Animating Expressive Faces Across Languages

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

***Abstract*—This paper describes a morphing-based audio driven facial animation system. Based on an incoming audio stream, a face image is animated with full lip synchronization and synthesized expressions. 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. The method presented here can also be used for text to audio-visual speech synthesis. Visemes in new expressions are synthesized to be able to generate animations with different facial expressions. An animation sequence using optical flow between visemes is constructed, given an incoming audio stream and still pictures of a face representing different visemes. The presented techniques give improved lip synchronization and naturalness to the animated video.**

***Index Terms*—Audio to video mapping, facial animation, facial expression synthesis, lip synchronization, translingual visual speech synthesis.**

I. Introduction

HUMANS communicate verbally using words and sentences. Humans also communicate nonverbally using expressions, gestures and prosody. Faces and voices hold a special significance for us. Humans recognize faces and voices and infer a lot of things from these [14]. Apart from identity, faces and voices convey temper, humor and various moods and emotions. We also infer attributes like personality and character from the faces and voices. The design and implementation of computer systems that can cover the whole range of human-human like interaction by using faces and voices is one of the challenging objectives of Computer Human Interaction research.

Efforts have been made to train the computer to recognize and interpret the various modes of communication used by humans, like speech, gestures, gaze, expressions etc. It does a multimodal processing of these signals to gain an understanding of what the human is communicating [14], [18]. In this work, we seek to cover some aspects of the computer's response in this interaction. In particular, we look at the derivation of visual information from speech to create audio-visual reality. We seek to animate a person's face speaking in synchronism with the incoming audio and show that it is also possible to synthesize various expressions on this face while it is speaking. Realism is obtained by automatically extracting relevant still images from a previously available footage of this person, modifying these images and using them to synthesize new images and then using these new images to animate in sync with incoming audio in possibly any language. The whole process it is shown, can be largely automated. Human intervention may be required when dealing with expressions, though this may also be automated to some extent.

Animation methods rely on a visemic alignment being generated from the incoming audio [17] or from synthetic speech [21]. Visemic alignment refers to the time duration and the transition times between visemes, which are different, distinguishable lip shapes [13, pp. 394–395], in a face animation sequence. 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 [2]. A phoneme to viseme mapping is used to derive the visemic alignment from the phonetic alignment. Once the phonetic alignment is generated, the mapping and the animation hardly have any language dependency in them. Translingual visual speech synthesis can be achieved if the first step of phonetic alignment generation is made language independent. Thus, we show that given a speech recognition system for one language it is possible to synthesize video with speech of any other language. In particular we look at generating alignments for Hindi speech given an English speech recognizer.

Given an incoming audio stream and pictures of a face representing different visemes an animation sequence is constructed. We record 12 face images corresponding to the 12 visemes. These viseme images are aligned, for head movement compensation, and optical flows for transition in-between these visemes are computed (12×11 total) and stored. At the time of synthesis, for an incoming audio stream, using speech recognition and phoneme to viseme mapping, the corresponding video frame is identified and transition frames between visemes are generated using stored optical flows. The phoneme to viseme mapping scheme described in this paper allows us to handle languages other than the one for which a speech recognition system is available. Morphing-based animation has been considered in the past [5], [11], [17]. In this paper, we seek to extend this to include animation with expression. For a richer scope of animation, it is necessary to be able to animate the face with appropriate expressions. In our system, given a single viseme in a new facial expression, a method is presented that can generate the remaining visemes in this new expression. In earlier work [5], the generation of novel images from a single example image was considered by learning the desired transformation from prototypical examples. However appearing and disappearing features

Manuscript received December 5, 2001; revised January 21, 2003. This work was presented in part at the IEEE International Conference on Multimedia and Expo 2000, New York, July–August 2000, and at the IEEE International Conference on Multimedia and Expo 2001, Tokyo, Japan, August 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Radu Serban Jasinschi.

A. Verma, L. V. Subramaniam, N. Rajput and T. A. Faruquie are with the IBM India Research Lab, Indian Institute of Technology, Hauz Khas, New Delhi 110 016, India (e-mail: vermar@in.ibm.com; lvsubram@in.ibm.com; rnitendr@in.ibm.com; ftanveer@in.ibm.com).

C. Neti is with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA (e-mail: cneti@us.ibm.com).

Digital Object Identifier 10.1109/TMM.2004.837256

like teeth were not handled. In [8], it is assumed that there exists a video database of the head to be synthesized, wherein, the subject is present in the expression to be synthesized at least once. In [12], the generation of intermediate combinations of visemes and facial expressions from extreme ones is considered. Generating new viseme-expression pairs was not considered there. In this paper, we are considering the generation of new viseme-expression combinations while taking care of appearing and disappearing features. The seven basic expressions considered are neutral, surprise, fear, disgust, anger, happiness, and sadness.

Various application scenarios motivate audio driven facial animation. These include bandwidth reduction for video teleconferencing, movie dubbing, user-interface agents and avatars, and multimedia telephones for hard of hearing people. Simple experiments have shown the value of the visual channel in speech comprehension [14], for example the McGurk effect. In many scenarios, it is possible that the listener is in a crowded and noisy environment. Vision adds redundancy to the signal and provides evidence of those cues that would be irreversibly masked by noise or hearing impairment [13], [14].

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

- *Visual e-mail:* At the receiving end the email is "read out" by the sender. The receiver mailbox activates the correct person, to read out the mail, by matching the address.
- *Newscast:* In many cases involving a field reporter, the audio is available but due to various reasons, the corresponding video is not available. Usually a photograph of the person is shown on the TV screen along with the audio. Using the system presented here, a video of the person speaking can be generated and shown along with the audio (see Fig. 13). Vision directs the listener's attention and sustains interest.
- *Entertainment:* Making people say things they normally would not. For example popular actors are made to say different things and "interact" with people.
- *User Interface Agents:* Call centers providing on-line problem solving and assistance can enhance their effectiveness if a visual rendering is done on the user's computer screen. The animated face can be that of the office IT services person known to the users in the office. Familiarity helps in immediately engaging the user and holding his or her attention even if it is an automated call center.

