USING VISEME BASED ACOUSTIC MODELS FOR SPEECH DRIVEN LIP SYNTHESIS

Ashish Verma, Nitendra Rajput, L. V. Subramaniam

IBM India Research Lab
Indian Institute of Technology
New Delhi - 110017
email:{vashish,rnitendr,lvsubram}@in.ibm.com

ABSTRACT

Speech driven lip synthesis is an interesting and important step toward human-computer interaction. An incoming speech signal is time aligned using a speech recognizer to generate phonetic sequence which is then converted to corresponding viseme sequence to be animated. In this paper, we present a novel method for generation of the viseme sequence, which uses viseme based acoustic models, instead of usual phone based acoustic models, to align the input speech signal. This results in higher accuracy and speed of the alignment procedure and allows a much simpler implementation of the speech driven lip synthesis system as it completely obviates the requirement of acoustic unit to visual unit conversion. We show through various experiments that the proposed method results in about 53% relative improvement in classification accuracy and about 52% reduction in time, required to compute alignments.

1. INTRODUCTION

For natural human-computer interaction, it is important that computers can receive information from humans in a natural way. Speech recognition and lipreading are major steps toward this goal. Similarly, it is also very important that computers can communicate with humans in a natural way. Speech synthesis and speech driven lip synthesis are steps to achieve this latter goal. Some of the major applications of speech driven lip synthesis include computer aided instruction, cartoon animation, video games, and multimedia telephony for the hearing impaired.

Some of the previous attempts toward speech driven lip synthesis can be traced from [1, 2, 3, 4]. The basic methodology underlying all these approaches can be described as follows. First the incoming speech signal is time aligned with its corresponding text. Alignment refers to generation of phonetic sequence and duration of each individual phoneme in the sequence for the given speech signal. This can be achieved by incorporating various approaches starting from simple vector quantization based techniques [1] to very sophisticated Hidden Markov Models (HMM) and neural network based speech recognizers [5, 6, 7]. Once the incoming speech signal has been phonetically aligned, there are various ways to map the phoneme sequence into corresponding viseme sequence in terms of image or lip parameters. In [8], they perform a complete speech recognition followed by explicit phoneme-to-viseme mapping, through a table lookup method. This approach is described in Figure 1. Bregler *et al.* in [9] create a database of triphone video segments (trivisemes) using time aligned phoneme-level transcription of a video database. At the time of synthesis, after aligning the input speech using trained acoustic HMMs, they
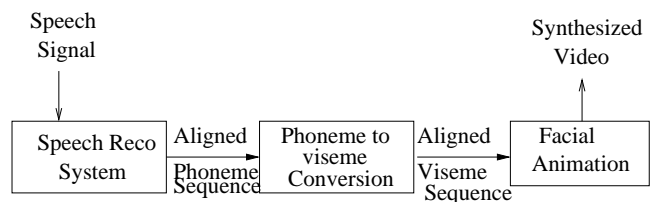

Fig. 1. Earlier Approaches

select the closest triviseme segments based on a distance metric. Yamamoto *et al.* in [5] have trained HMMs from audio-visual database and implicitly mapped the HMM states to lip image parameters, so that at the time of synthesis, aligned HMM state sequence provides lip image parameter sequence to be synthesized. They have also proposed to use viseme context in mapping HMM states to image parameters. Finally, video is synthesized from this viseme sequence using either image morphing or 3-D animation based approaches [10, 11]. In image morphing based approaches, each viseme is represented by an image of the lip, with its shape corresponding to the viseme and in 3-D animation based approaches, a viseme is represented by certain geometrical parameters of the lip. We have reported some results in [11, 12] on an image based approach using optical flows. In [13], it has been shown that translingual visual speech synthesis can be achieved by using speech recognition system of a language different than that of the incoming speech signal, for the alignment procedure.

