

Phoneme and Viseme based Approach for Lip Synchronization

Ms Namrata Dave¹ and Dr Narendra M. Patel²

¹G. H. Patel College of Engineering & Technology, Gujarat Technological University

²Birla Vishvakarma Mahavidyalaya, Gujarat Technological University

¹namrata.dave@gmail.com, ²nmp_bvm@yahoo.com

Abstract

Phoneme to Viseme mapping has great application in Visual Speech Recognition, Lip Synchronization, Talking Head Applications, movies, news reading, and film industries. Lot of work has been done in area of various face component detection and recognition. Apart from eye detection, ear detection, iris detection etc, lip tracking and lip detection is one of the favourite topics for researchers. Various algorithms and techniques have been implemented so far to achieve better and better performance. Normalized RGB colour scheme, HSV colour model, Lip detection using HUE segmentation and many more techniques have been implemented and are in the boom. All methods are having their own pros and cons. We are aiming to extract out phonemes from speech as well as we extract visual feature i.e. visemes from face by using hue and saturation values. The reason behind the selection of this algorithm is that, it performs well under various illumination conditions, which is the one of the dimension of difficulty in the area of lip detection. We are aiming to carry out the work on in-house database with varying lighting and noisy conditions.

Keywords: Phoneme, Viseme, Lip Synchronization

1. Introduction

Nature of human speech is bimodal [1]. Speech observed by a person depends on audio features, as well as on visual features like lip synchronization or facial expressions. Visual features of speech can compensate for a possible loss in acoustic features of speech due to noisy environments. This combination of auditory and visual speech recognition is more accurate than audio only or visual only features. Perception of speech can be enhanced with use of multiple information sources like audio and video features of speech. Since last two decades, lots of research is done with use of bimodal nature of speech into area of real time speech driven facial animations.

The goal is to animate the face of a speaking avatar (i.e. a synthetic 3D human face) in such a way that it realistically pronounces the given text, which is based only on the speech input. Especially important component of facial animation is the movement of the lips and tongue during speech [2]. For a realistic result, lip movements must be perfectly synchronized with the audio input.

Lip synchronization is the determination of the mouth motion and tongue during speech [3]. Speech sound is produced by the vibration of the vocal cords in the case of voiced sounds and air turbulence in the case of whispered sounds [4]. Produced sound is modeled based on vocal tract which includes throat, tongue, mouth, teeth, lips and nasal cavity. Vowels are created by the relatively free passage of breath through the larynx and oral cavity, while consonants are produced by a partial or complete obstruction of the air stream by any of various constrictions of the speech organs.

Intonation characteristics are pitch, amplitude and voiced/whispered quality and they are dependent on the sound source, while vocal tract determines the phoneme. A phoneme is the basic unit of acoustic speech. Visual representation of phonemes is called viseme. There are many acoustic sounds that are visually ambiguous. Therefore, there is a many-to-one mapping between phonemes and visemes. To make lip synchronization possible, position of the mouth and tongue must be related to characteristics of the speech sound (basic idea of lip synchronization is shown on Figure 1. Whereas, positions of the mouth and tongue are functions of the phoneme and are independent of intonation characteristics of speech sound.

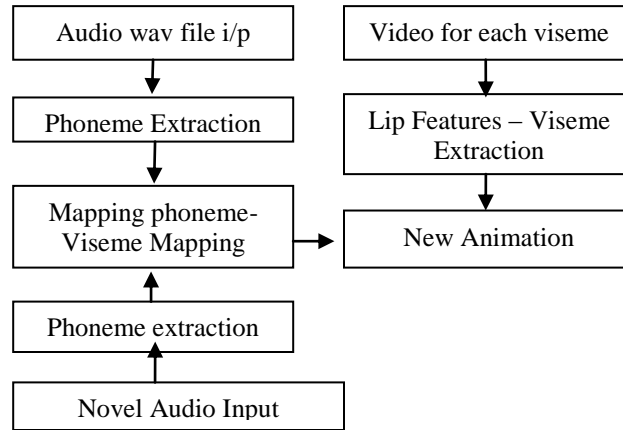


Figure 1. Basic Idea of Lip Synchronization

Various techniques have been proposed to convert human voice into the facial motion and then to drive the facial animation from speech signals [3-10, and 12]. The key issue is audio-to-visual mapping that converts acoustic speech to mouth shape parameters. This problem can be solved at several different levels, as it is explained in the next Section. Section 2 gives an overview of the most used approaches that attempt to extract the mouth shape information from the speech signal. Section 3 explains our approach of phoneme to viseme mapping for lip synchronization, while Section 4 and 5 briefly explains database, our proposed plan for an implementation and results respectively. The paper closes with a conclusion and a discussion of the future work.