Many other uses of this system can be thought of. A talking face has the advantage of directing the listener's attention and sustaining interest. An audio-visual reality is created if the animated face is able to hold human attention and successfully engage the person in useful conversation or task.

This paper is organized as follows. In Section II, we present the audio-driven facial animation system model. In Section III, the audio to visual mapping rule is presented. There we describe, in detail, the method used to adapt the speech recognition system of one language to generate phonetic and visemic alignments in a new language. The specific case of adding Hindi words to an English speech recognition system is considered. In Section IV, the method of generating the animation is discussed. The normalization of viseme images, facial expression synthesis, and lip synchronization with audio are detailed in this section. In Section V, we present an evaluation of the system. Finally, Conclusions are presented in Section VI.

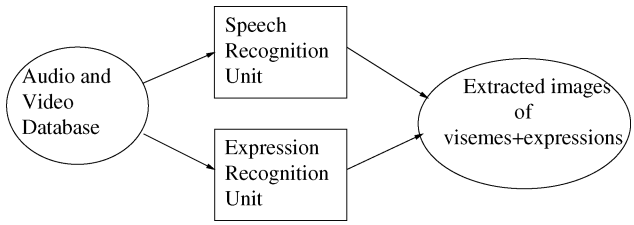


Fig. 1. Extraction module.

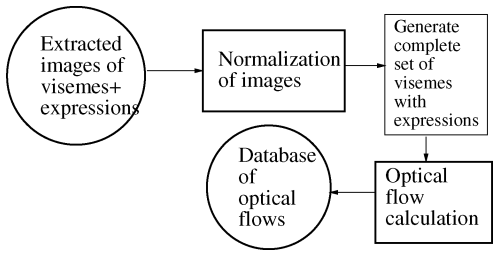


Fig. 2. Background processing module.

II. System Model

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

Fig. 1 shows the viseme extraction module. For an incoming stream of synchronized audio+video, we first recognize the phoneme from the audio and then map this phoneme to its corresponding viseme and select the corresponding video frame to represent this viseme. The expression recognition unit can be either audio based [7], [19] or video based [10]. We have found a 15-s video to be sufficient in most cases for obtaining all the viseme frames in neutral expression. A short sentence like "The sharp quick brown fox jumped over the lazy dog" captures all the 12 visemes. We need to capture the 12 viseme frames in each expression to animate sentences in that expression. Therefore, for each expression we try to capture as many of the 12 visemes as possible from existing video.

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

Fig. 3 shows the synthesis module. From an incoming audio stream, timing information, phoneme transitions and expressions are extracted. The timing information and phoneme transition can also be extracted for a novel language whose speech recognition engine is not available as described in Section III. The phonemes are then mapped to the corresponding visemes. This mapping is described in the following section. The expression recognition unit based on audio gives the correct expres-

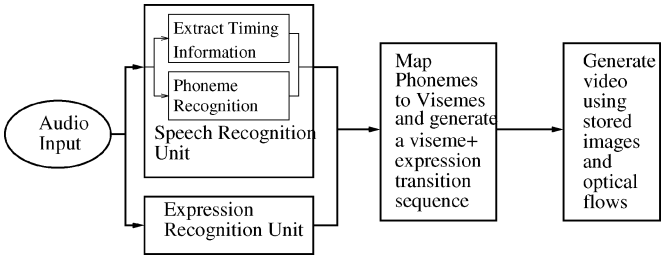


Fig. 3. Synthesis module.

sion. However, in our case the expression maps have been explicitly provided. Together, the viseme+expression combination determines the image 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.

III. Audio to Visual Mapping

By giving a full audio-visual output from an audio only input, the communication modality is being changed and enhanced while preserving the conveyed information. In this section we describe the exact modality for doing a mapping from audio to the visual space.

For an input audio stream, a speech recognition system is used to generate 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 in the input audio. Each phoneme is mapped to its corresponding viseme. In the English language, there are about 50–60 phonemes and these are mapped onto a set of visemes. For the facial animation system, it is not necessary to recognize 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 these visemes is shown in Table I.

TABLE I
Phoneme to Viseme Mapping Rule

Phoneme	Viseme
$AA, AE, AH, AO, AW, AX, AXR, AY, HH$	$Viseme1$
EH, ER, EY, IH, IX, IY	$Viseme2$
L	$Viseme3$
R	$Viseme4$
OW, OY, UH, UW, W	$Viseme5$
B, BD, M, P, PD	$Viseme6$
$D, DD, DH, DX, G, GD, K, KD, N, NG, T, TD, TS, X, Y$	$Viseme7$
F	$Viseme8$
JH, S, Z, ZH	$Viseme9$
CH, SH	$Viseme10$
TH	$Viseme11$
$D\$$	$Viseme12$

A. Translingual Mapping

Audio driven facial animation has typically used a speech recognition engine in the language to be animated. Building a speech recognition system is data intensive and is a very tedious and time consuming task [3]. We present a translingual speech synthesis system to implement a language independent system for audio-driven facial animation given a speech recognition system in one language, in our case, English. In order to understand the proposed translingual visual speech synthesis system, we enumerate the crucial points in translingual visual speech synthesis as follows.

1) From the given input audio and the transcribed truth, we generate the phonetic alignment. This requires a speech alignment system which could understand the phonetic baseforms of the text. This would work fine if the input audio is in the same language as the language used for training the recognition system.

2) 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 language for which the recognition system is built. But the alignment generation system generates 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 phonetic vocabulary of one language to a vocabulary using the phonemes of the other language to get an effective alignment in the phone set of the new language.

3) A phoneme to viseme mapping such as in Table I can then be used to get the corresponding visemic alignment which generates the sequence of visemes and their time durations which are to be animated to get the desired video.

4) Animating the sequence of viseme images to get the desired video output aligned with the input audio signals can now be done, as described in Section IV.

A new 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. If a novel language word is presented to the alignment generator, then the alignment generator will not be able to generate the alignments for such a word as the word is not in the phonetic vocabulary of the base language training system. Moreover the phonetic spelling of a word in the novel language may not be represented completely by the phone set of the base language. We present below a technique to overcome these problems resulting in a language independent alignment generation system. We build a system that will have the alignment generator and the viseme images for the base language that can be used to generate the animation for audio input in any language.

