ANIMATING EXPRESSIVE FACES TO SPEAK IN INDIAN LANGUAGES

Tanveer A. Faruquie, Chalapathy Neti, Nitendra Rajput, L. Venkata Subramaniam, Ashish Verma*

IBM India Research Lab, New Delhi, India.
*IBM T. J Watson Research Center, White Plains, NY, USA.

ABSTRACT

This paper describes a morphing based automated audio driven facial animation system. A novel scheme to implement a language independent system for audio-driven facial animation given a speech recognition system for just one language, in our case, English, is presented. New viseme and expression combinations are synthesized to be able to generate animations with new facial expressions.

1. INTRODUCTION

A talking face, with lip movements in synchronization with the spoken words along with correct expressions, greatly enhances communication. Many methods have been presented to animate faces in sync with audio [2][5]. These methods rely on a viseme based alignment being generated from the incoming audio, where visemes are different, distinguishable lip shapes [6, pp. 394-395]. For this a speech recognition system is used to generate the phonetic alignment from the incoming audio. Phonetic alignment refers to the time duration and the transition times between phonemes in an audio sequence [1]. A phoneme to viseme mapping then generates the visemic alignment from the phonetic alignment. Given visemes with two or more different facial expressions the remaining visemes in these facial expressions are synthesized [5] so that all viseme and expression combinations become available. Transition frames between visemes are generated using optical flows to get an animated sequence.

Audio driven facial animation requires training of a speech recognition system which is used for generating alignments from the input speech. Once the phonetic alignment is generated, the mapping and the animation hardly have any language dependency in them. Translingual visual speech synthesis is achieved if the first step of alignment generation is made language independent. Thus, given a speech recognition system for one language it is possible to synthesize video with speech of any other language [4]. In particular we look at generating alignments for Hindi speech given an English speech recognizer.

2. SYSTEM MODEL

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

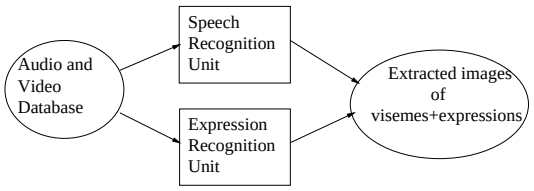

Figure 1: Extraction Module

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. A short sentence like "The sharp quick brown fox jumped over the lazy dog." captures all the 12 visemes.

In the background processing module, shown

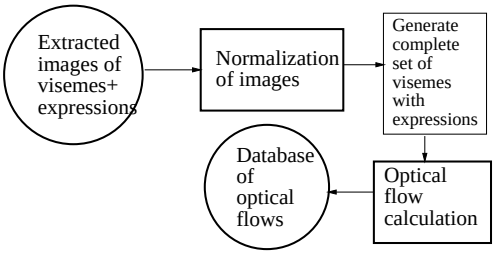

Figure 2: Background Processing Module

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

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, $n = 1, \ldots, 7$, are the seven basic expressions and, $m = 1, 2, \ldots, 12$, the viseme set. Optical flows between different visemes within an expression and between the expressions are computed and stored.

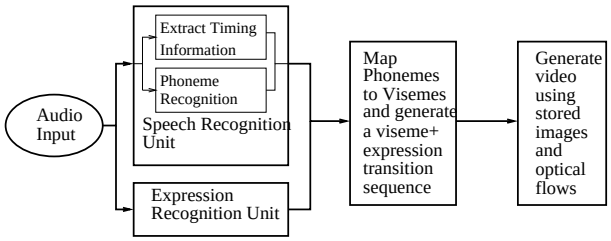

Figure 3: Synthesis Module

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 described in the following section. The timing information and phoneme transition can also be extracted for a novel language whose speech recognition engine is not available as described in the following section. 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 TO VISUAL MAPPING

As the audio stream comes in, a speech recognition system is used to give the phoneme transitions. The speech recognition system also provides the timing information, i.e. the duration of each phoneme and the duration of the transitions. Each phoneme is mapped to its corresponding viseme. For the facial animation system, it is not necessary to know the exact word being spoken or even the exact phoneme from a word recognition point of view, it is sufficient to know the viseme. Our system uses a set of 12 visemes and the mapping from phonemes to visemes is shown in Table I.

3.1. Translingual Mapping

If the language in which the video is to be synthesized is a new language, then the phoneme set of the new language may be different from that of the training language. But the alignment generation system gives the alignments based on the best phone boundaries using its own set of phonemes (corresponding to the language used in the training). Therefore, a mapping is required to convert the phonemes from one language to the visemes of the other language so as to get an effective alignment.