2. Feature Extraction

Feature extraction method includes two level of feature extraction. First, phoneme features are extracted from recorded audio and next is to extract visual features from recorded video for each phoneme or group of phoneme.

2.1. Audio Feature Extraction Methods

First step is to extract phonemes. Phonemes are best described as linguistic units. They are the sounds that group together to form our words, although quite how a phoneme converts into sound depends on many factors including the surrounding phonemes, speaker accent and age.

English uses 40 phonemes to convey the 500,000 or more words for making good data item on which speech engines work. There are many ways to extract speech

features. These levels are Front end (signal level), Acoustic model (phoneme level) and Language model (word level). Each of the three levels can be applied within speech-driven face animation system. However, the choice will depend on a specific application, considering characteristics of the individual solution. In our system we have used phoneme level feature extraction approach.

Table 1. List of Phonemes & Visemes Used for Digits One to Ten

Word	Phoneme	Viseme
One	w ah n w an n v ao n	v ao n
Two	t uw d uw	t uw
Three	th r iy dh iy	th r iy
Four	f ao r	f ao r
Five	f ay v	f ay v
Six	s ih k s	s ih k s
Seven	s eh v ax n	s eh v ax n
Eight	ey t	ey t
Nine	n ay n	n ay n
Ten	t eh n	t eh n

Phonemes are extracted by applying Fourier Transform on speech waveform. It allows the waveform to be analyzed in the frequency domain. Table 1 shows the phonemes extracted for one to ten digits and corresponding visemes in our research work.

2.2. Visual Feature Extraction Methods

The visual front-end stage encodes stimuli coming from the visual cues (mainly the lips) of a speaker and transforms it into a suitable representation that is compatible with that of the recognition module [11]. A number of approaches have been proposed in literature for this purpose. These approaches can be categorized as either appearance-based, shape-based, or both [18].

Since the visemes used are not developed for a specific language, the English phonemes are divided into the viseme class that best describes the phoneme. The 10 viseme classes with related phonemes of ten digits are described in Table 1. However, prior to this feature extraction process, a number of pre-processing steps have to be done. This involves face detection followed by ROI extraction. Then, the lips of the speaker are tracked in consecutive frames. Following these steps, and given an informative set of features, the visual front-end module can proceed with feature extraction.

3. Phoneme to Viseme Mapping

The One key issue in bimodal speech processing is the audio to visual mapping. Many approaches have been proposed in an attempt to solve the problem of extracting the mouth shape information from the speech signal. The most used techniques are vector quantization, neural network, Gaussian mixer model and hidden markov model. Hidden Markov Model takes into consideration audio contextual information, which is very important for modeling mouth co-articulation during speech. That is not the case with vector quantization and the Gaussian mixture.

3.1. Neural Networks

Mapping between the acoustic speech and the appropriate visual speech movements can be determined by training a neural network [10]. In the training phase, input patterns and output patterns are presented to the network. Suitable technology, as well as the number of hidden layers and the number of nodes per layer should be determined. Single network can be trained to reproduce all the visual parameters or many networks can be trained so that each network estimates a single visual parameter.

Neural networks can be trained for audio to visual mapping so that they take into account the audio contextual information (*e.g.*, Time-delay neural networks - TDNN). TDNN is more computationally efficient than HMM, but requires a large number of hidden units, which results in high computational complexity during training phase. Many approaches use a combination of the different techniques. We have used neural network for mapping between phoneme and viseme.

4. Dataset

We have created our own audio visual database for our experiment. We have recorded sound files on normal computer with inbuilt mike and in normal noisy condition. Sound files are recorded at laboratories as well as home, which are not quite noise proof. Speakers belong to age group of 18 to 35 year.

Table 2. Audio Property for each .wav File

Sr. No	Property	Value
01	Bit Rate	128 Kbps
02	Audio Sample rate	8 KHz
03	Audio Sample Size	16 Bits
04	Audio Format	PCM
05	Channel	1 Mono

We have used audio recorded for total 10 Speakers: 5 male and 5 female. The rationale behind collecting the speech samples from noisy environments is to represent a real world speech samples collection, because most speech recognition systems are meant to be used in such a little bit noise environments. Therefore, collecting speech samples from noisy environments was purposely done. Audio property for each recorded audio is given in Table 2.