1) Phonetic Vocabulary Adaptation Layer: The base language vocabulary does not include words from the novel language. Hence, when a word from the novel language is presented to the speech recognition system that has been trained in the base language, it will fail to give the phonetic baseforms of the word. In order to generate alignments for words in the novel language, first a phonetic vocabulary of this language is created wherein words are represented in the phonetic baseforms using the phone set of the novel language. Since the recognition system is trained on the phone 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 phone set. Such a modification is made possible by the Phonetic Vocabulary Adaptation Layer. This layer works by using a mapping from the phone set of one language to the other language. For illustration, a mapping from the Hindi character to the English phones is as shown in Table II. There are three possible cases.

1) The word in the novel language can be represented by the phonemes in the base language; for such words, the baseforms can be simply written using the base language phone set.
2) The word in the novel language cannot be represented by the base language phone set; then the word is written using the best approximation achieved through the above mapping.
3) A phoneme in the base language never appears in the novel language; in such a case, that particular phoneme in the base language is redundant and is left as it is.

Since the aim of mapping the phone set is to generate the best phoneme boundaries through acoustic alignment, the mapping is based on acoustically-similar phonemes, i.e., if there is no phoneme in the base language which can be associated with the phoneme in the novel language, then that base language phoneme is chosen which is acoustically closest. Both, however, may map to a different viseme. This problem is addressed in the next subsection.

The phonetic vocabulary adaptation layer helps in generating the base language alignments for the novel language audio. In Table II, an example of mapping the phones of Hindi language to the English language phone set is presented. As 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.

Next, we show how to extract the base language visemic alignments for animation in the novel language.

2) Visemic Vocabulary Adaptation Layer: Since the system has to work for any novel language using the alignment generator and the viseme set in the base language, visemic alignment cannot be simply generated from the phonetic alignment using direct phoneme to viseme mapping. As was shown above, the phonetic vocabulary modification layer was built on the mapping based on acoustically similar phonemes. However, this mapping may distort the visemic alignment as it does not take into consideration the visemes corresponding to each such phoneme. So 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 Table II, it uses a mapping based on the visual similarity of the two phonemes. In Table III, we show this mapping based on visual similarity for those Hindi phones that do not map to their acoustically similar phones. The other Hindi phones use the mapping of Table II. 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. As can be seen in Table II, the phoneme mapping between the two languages is not one-to-one. So a single phone in the base language may represent more than one phone in the novel language. This however creates no confusion as the Phonetic Vocabulary Modification Layer outputs the alignment in the novel language after taking into account the many-to-one mapping.

Alternately, if the viseme set images are 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. Here the

TABLE II
ACOUSTICALLY CLOSEST MAPPING FROM HINDI TO ENGLISH

Hindi Phone	Hindi Alph	English Phone	Hindi Phone (y)	Hindi Alph	English Phone
AA	आ	AA ***	JH	ज	JH ***
AAN	आं	AA **	JHH	झ	JH **
AE	ॅ	AE ***	K	क	K ***
AEN	ॅं	AE **	KD	क्	KD ***
AW	ॉ	AW ***	KH	ख	KD HH *
AWN	ॉं	AW **	L	ल	L ***
AX	अ	AX ***	M	म	M ***
AXN	अं	AX **	N	न	N ***
B	ब	B ***	NG	ङ	NG ***
BD	ब्	BD **	OW	ो	OW ***
BH	भ	BD HH *	OWN	ोॅ	OW **
CH	च	CH ***	P	प	P ***
CHH	छ	CH **	PD	प्	PD ***
D	ड	D ***	PH	फ	PD HH *
DD	ड्	DD ***	R	र	R ***
DDN	ड	DD **	S	स	S ***
DH	द	DH ***	SH	श	SH ***
DHH	ध	DH **	T	ट	T ***
DN	ण	DX **	TH	थ	TH ***
DXX	ड़	DX **	THH	ठ	TH **
D$	SIL	D$ ***	TX	त	TH **
EY	ए	EY ***	UH	ु	UH ***
EYN	ॅ	EY **	UHN	ठ	UH **
F	़फ	F ***	UW	ू	UW ***
G	ग	G ***	UWN	ूं	UW **
GH	घ	GD HH *	V	व	V ***
HH	ह	HH ***	X	SIL	X ***
IH	ि	IH ***	Y	य	Y ***
IY	ी	IY ***	Z	ज	Z ***
IYN	ीं	IY **			

TABLE III
VISUALLY CLOSEST MAPPING FROM HINDI TO ENGLISH

Hindi Phone	Hindi Alph	Visually closest English Phone
CHH	छ	SH
EYN	ॅ	IY

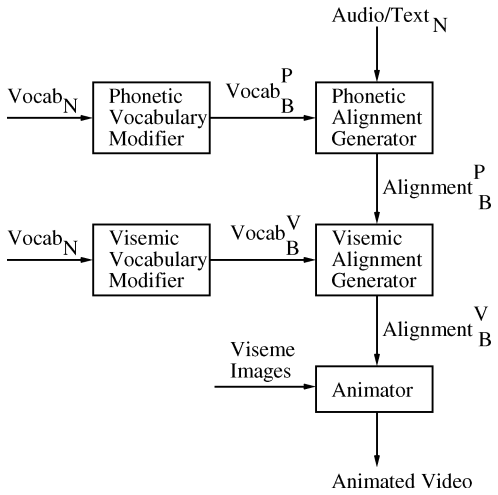


Fig. 4. Block diagram showing the modification layers.

phonetic alignment generated in the base language is converted to the novel language by comparing the phonetic spelling of each word in the two vocabularies. This comparison is used to map the base language phonemes in the generated alignment to corresponding phones in the novel language. Then the phoneme to viseme mapping of the novel language is applied. Note that the visemic alignment so generated is in the novel language and this was desired as the viseme images are available in that language and not in the base language. If the viseme set of the novel language is very different from the viseme set of the base language, then this modified system would be especially useful.

Fig. 4 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 phonetic and visemic vocabulary modifiers are appended to the speech recognition system to generate the visemic alignments corresponding to the novel language. In case the viseme images 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.