In this section an approach to synthesizing visual speech from a given audio signal in any language, with the help of a speech recognition system in another language, is presented. From here onwards, we refer to the language used in training

the speech recognition system as the *base language* and the language in which the video is to be synthesized as the *novel language*. In the illustrations, Hindi has been chosen as the novel language and English as the base language.

3.1.1. Phonetic Vocabulary Adaptation Layer

Since, the recognition system is trained on the phoneme set of the base language, the vocabulary needs to be modified so that the words from the novel language now represent the baseforms in the base language phoneme set. Such a modification is made possible by the Phonetic Vocabulary Adaptation Layer. This layer works by using a mapping from the phoneme set of one language to the other language. For illustration, a mapping from the Hindi phones to the English phones is as shown in Figure 4. As is seen, not all the English phonemes are used by the novel language. Also there exists an exact mapping for a large number of phones. These are shown by a *** sign on that row. A ** in the row implies that the mapping is not exact but that it is the acoustically closest map. A * in the mapping implies that the novel language phoneme has been approximated by a string of more than one phoneme from the English language for acoustic similarity.

3.1.2. Visemic Vocabulary Adaptation Layer

An additional vocabulary which represents the words of the novel language in the phoneme set of the base language is created but this does not use the mapping in Figure 4, it uses a mapping based on the visemic similarity of the two phonemes. We call this mapping based on visemic similarity the visemic vocabulary modification layer. Using this additional vocabulary, the base language alignments and the base language phoneme to viseme mapping, we get the visemic alignments. This visemic alignment is used to generate the animated video sequence. Alternately, if the viseme set is available for the novel language, then the visemic vocabulary modification layer can be modified to directly give the visemic alignment using the phoneme to viseme mapping in the novel language.

अ	AX ***	छ	CH **	ब	B ***
आ	AA ***	ज	JH ***	भ	B **
इ	IX ***	झ	J **	म	M ***
ई	IY ***	ञ	N G *	य	Y ***
उ	UH ***	ट	T ***	र	R ***
ऊ	UW ***	ठ	T **	ल	L ***
ए	EY ***	ड	D ***	व	V ***
ऐ	AE ***	ढ	D **	श	SH ***
ओ	OW ***	ण	N **	ष	SH ***
औ	AU ***	त	DH **	स	S ***
क	K ***	थ	TH ***	ह	HH ***
ख	K **	द	DH ***	क्ष	K SH *
ग	G ***	ध	DH **	त्र	T R AX *
घ	G **	न	N ***	ज्ञ	G Y *
ङ	N **	प	P ***		
च	CH ***	फ	F ***		

Figure 4: Phoneme mapping from English to Hindi

Figure 5 shows the block diagram of the modification layers described above to achieve translingual visual speech synthesis. In the figure the subscripts B and N refer to the base language and the novel language respectively. The superscripts P and V refer to phonemes and visemes respectively. The speech recognition system is modified to generate visemic alignments corresponding to the novel language using the phonetic and visemic vocabulary modifiers. In case the visemes for the novel language are available the visemic vocabulary modifier is not required and a direct phoneme to viseme mapping in the novel language may be used to give visemic alignments.

4. 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 viseme v_1 with facial expression e_2, how to generate viseme v_2

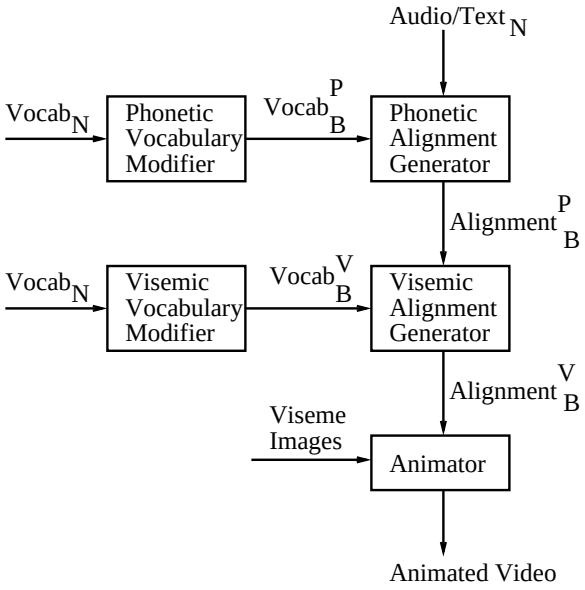

Figure 5: Block diagram showing the modification layers

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 6). 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 $flow_{new}$.

To introduce the new features that appear in viseme v_2 (see Figure 7), 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).

