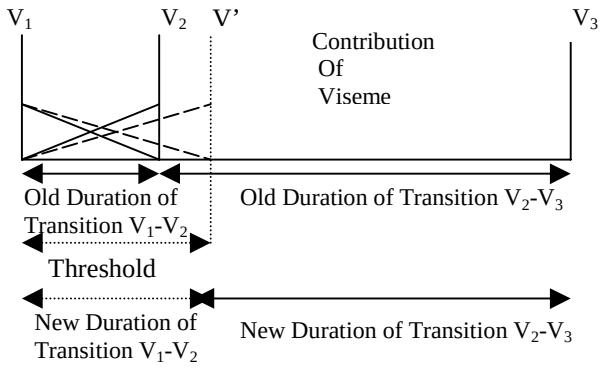

Figure 7. Temporal smoothing when the duration of next viseme >= 2*Threshold

Figure 7 shows the case when the current viseme transition duration is less than threshold and the duration of next viseme is greater than 2*Threshold. In this case the duration of the current viseme transition is increased to the threshold and the duration of the next viseme transition is reduced accordingly. Note that the slope of the lines in corresponding to new viseme duration is less than the slope corresponding to the earlier duration hence causes smoothening of the animation.

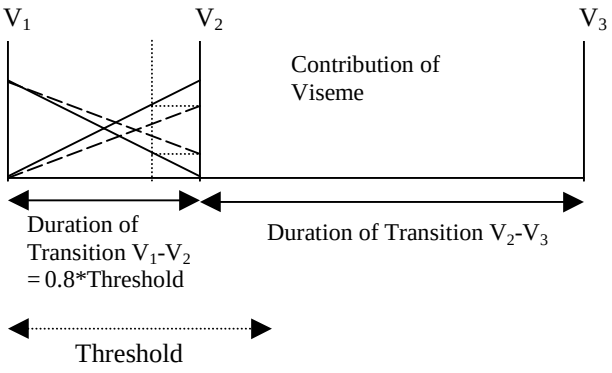

Figure 8. Temporal smoothing when the duration of next viseme < 2*Threshold.

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. For example in Figure 8, the duration of current viseme transition is 0.8 times the Threshold and also the duration of the next viseme transition is less than twice the Threshold, so the extent of morph is limited by the factor 0.8. That means that only 80% of morphing will be done but in the same duration as of the current transition. It can be observed that slopes of the lines for the temporal smoothened case is less than the earlier one.

6 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.

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

Where, $l(V_{i \text{ to } i+1})$ 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 $s(V_{i \text{ to } i+1})$ is computed as

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

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

Let $t = \frac{s(F_k) - s(V_{i \text{ to } i+1})}{l(V_{i \text{ to } i+1})} * \text{extent\_of\_morph}_{Vi\text{-}Vi+1}$

gamma = gamma[V_i] + t * (gamma[V_{i+1}] - gamma[V_i])

beta = t *(1.0 – gamma)

alpha = 1.0 – beta – gamma

The morph parameters alpha, beta and gamma 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 using the polymorphing between I_i, I_{i+1} and I_v using optical flows between these as warp functions and using alpha, beta and gamma 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.

7 Results

Figure 9 shows a sequence of frames generated without co-articulation and temporal smoothing for speaking the word ‘stew’. While enunciating the phonemes /s/ /t/ the lips should also be curled due to the presence of /u/. Figure 10 shows another sequence generated using our method of co-articulation. A careful observation reveals that the lip shapes in this sequence actually get curled while speaking /s/ and /t/ due to co-articulation thus render more realistic and smooth animation.

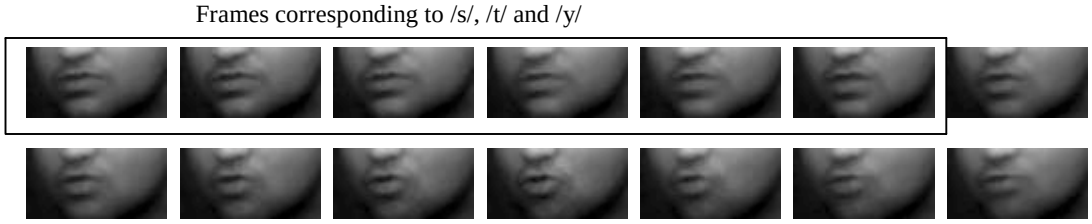

Figure 9. Frames generated without co-articulation for speaking ‘STEW’.

Figure 10. Frames generated with co-articulation for speaking ‘STEW’.

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. These results can be accessed at **www.cse.iitd.ernet.in/~pkalra/ICVGIP2000.**

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 by modeling co-articulation and introducing timing constraint for temporal smoothing.

The work can be extended to introduce smoothing for spatial constraints in the facial model [15]. The spatial smoothing needs to be incorporated due the difference in viseme shapes. 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.

Acknowledgement

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, pp. 837-852, 1998.

[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*, pp. 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, 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, Ming Ouhyoung, "Speech Driven Facial Animation," *Proc. of Computer Animation and Simulation Eurographics 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, pp. 256-269.

[10] Seungyong Lee, George Wolberg, Sung Yong Shin. "Polymorph: Morphing along Multiple Image," *IEEE Computer Graphics and Applications*, Vol. 18, No. 1, pp. 58-71, Jan 1998.

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

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

[13] 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.

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

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

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

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

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