The system uses the generated visemic alignment for the purpose of animation. For animation, morphing is done from one viseme image to another as given by the visemic alignments. Due to nonaccurate mapping of phonemes, the alignment may not represent the exact phone boundaries. However, this is not observed in the animated video as a viseme is always in transition during these boundaries. A smooth and continuous video is thus generated which does not reflect any inaccurate phoneme boundaries. Also in Section VI, we show that visemic classification results are much better than phonetic classification results. The method presented here can also be used for text to audio-visual speech synthesis. Text to speech synthesizers work by producing a phonetic alignment of the text to be pronounced and then by generating smooth transitions in between adjacent phones to get the desired sentence [9]. Using phoneme-to-viseme mapping and text-to-speech synthesis, a text-to-video synthesizer can be built.

It is also possible to train classifiers based on visemic classes. In [4], we have proposed an HMM model that uses viseme-based training models. The audio data is grouped into a smaller number of visually distinct visemes rather than the larger number of phonemes. Such a classifier can also be trained with multilingual input. However, it is not clear how it will compare with the translingual system presented here and remains a scope for future work.

IV. Audio–Visual Animation

In this section, we show how the animation is achieved. We also present a scheme to synthesize new expressions to achieve realism. The voice of the speaker, in the form of an audio stream is left unaltered. Hence the spoken words and the prosody of the speaker is conveyed as is. In addition the system animates the face of the speaker to make it speak in sync with the audio and to generate expressions in sync with the prosody in speech. An additional channel for communication is thus created.

A. Normalization of Images

The system waits for the first occurrence of a viseme+expression combination and extracts all possible combinations from the audio+video footage. The images so obtained may not be aligned. If these images are used for animation, then the resulting sequence will have disturbing and unintended head motions. We, therefore, need to align the images. We use a method similar to [6] to normalize the images. There are two components of motion between the images, 3-D rigid body motion and non rigid motion. The rigid component is due to the head rotation, translation etc. and the nonrigid component is due to changes in expression and lip shape. The face can be approximated as a single plane viewed under a perspective projection [16]. As a result, it is possible to describe the optical flows using a parametric model, where the parameters are estimated as suggested in [20]. This model, which ignores the nonrigid motion of facial features, is used to extract the three-dimensional (3-D) rigid body component of motion to align the images.

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

B. Facial Expression Synthesis

In the background processing module, we generate the complete set of viseme+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 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 that appear in viseme v_2 (not present in v_1) and to delete the facial features

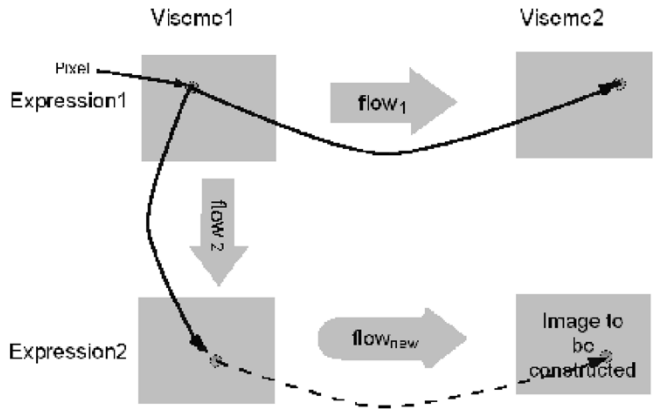


Fig. 5. New viseme-expression pair generation.

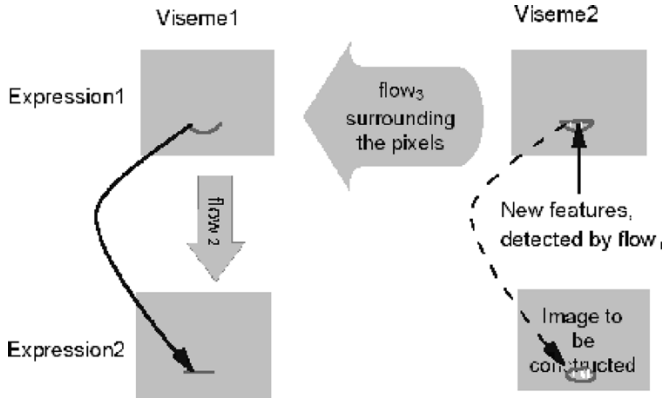


Fig. 6. Introducing new features.

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

To introduce the new features that appear in viseme v_2 (see Fig. 6), detect the facial features that appear in (e_1, v_2) that were not there in (e_1, v_1) using $flow_1$. The pixels in (e_1, v_2) that 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).

To suppress the facial features disappearing in viseme v_2 (see Fig. 7), detect the features that are present in (e_1, v_1) but that disappear in (e_1, v_2) using $flow_3$. The pixels in (e_1, v_1) that 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.

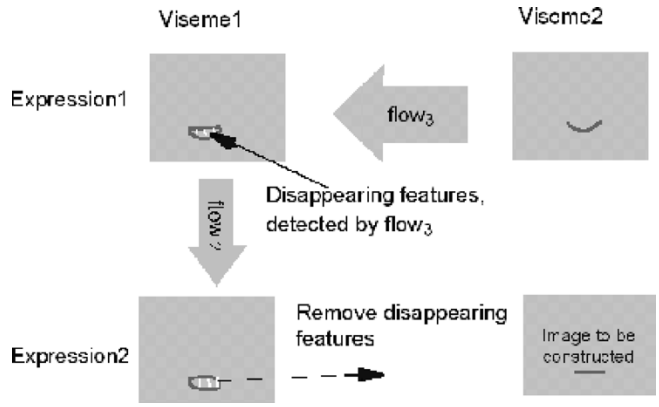


Fig. 7. Suppressing disappearing features.

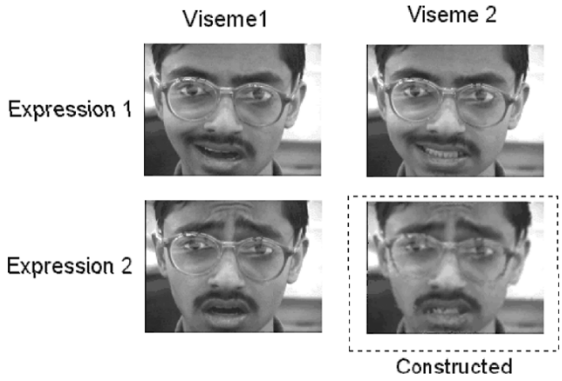


Fig. 8. Existing images and the constructed image with new features appearing.