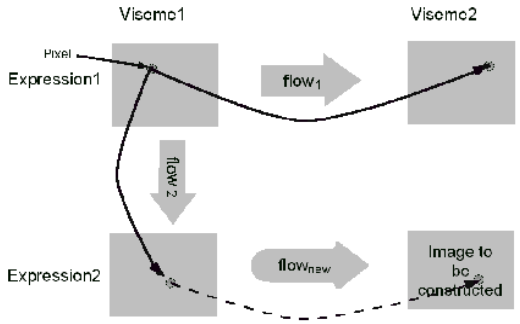

Figure 6: New Viseme-Expression Pair Generation

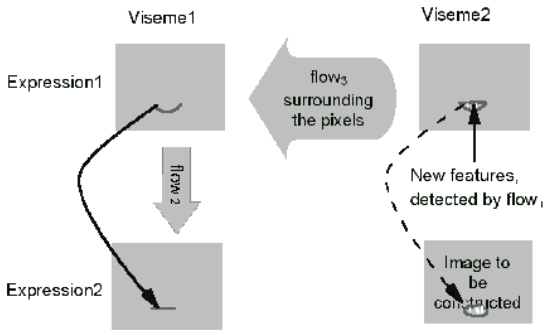

Figure 7: Introducing New Features

To suppress the facial features disappearing in viseme v_2 (see Figure 8), 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 9 and Figure 10 are examples of new viseme+expression combinations generated from the existing ones.

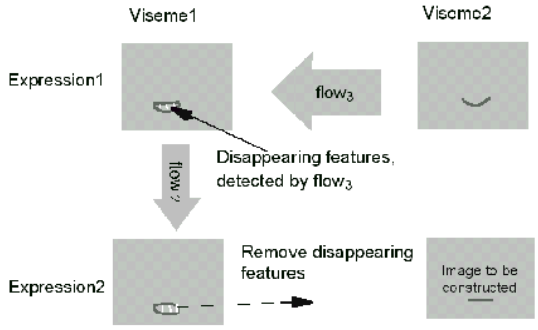

Figure 8: Suppressing Disappearing Features

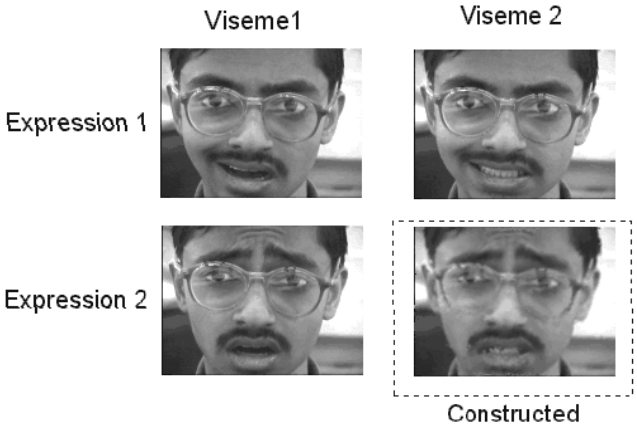

Figure 9: Existing Images and the Constructed Image with New Features Appearing

5. SYSTEM EVALUATION AND CONCLUSIONS

An automated system for creating an additional channel for communication is presented. 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 appear. Using the method described here animations were generated in Hindi and Telugu. Given a speech recognition system for one language, one can easily and quickly customize the alignment layers to get a synchronized video in any other language. A text-to-video synthesizer can also be built by using the phonetic alignments generated by the text to speech synthesizer.

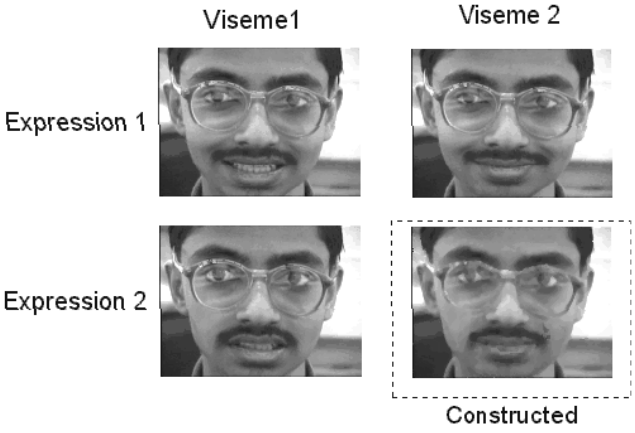

Figure 10: Existing Images and the Constructed Image with Disappearing Features

REFERENCES

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer, "Speech Recognition with continuous parameter hidden markov models," *Proceedings ICASSP-88*, New York, May 1988, pp. 40-43.

[2] 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.

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

[4] T. A. Faruquie, A. Kapoor, R. Kate, N. Rajput, L. V. Subramaniam and A. Verma, "Audio driven facial animation for audio-visual reality," *IEEE International conference on Multimedia and Exposition*, Tokyo, Japan, 22-25 Aug, 2001.

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