2. PROPOSED APPROACH

All of the approaches cited in the previous section, use phoneme based acoustic models (VQ, HMM or Neural Networks) to time align the incoming speech signal and then use either explicit or implicit phoneme to viseme mapping to synthesize the viseme sequence. However, for the given application, only the visemic sequence corresponding to the incoming speech signal is of real interest and the corresponding phonetic sequence is of smaller significance. In other words, the phoneme sequence gives much finer level information than is required by the application. For example, if phonemes {/p/,/b/,/m/} map to the same viseme, we do not have to worry about as to which of these phonemes was actually spoken, as long as we know that one of these three was spoken. Hence, we can build just one acoustic model instead of three separate acoustic models representing each of these phonemes. This

gives us motivation to build viseme based acoustic models. Viseme based acoustic models can be used to time align the speech signal directly into the viseme sequence and thereby completely avoiding the audio to visual information mapping used in the previous approaches. These models exploit the physiological relationship between the shape of the vocal tract (including lips) and the corresponding speech signal. In particular, we use three-state viseme based HMM to represent a particular viseme, whose output probability distribution will determine the distribution of the speech signal given this particular viseme was uttered. Viseme based acoustic models incorporate explicit viseme context dependency in acoustic domain to form allovisemes, equivalent to allophones in phoneme based acoustic models, to capture co-articulation effects.

2.1. Training of viseme based acoustic models

For the purpose of training of viseme based acoustic models, acoustic training data needs to be classified into viseme classes. This can be achieved using two approaches. One possible approach is to collect audio-visual data, in order to have acoustic data synchronized with the visual data. Visual data can be time aligned either manually or automatically (through neural networks, HMMs) and then corresponding audio data will be automatically classified into visemes. This however requires accurate video models which are not available with us.

We have used another approach to generate visemically classified data. Phonetically aligned speech database is first generated using phone based HMMs of a speech recognition engine. A phoneme-to-viseme mapping is then applied to convert phonetic alignments into visemic alignments. This gives a speech database aligned to a set of visemes. This approach is shown in Figure 2.
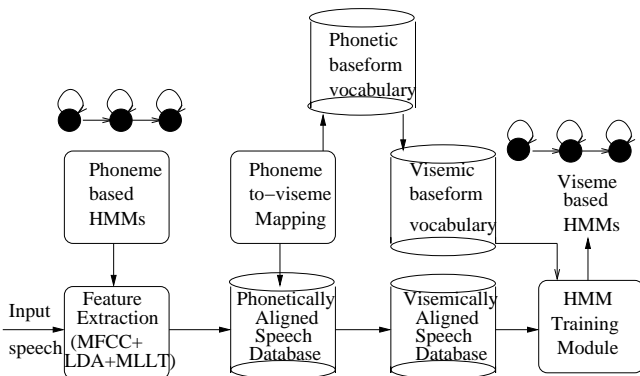

Fig. 2. Training of Viseme HMMs

To train viseme based acoustic models, visemic transcription of the training text corpus is required as compared to phonetic transcription for training phoneme based models. A vocabulary consisting of visemic transcriptions (or visemic baseforms) of Hindi words was created by applying phoneme-to-viseme mapping to the phonetic transcriptions. It is to be noted that the phoneme-to-viseme mapping is used only during the training of HMMs and it would not be required at all during the visual speech synthesis procedure.
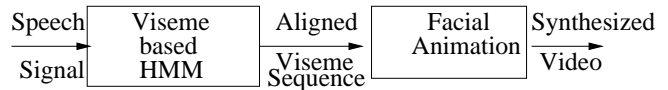

Fig. 3. Visual speech synthesis using viseme based HMMs

2.2. Visual speech synthesis using viseme acoustic models

At the time of synthesis, the incoming speech signal is time-aligned through viseme based HMMs using Viterbi alignment. Since these HMMs are viseme based, the aligned HMM state sequence directly represents the viseme sequence to be synthesized, thereby completely avoiding any explicit or implicit phoneme-to-viseme conversion, as required in earlier approaches. Note that actual visemes may be represented either by image parameters, in case of image morphing based approaches, or geometrical parameters in case of 3-D animation based approaches. The training and use of viseme based acoustic models remains totally independent of the particular kind of viseme representation. The viseme sequence so obtained is fed to facial animation module, which synthesizes the desired video through either image morphing or graphical animation based approach. The specific approach used in our experiments is described in Section 3.