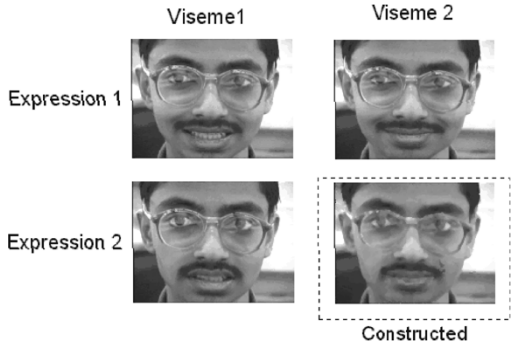


Fig. 9. Existing images and the constructed image with disappearing features.

Figs. 8 and 9 are examples of new viseme+expression combinations generated from the existing ones.

Holes that are produced during expression synthesis are filled using extrapolation diffusion. Holes appear in the final image if the feature at that location has no correspondence from the initial image from which it is generated. We propose to use extrapolation diffusion to intelligently fill the holes after reading the neighboring distribution of pixel values. When the viseme images have holes, the animation generated from them will not be pleasing to the viewer. Using a low-pass filter over the image to fill the noise could have worked only if the holes were a few pixels in area and the assumption that the holes do not have any edges is satisfied. Moreover such a filter would adversely affect other portions of the image, thus further bring down the quality of the synthesized viseme image. We illustrate the extrapolation

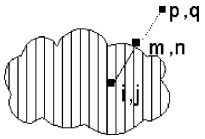


Fig. 10. Extrapolation diffusion.

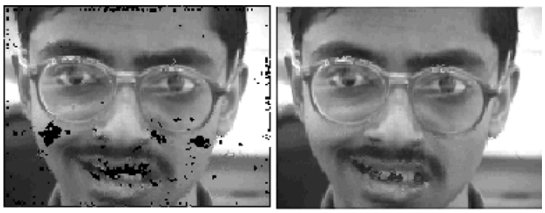


Fig. 11. Using extrapolation diffusion for filling out the holes in the image on the left.

diffusion technique with the help of Fig. 10. Here the shaded region represents the hole and the points outside the shaded region are the normal image points. To fill a point (i,j) inside a hole, we look at the closest normal (m,n) to the hole boundary from (i,j). We extrapolate this normal to the region (p,q) inside the image up to a fixed distance. The $(\mathrm{r},\mathrm{g},\mathrm{b})$ values at (i,j) are calculated as follows:

$$r_{ij} = r_{mn} - \frac{\sqrt{(p-i)^2+(q-j)^2}}{\sqrt{(p-m)^2+(q-n)^2)}}(r_{mn} - r_{pq})$$

$$g_{ij} = g_{mn} - \frac{\sqrt{(p-i)^2+(q-j)^2}}{\sqrt{(p-m)^2+(q-n)^2)}}(g_{mn} - g_{pq})$$

$$b_{ij} = b_{mn} - \frac{\sqrt{(p-i)^2+(q-j)^2}}{\sqrt{(p-m)^2+(q-n)^2)}}(b_{mn} - b_{pq}).$$

The results of using diffusion for holefilling from nearby regions are shown in Fig. 11.

C. Lip Synchronization With Audio

The synchronization with audio is the key task in producing realistic audio-visual output. The trivial method is to generate the frames as the transition occurs, but the animation produced by this method is far from smooth. This is due to the fact that while speaking some words there may be too many viseme transitions in a very short period of time. For example, in the introduction, there are 11 viseme transitions. This provides an unrealistically short time for the lips to change the shapes corresponding to different visemes. If a complete transition from one viseme to another is allowed in this short duration, the animated sequence is jerky. Also the speed at which the speaker is speaking will determine how pronounced the lip movements are. While speaking fast a speaker is not able to produce lip shapes corresponding to each viseme completely. Coarticulation also plays an important role. Coarticulation refers to the influence of past and future sounds on the current sound. Due to coarticulation, lip shapes get modified and the motions overlap. For example in saying the word "bull" the lips get rounded right from the beginning of the word due to the following phoneme /o/ whereas in saying the word "bill" the lips remain closed. Another example is the word "stew" and "still".

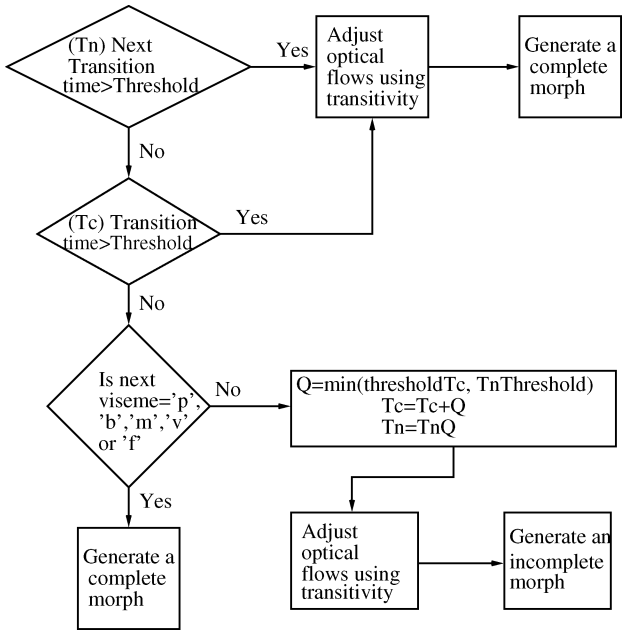


Fig. 12. Audio synchronization.