Video is recorded in computer department laboratory in natural lighting conditions without any lipstick or external makeup. We have recorded video for each group of visemes by pronouncing phonemes. We then extracted mouth parameters from each frame

and then find averaged parameters. Frame rate for recording video is 30 frames per second, file format is avi and resolution of recorded video was set to 640 by 480.

5. Implementation and Results

Our system is mainly divided into three phases: Phoneme Extraction, Viseme Generation using Lip Extraction Algorithm and Digit Recognition using neural network. Finally we map phonemes to viseme by using phoneme to viseme look up table. We will explain result of each stage of our system. [15-17]

For the phoneme extraction we are obtaining results with approximately 80 percent of accuracy. We have extracted phoneme information from recorded audio files. Which is mapped to unique index and those indices for each digit file is stored in index files. Phoneme extracted for a recorded sound file of word ‘one’ is listed in Table 3, which is output of phoneme extractor program that we have used in our work.

Table 3. Phoneme Extraction Result for Word “one”

Digit ‘ONE’		INDEX
[o1.wav]	1047	
w	538	37
ah	665	3
n	919	24
_	1046	

However, prior to this visual feature extraction process, a number of pre-processing steps have to be done as shown in Figure 2. This involves face detection followed by ROI extraction. Then, the lips of the speaker are tracked in consecutive frames. Following these steps, and given an informative set of features, the visual front-end module can proceed with feature extraction.

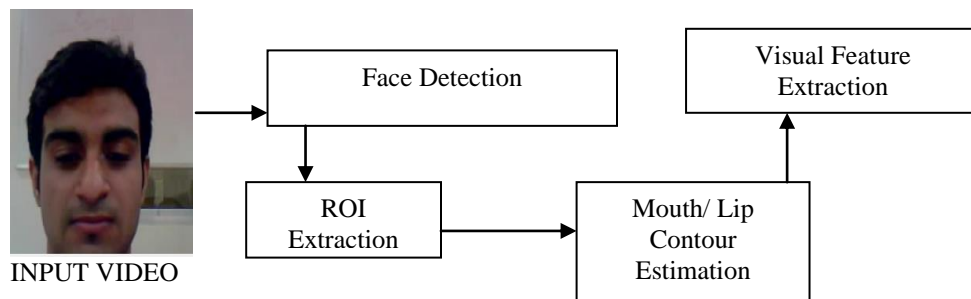


Figure 2. Visual Feature Extraction Process

The choice of viseme features depends on the specifics of the lip movement synthesis. The simplest set required for synthesis consists of the height and width of the outer lips as shown in Figure 3. With these two features, it is possible to derive synthetic lip movements that are believable. These features are sufficient enough to drive a face model. This is also the minimum set sufficient to derive coordinates for an arbitrary, predefined lip shape. For development and evaluation purposes, we focus on the accuracy of these two features

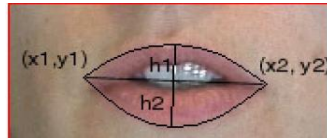


Figure 3. Mouth Region Height and Width for Viseme

To extract visual features, we implemented lip detection method to extract the outer lip contour that combines edge based and region based algorithms. The results from the two methods are then combined. The novelty lies in the fusion of two methods, which have different characteristics and thus exhibit different type of strengths and weaknesses. The other significance of this study lies in the extensive testing and evaluation of the detection algorithm on a realistic database.

Figure 4 gives an overview of the lip detection algorithm. Video is divided into frames. Then each frame is processed as an image. The first step is to extract face using skin based detection algorithm, second step is to select the mouth Region of Interest (ROI) using the lower one third of the detected face. The next step involves the outer lip point's detection where the same mouth ROI is provided to the edge and region based methods. Finally the results from the two methods are fused to obtain the final outer lip points.

In order to increase the accuracy of feature extraction method a combined method is developed in which the results obtained by applying different techniques are combined to yield a higher accuracy algorithm. Figure 4 shows the proposed method.