Since visemes are fewer in number than phonemes, we need lower amount of speech data for training of viseme based acoustic models to achieve same level of accuracy as of the phoneme based acoustic models. Moreover, due to lower number of units, the viseme based acoustic model has much lower number of parameters, for example, lower number of distributions in HMM based models and lower number of output nodes and weights in neural network based models. This results in faster alignment of input speech as compared to using phoneme based acoustic models.

2.2.1. Translingual Visual Speech Synthesis

We have reported in [13] that among languages which have a large number of common phonemes, we can use speech recognition system of a language different from that of the incoming speech signal, to align the input speech. This is performed by writing words of the target language using phone set of one of these language in which speech recognition is available. Due to the presence of large number of common phones between these languages, reasonably good phonetic alignment is produced of the input speech signal. Since visemes are fewer in number than phonemes, the percentage of common visemes among these languages is much higher than the percentage of common phonemes. Hence using viseme based acoustic models, will result in further reduced alignment errors for translingual visual speech synthesis.

3. EXPERIMENTAL SETUP

We have performed various experiments to measure relative performance of phoneme and viseme based acoustic models on a large speech database. The database contains 130 hours of speech data from 500 male and female speakers. The database is in the form of about 100,000 utterances of 4-7 seconds duration each in Hindi. This database is part of an ongoing data collection effort at IBM India Research Lab, New Delhi, for development of Large Vocabulary Continuous Hindi speech recognition system [14].

3.1. Acoustic feature extraction

24-dimensional mel-cepstral coefficients were extracted from the speech signal, sampled at 11kHz. Four previous and four following such frames were concatenated to capture the dynamics of the speech signal. These concatenated feature vectors were then multiplied by LDA (Linear Discriminant Analysis) and MLLT (Maximum Likelihood Linear Transformation) matrices to enhance the discrimination capability of these feature vectors among acoustic classes. 60-dimensional feature vectors, so obtained were used to train the acoustic models. We have experimented with HMM based acoustic models, trained on about 100 hours of speech data comprising of 400 speakers from the database.

3.2. Phoneme based models

To compare the results of viseme based acoustic models with that of phoneme based models, we have used the acoustic models of our Hindi speech recognition system [14] as the phoneme based models. These models consist of 61 phones and 3718 allophones (or leaves), obtained using context-dependent trees, modeled with 71,065 Gaussian mixture components. These acoustic models, in combination with the language model, produce word error rates of less than 12% on a general dictation task.

3.3. Viseme based models

A vocabulary of 13,800 Hindi words was built, providing visemic baseforms of the words in the database. This was created by applying phoneme-to-viseme mapping to the usual vocabulary containing phonetic baseforms, used in the training of Hindi speech recognition system. We have used 12 visemes in the experiments, each of which is represented by a full face image having a particular lip shape. These viseme images were extracted from a video clip of the speaker, speaking the sentence, *"The sharp quick brown fox jumped over a lazy dog,"* capturing all 12 visemes, with as little head movement as possible.

Phoneme	Viseme	Phoneme	Viseme
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

We mapped phonetically aligned speech data to viseme classes using a mapping similar to mapping given in Table 1. Incorporating the context of 5 previous and 5 following visemes, 1623 allovisemes (context-dependent visemes) were constructed from the 12 basic visemes through context-dependent trees. These allovisemes were modeled with 26400 Gaussian mixture components. These models were refined afterwards using forward-backward algorithm on same amount of training data as phoneme based models.