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

We restrict the extent of the morph depending upon the viseme and the duration of viseme transition. Fig. 12 shows the rules used by our system. Consider a viseme transition between v_a and v_b in duration Tc. Now, if $Tc < Th$, where Th is a threshold that is heuristically set, we generate the morph until $t = Tc/Th$. But there is a catch, consider a transition from viseme v_b to v_c in duration Tn. If $Tn > Th$, then viseme v_b needs to be emphasized and hence the morph to v_b should be complete. In this case we extend the duration of transition $v_a - v_b$ and reduce the duration of transition $v_b - v_c$ by Q, where $Q = \mathrm{Min}(Th - Tc, Tn - Th)$. If the transition $v_b - v_c$ was long enough, then viseme v_b would be morphed from v_a. Further, visemes that represent p, b, m and v, f have to be morphed completely because these visemes involve lip closure or near closure. So if transition occurs to any of these visemes, the morph is completed irrespective of the duration.

Suppose v_b was not completely morphed, then to generate the morph to viseme v_c, we cannot use the optical flows between v_b and v_c computed using the images in our database. We need to know the optical flow between the generated (and incomplete) viseme v_b and v_c. Since the optical flow computations are too costly and almost impossible in real time, we use the transitivity between the optical flows $v_a - v_b$ and $v_b - v_c$ to calculate an

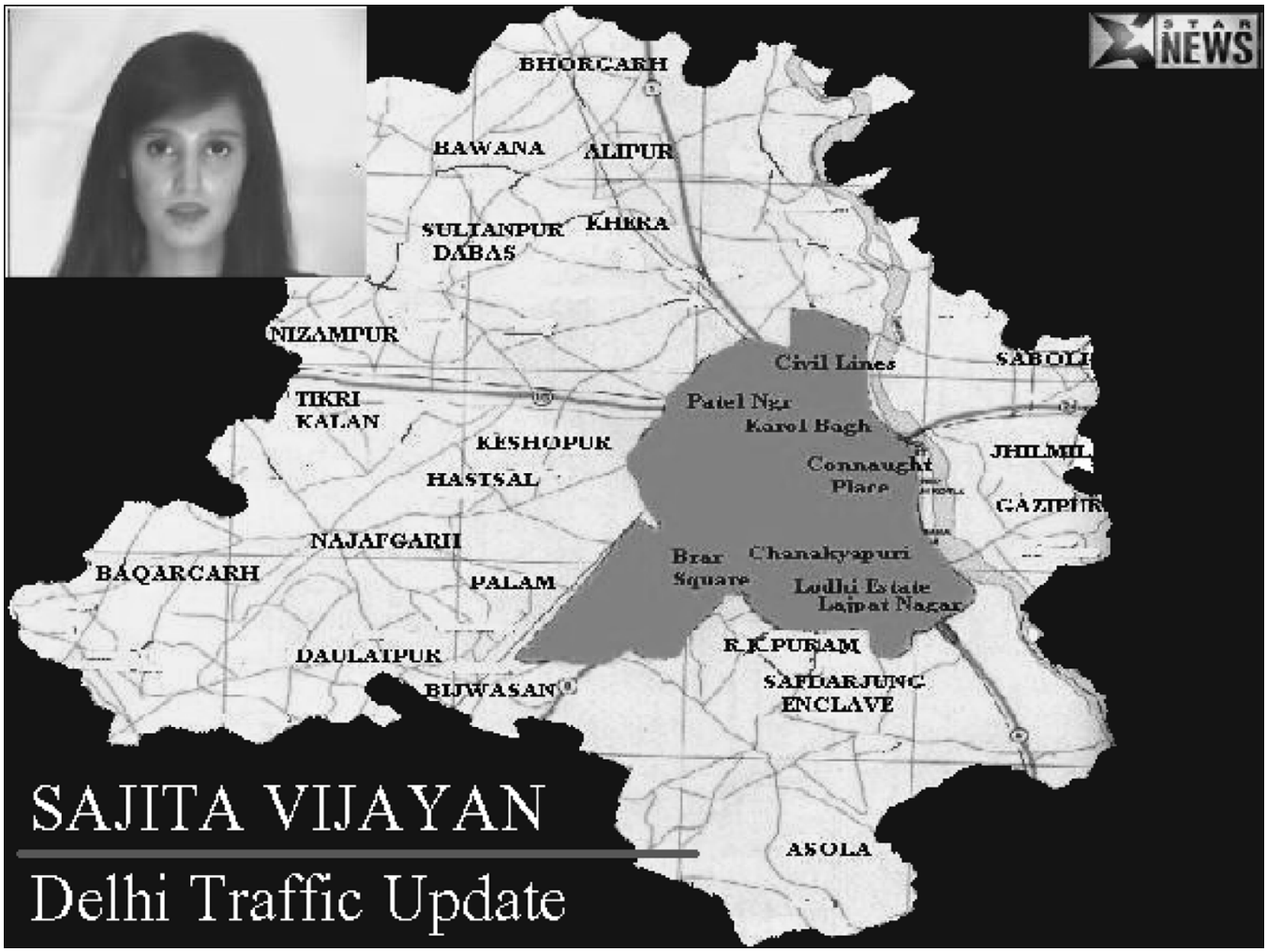


Fig. 13. Newscast example.

approximate optical flow, which is used to generate the morph. Our system uses a threshold $Th = 100$ ms at 30 *fps*.

In this paper, we have not addressed the coarticulation problem directly; however, it must be noted that the simple technique proposed here to limit the extent of morph controls the resulting lip shapes in such a way as to account for the neighboring visemes.

V. System Evaluation

To judge the system we did experiments to evaluate the performance of each of the major components, i.e., the animation system, the translingual system, and the facial expression system.

For evaluating the animation system, 20 evaluators proficient in English were used to judge the perceived quality of synthesized video. Twelve monosyllabic English words chosen to cover different consonant viseme classes were used for the test. A speaker was first recorded speaking each of these 12 words. Next animations of the same speaker saying each of the 12 words were generated. For generating the animation, 12 visemic images of the speaker in the neutral facial expression were extracted from a clip of the speaker uttering the sentence "the sharp quick brown fox jumped over the lazy dog." During the trial the evaluators were presented silent words from the two faces. A clip of a silent word was shown to the evaluator alongside four written words. Each of the 20 evaluators was asked to select the correct written word corresponding to the silent word shown. It was not revealed to the evaluators whether a face was natural or animated. Each evaluator was presented the total set of 24 silent words. The results of this experiment are shown in Table IV. The overall recognition accuracy for natural words was 83% and for the animated words 72%. In most cases, the evaluators could better recognize the natural words compared to the animated words. The labials (Viseme 6) are particularly well recognized in both cases. Interestingly, alveolar fricatives (Viseme 9) and /w/ (Viseme 5) were better recognized in the animated clips.

TABLE IV
Recognition Accuracies in Natural and Animated Faces

	Natural Face	Animated Face
Labials (Viseme 6)	100%	97%
LLL (Viseme 3)	90%	60%
RRR (Viseme 4)	55%	50%
Alveolar Fricatives (Viseme 9)	30%	53%
Lingual Stops (Viseme 7)	93%	55%
WWW (Viseme 5)	65%	85%
Palatal Alveolars (Viseme 10)	100%	70%
Overall	83%	72%

TABLE V
PHONETIC AND VISEMIC CLASSIFICATION RATES

	Translingual Classification Rate	Hindi Classification Rate
Phonetic	24.81%	35.74%
Visemic	40.79%	51.23%

To evaluate the performance of the translingual system, two experiments were performed. Classification experiments were first performed to measure the degradation of phonetic and visemic classification due to the translingual system. Two test sets were generated for this purpose. The first test set used for this experiment consisted of nine hours of continuous speech Hindi sentences. The Hindi transcripts of the sentence is passed on to the English recognition system which aligns the Hindi speech using the translingual technique described in Section III. This aligned and segmented speech was used as the test set to obtain the translingual classification rate. The second test set was obtained using a Hindi recognition system to align the same nine hours of Hindi speech. This was used to generate the Hindi classification rate. The English acoustic models that were used to align the speech were trained with 200 hours of continuous English speech. The Hindi acoustic models were trained on 31 hours of continuous Hindi speech. For phonetic classification, the English phonetic space trained with the same 200 hours of continuous English speech was used. The whole experiment was then performed by dividing the space into visemes classes to get visemic classification rates. The results of the classification experiments are presented in Table V. The results show that there is degradation due to the translingual system. The percentage degradation of phonetic classification is 30.6% due to the translingual system and the percentage degradation of the visemic classification is 20.38% due to the translingual system. As expected, visemic classification results are better than phonetic classification results in both cases. The phoneme to viseme mapping absorbs some of the alignment errors. It is to be noted that the phonetic classification rate is normally in the range of 40%–50% for most languages [1]. A large vocabulary continuous speech recognition system that gives a word level recognition accuracy of about 96% would have a phonetic classification rate in that range [15].

The next set of experiments involved field trials to assess the subjective quality of the translingual system. Ten evaluators proficient in Hindi were used to evaluate the system. Animations for ten monosyllabic Hindi words were presented to the evaluators. Emphasis was laid on evaluating the quality of the animations for Hindi words containing phones that do not have exact mappings in English, i.e., Hindi phones in Table II that are mapped to their closest English equivalent and Hindi phones that are approximated by a string of more than one English phone; six Hindi words were chosen that had Hindi phones with a former mapping and four words that have a latter mapping. Animations of these ten words were generated using one speaker and a viseme set of images. As in the earlier experiment the evaluators were presented four written options for each animation and were asked to select the written word corresponding to the silent word shown. In this experiment, only animated faces (with no speech) were shown to the subject. The results of this experiment are shown in Table V. Taken together the classification experiments and the subjective quality experiments suggest that the translingual system does suffer degradation. However, the degradation may be lower than the 20% suggested by the classification experiment as held out by the subjective evaluation.

TABLE VI
RECOGNITION ACCURACIES FOR HINDI TRANSLINGUAL SYSTEM

	Translingual Animated Face
Closest Mapped Phones	53%
Multiple Mapped Phones	82%
Overall	65%

The Translingual system was also used for animating in Telugu, a southern Indian language, very different from Hindi. Though, no subjective evaluation was done on this the authors believe the animation quality to be comparable to that in Hindi.[1]

To evaluate the facial expression synthesis module, subjects were asked to classify an animated face according to the expression being conveyed. In all cases, at least four of the visemes were synthesized. The speaker is made to say the by now famous sentence "the sharp quick brown fox jumped over the lazy dog" in the different expressions. For the neutral expression, all the 12 visemes used for the animation are natural. For the other expressions, at least half the visemes were synthesized. Clips were now animated using these visemes. In all cases, the correct expression was easily recognized by the subject. However, artifacts start appearing when more and more synthesized visemes are used. These artifacts are concentrated in the mouth region and are distracting to the viewer.

To judge the usefulness of the system for hard of hearing people, the system was tested over one person with hearing impairment using different English sentences. Six sentences each of about seven words length were presented to the subject. The audio was left clean in all cases. It was found that the addition of video improved speech understanding by at least 50%. Though this sample size is small, it does point to the usefulness of the system.

VI. CONCLUSION

An automated system for creating an additional channel for communication is presented. From audio and a few images of a person, a facial animation with lip sync and appropriate expressions is generated. Given a speech recognition system for one language, a method to easily and quickly customize the phonetic and visemic alignments to synthesize video in any other language is presented. Since actual images of a person are used, the animation looks realistic and individual variability is preserved. It is also possible to generate new lip shapes in expressions previously not seen by the system. For the future, it would be worthwhile to consider including other features to the animation system

[1]Native Telugu speakers are invited to evaluate the quality of the animation online at http://http://nitendrarajput.tripod.com/animationexamples.htm. Many online synthesis examples based on the work in this paper are given at this site. The reader is invited to go through these examples.

like correct gaze following, controlled pose variation, eyebrow movement, and eye blinking in the animation system.

REFERENCES

[1] O. Anderson, P. Dalsgaard, and W. Barry, "On the use of data-driven clustering technique for identification of poly- and mono-phonemes for four European languages," in *Proc. ICASSP-94*, pp. 1/121–1/124.

[2] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Speech recognition with continuous parameter hidden Markov models," in *Proc. ICASSP-88*, New York, May 1988, pp. 40–43.

[3] L. R. Bahl, "Large vocabulary natural language continuous speech recognition," in *Proc. ICASSP-89*, Glasgow, Scotland, May 1989, pp. 465–467.