### 3.4. Visual speech synthesis

We synthesized visual speech, corresponding to given Hindi sentences by first using phoneme-based and then viseme-based acoustic models. Images corresponding to the 12 viseme given in Table 1 are shown in Figure 4 for a particular speaker. Since the presence of head movement in the viseme images affects the quality of the synthesized video, utmost care is taken to have minimum head movement while shooting the training video sequence. Viseme images are then normalized to remove head movement present still present in the images by compensating for any rigid body motion. For visual speech synthesis, we used optical flow based image morphing technique, presented in [11, 12, 13]. In this technique, optical flow is computed and stored for every pair of viseme images. At the time of synthesis, the stored optical flow is used to generate intermediate image frames between two visemes in the video sequence.

We also synthesized video corresponding to English speech, using our Hindi speech recognition system, through translingual visual speech synthesis approach, presented in [13]. All synthesized video clips can be accessed at *http://www.cse.iitd.ernet.in/~lvs/animations/vhmm.htm.*

In order to have a quantitative comparison of the alignment accuracy for the two approaches, we performed classification experiments using both acoustic models. For qualitative assessment, we performed Mean Opinion Test on 5 subjects, showing them 2 synthesized video clips each, synthesized using phoneme and viseme based HMMs. For translingual visual speech synthesis, one video clip each was shown for phoneme and viseme based HMMs. Results of various experiments are given in Section 4.
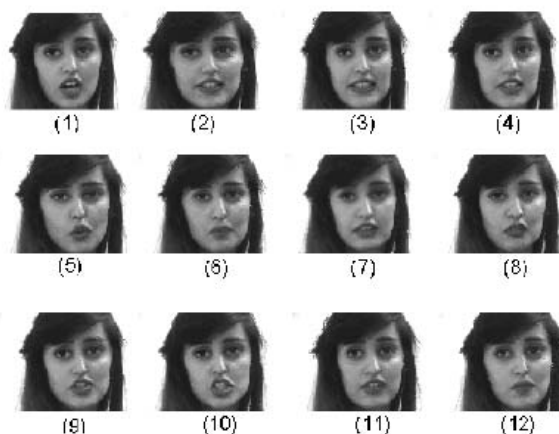
Fig. 4. Viseme Images

3.5. Experiment 1: Phone based models

For phone based models, 26000 utterances, comprising of 130 speakers, were time-aligned with their corresponding phoneme sequences using phone based acoustic models. Out of these, feature vectors from 16,000 utterances, involving 80 speakers, were used to train 61 (number of phones in our speech recognition system) GMM based classifiers and vectors from the remaining 10,000 utterances, from 50 speakers, were used to test these classifiers. To create viseme based classifiers, *phonetically aligned* data was grouped into 12 viseme classes using a similar mapping as given in Table 1, so that both the experiments have same number of classes. Now, parameters of the 12 GMM classifiers were trained from this data using Expectation Maximization (EM) algorithm.

3.6. Experiment 2: Viseme based models

For viseme based models, same amount of data was time-aligned using *viseme based acoustic models*. 12 viseme classifiers were trained from the feature vectors obtained from 16,000 aligned utterances. Note that in this case phoneme-to-viseme mapping is not required as visemes are directly aligned. The accuracy of these classifiers was tested using vectors from remaining 10,000 utterances. Results of these experiments are discussed in the following section.

4. RESULTS AND DISCUSSION

Results of the experiments are shown in Table 2. Phone based GMM classifiers give 35.29% classification. The classification accuracy increases to 46.89% when we merge the phonetically aligned data into viseme based classes using phoneme-to-viseme mapping. This gain in accuracy is a direct result of reducing the number of classes from 61 to 12. However, when we generate visemically aligned data, by using viseme based HMMs, to train the same number of classifiers, the accuracy increases to 54.04%. This improvement comes from the fact that viseme based HMMs make use of viseme based context-dependency which is more relevant here than phone based context-dependency as used in phoneme based HMMs.