[4] S. Basu, T. A. Faruquie, C. V. Neti, N. Rajput, A. W. Senior, L. V. Subramaniam, and A. Verma, "Speech Driven Lip Synthesis Using Viseme Based Hidden Markov Models," U.S. Patent 6 366 885, Apr. 2, 2002.

[5] D. Beymer, A. Shashua, and T. Poggio, "Example Based Image Analysis and Synthesis," M. I. T. A.I. Laboratory, AI Memo 1431, CBCL Paper no. 80, 1993.

[6] M. J. Black and Y. Yacoob, "Tracking and recognizing rigid and nonrigid facial motions using local parametric models of image motion," in *Proc. 5th Int. Conf. Computer Vision*, 1995, pp. 374–381.

[7] J. F. Cohn and G. S. Katz, "Bimodal expression of emotion by face and voice," in *Proc. 6th ACM Int. Multimedia Conf. Face/Gesture Recognition and Their Applications*, 1998, pp. 41–44.

[8] E. Cosatto and H. P. Graf, "Photo-realistic talking-heads from image samples," *IEEE Trans. Multimedia*, vol. 2, pp. 152–163, Sept. 2000.

[9] R. E. Donovan and E. M. Eide, "The IBM trainable speech synthesis system," in *Proc. Int. Conf. Speech and Language Processing*, 1998.

[10] A. A. Essa and A. P. Pentland, "Coding, analysis, interpretation and recognition of facial expressions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 757–763, July 1997.

[11] T. Ezzat and T. Poggio, "Miketalk: A talking facial display based on morphing visemes," in *Proc. IEEE Computer Animation '98*, 1998, pp. 96–102.

[12] ——, "Facial analysis and synthesis using image based models," in *Proc. 2nd IEEE Int. Conf. Automatic Face and Gesture Recognition*, Oct. 1996.

[13] F. Lavagetto, Arzarello, and M. Caranzano, "Lipreadable frame animation driven by speech parameters," in *Proc. Int. Symp. Speech, Image Processing and Neural Networks*, 1994.

[14] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to Behavioral Principles*. Cambridge, MA: MIT Press, 1998.

[15] C. Neti, N. Rajput, and A. Verma, "A large vocabulary continuous speech recognition system for Hindi," in *National Conf. Communications 2002*, Mumbai, India, Jan 2002.

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

[17] K. Scott, D. Kagels, S. Watson, H. Rom, J. Wright, M. Lee, and K. Hussey, "Synthesis of speaker facial movement to match selected speech sequences," in *Proc. 5th Australian Conf. Speech Science and Technology*, vol. 2, Dec. 1994, pp. 620–625.

[18] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE*, vol. 86, pp. 853–869, May 1998.

[19] V. C. Tartter and D. Braun, "Hearing smiles and frowns in normal and whisper register," *J. Acoust. Soc. Amer.*, vol. 96, pp. 2101–2107, 1998.

[20] R. Y. Tsai and T. S. Huang, "Estimating three-dimensional motion parameters of a rigid planar patch," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 1147–1152, 1981.

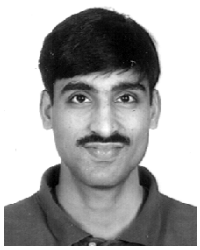
[21] K. Waters and T. Levergood, "An automatic lip-synchronization algorithm for synthetic faces," in *Proc. Multimedia*, Oct. 1994, pp. 149–156.

Ashish Verma received the M.E. degree in electrical communication engineering from the Indian Institute of Science, Bangalore, India, in 1997. He is currently pursuing the Ph.D. at the Indian Institute of Technology, New Delhi, India, in the area of speech personality transformation.

He worked at Hughes Software Systems, Gurgaon, India, from 1997 to 1998. Since 1998, he has been with IBM India Research Lab, New Delhi, India, in the field of speech recognition, audio-visual speech recognition, speech synthesis, audio-driven facial animation, etc.

L. Venkata Subramaniam received the B.E. degree in electronics and communications engineering from the University of Mysore, Mysore, India, the M.S. degree in electrical engineering from Washington University, St. Louis, MO, and the Ph.D. degree in electronics from the Indian Institute of Technology, New Delhi, India, in 1991, 1993, and 1998, respectively.

He has been with IBM India Research Lab, New Delhi, India,since 1998 as a research staff member. His main research interests are in the area of audio-visual speech recognition and synthesis and statistical natural language processing.



Nitendra Rajput received the B.E. degree in electronics and telecommunications engineering from Government Engineering College, Jabalpur, Madhya Pradesh, India, in 1996 and the M.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay, in 1998.

He previously worked on image compression and various digital signal processing algorithms. Since 1998, he has been with IBM India Research Lab, New Delhi, India, as a Research Staff Member. For the last four years, he has been working on audio-visual speech recognition and visual speech synthesis. Currently, he is working on building a speech recognition system for the Hindi language. His research interests are in statistical signal analysis and image processing.

Chalapathy Neti (S'83–M'91) received the B.S. degree in electrical engineering from the India Institute of Technology, Kanpur, India, and the M.S degree in electrical engineering from Washington State University, St. Louis, MO, and the Ph.D. degree in biomedical engineering from The Johns Hopkins University, Baltimore, MD.

He has been with IBM since 1990. He is currently a Research Manager in the Human Language Technologies Department at IBM T. J. Watson Research Center, Yorktown Heights, NY. In this role, he is leading a group of researchers in developing algorithms and technologies for joint use of audio and visual information for robust speech recognition, speaker recognition and multimedia content analysis and mining. His main research interests are in the area of perceptual computing (using a variety of sensory information sources to recognize humans, their activity and intent), speech recognition, multimodal conversational systems for information interaction and multimedia content analysis for search and retrieval. He has authored over 30 articles in these fields and has five patents, with several pending.

Dr. Neti is a member of the IEEE Multimedia Signal Processing Technical Committee and an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA.

Tanveer A. Faruquie received the B.E. degree in electronics and telecommunications from Rani Durgawati University, India, in 1996, and the M. Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay, India, in 1998.

He has been a Research Staff Member in the IBM India Research Laboratory, New Delhi, India, since March 1998. His present research interests include signal processing, statistical learning, machine learning, and image processing.