	Phoneme Models	Viseme Models
Phone Accuracy	35.29%	-
Viseme Accuracy	46.89%	54.04%
Alignment time	1357 seconds	652 seconds
MOS	6.08	6.24
MOS (Translingual)	6.02	6.86

Table 2. Results for phone and viseme based acoustic models

Furthermore, the speed with which the input speech signal is aligned, also increases dramatically when using viseme based HMMs. Third row of Table 2 shows the time taken by Viterbi alignment algorithm to align about 17 minutes of speech signal through phone based and viseme based HMMs respectively. We can see that the time required for alignment reduces by 52% when using viseme based HMMs. This is due to the fact that there are lower number of base classes in the system, and so the complexity of the engine goes down drastically. As compared to 71,065 GMM components to model phoneme based HMMs, we require only about 26,400 GMM components to model viseme based HMMs. Last two rows in Table 2, show mean opinion scores on a scale of 0 to 10, obtained from subjective tests. The results show that synthesized video clips are perceived better when viseme based HMMs are used.

5. CONCLUSION

We have proposed the use of viseme based acoustic models to align input speech signal in the context of speech driven lip synthesis. We have implemented viseme based HMMs and shown that they produce better results as compared to phoneme based HMMs, both in terms of alignment accuracy and speed. Viseme based context-dependency is also inherently captured in these acoustic models due to their very nature. Another important point that is in agreement with the results is that there are more common lip shapes across languages than common phones, hence viseme based acoustic models are much better suited for use in translingual visual speech synthesis, as compared to their phonetic counterparts.

6. ACKNOWLEDGMENT

7. REFERENCES

[1] Morishima S. and Harashima H., "A media conversion from speech to facial image for intelligent man-machine interface," *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 4, pp. 594-600, 1991

[2] Chen T. and Rao R. R., "Audio-visual integration in multimodal communication," in *Proceedings of IEEE*, Vol. 86, No. 5, May 1998.

[3] Parke F. I. and Waters K., *Computer facial animation*, Wellesley MA: A K Peters, 1996.

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

[5] Yamamoto E., Nakamura S. and Shikano K., "Speech to lip movement synthesis by HMM," *Proceedings of AVSP'97*, Rhodes, Greece, September 1997.

[6] Lavagetto F., "Time-delay neural networks for estimating lip movements from speech analysis: A useful tool in audio-Video," *IEEE Transaction on Circuits, Systems, Video Technology*, Vol. 7, pp. 786-800, October 1997.

[7] Lavagetto F., "Converting speech into lip movements: A multimedia telephone for hard of hearing people," *IEEE Transaction on Rehabilitation Engineering*, Vol. 3, pp. 1-14, March, 1995.

[8] Chen T., Graf H. P. and Wang K., "Lip-synchronization using speech-assisted video processing," *IEEE Signal Processing Letters*, Vol. 2, pp. 57-59, April, 1995.

[9] Bregler C., Covell M. and Slaney M., "Video rewrite: Driving visual speech with audio," *Proceedings of ACM SIGGRAPH'97*, pp. 353-360.

[10] Parke F. I., "Parameterized models for facial animation," *IEEE Computer Graphics Applications Magazine*, Vol. 12, pp. 61-68, November 1982.

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

[12] Faruquie T. A., Neti C., Rajput N., Subramaniam L. V. and Verma A., "Animating expressive faces to speak in Indian languages," *National Conference on Communication*, Mumbai, January, 2002.

[13] Faruquie T. A., Neti C., Rajput N., Subramaniam L.V. and Verma A., "Translingual visual speech synthesis," *IEEE International Conference on Multimedia and Exposition*, New York, July 2000.

[14] Neti C., Rajput N. and Verma A., "A large vocabulary continuous speech recognition system for Hindi," *National Conference on Communication*, Mumbai, January, 2002.