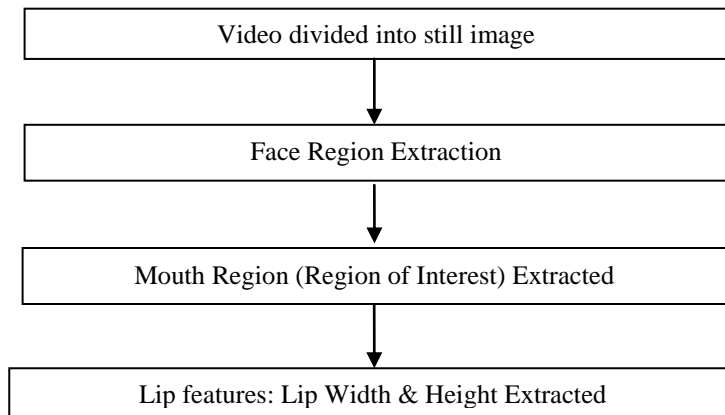


Figure 4. Flow of Proposed Algorithm for Lip Feature Extraction

In each stage component labelling is exploited to separate individual parts from each other and so to eliminate the excessive and noisy parts. Figure 5 shows result of viseme extraction procedure. We have extracted visemes for all 12 classes and one for silence.





Figure 5. Viseme Classes for Speaker "x"

We have used neural network as classifier to map phonemes to corresponding visemes. It is not necessary to have a unique viseme for every phoneme, since there are several phonemes that have the same or similar facial expression. There are 15 predefined viseme Facial Animation Parameters (FAP group 1) in the MPEG-4 standard [11].

A Two-layer back-propagation neural network as shown in Figure 6 is used with sigmoid activation function in hidden layer and output layer. Neural network classifies input phonemes to generate corresponding viseme. We have created 4 different neural network to test and train phonemes. Each NN was created with different combinations of input neurons, hidden layer neurons and single output neuron.

We have analysed system with different number of hidden layer neurons and the combination for which we get good results are considered here and saved for further processing. Reason behind using different neural network is number of phonemes generated for each digit is different. Neural network can be trained for inputs of same size. Also digits which generate similar phonemes we have not used them to train same neural network. For example digit 'four' generate 'f' as well as digit 'five' also generate 'f' as its first phoneme.

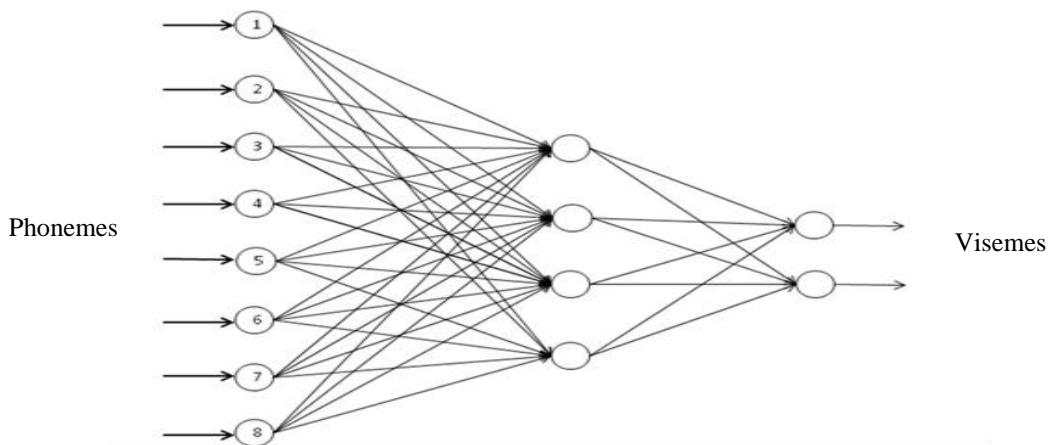


Figure 6. Neural Network Architecture

The recognition part is straightforward. Each of the 4 networks takes the speech as input and the recognized phoneme is simply the one that corresponds to the network that produces correct output as per the given target vector. As soon as a digit of the

incoming speech is classified into the correct digit, the corresponding visemes for each of the phoneme of that digit is to be decided and sent to the animated face model. Ideally, the network corresponding to the correct phoneme should return 1 and the rest 0, but that is seldom the case. Instead, the output lies within some specific interval and the network that produces the *largest* output is picked as the correct phoneme. Errors may occur, *i.e.*, an incorrect phoneme can be identified as the correct one by the neural networks. If a viseme should be picked in *each* frame, an error would cause a sudden discontinuous facial animation. Therefore, one frame for each viseme of every speaker is summarized.

Table 4. Results of Viseme ‘v’, ‘ao’ Extracted for Speaker1



File Name	Viseme Letter	Width	Height	Ratio	Viseme Image
av1.avi	V F	65	34	1.92	
Aao1.avi	ao	47	28	1.67	

Table 5. Confusion Matrix for Phoneme based Digit Recognition

Exp. Class Digits Percentage	Recognized Class									
	1	2	3	4	5	6	7	8	9	10
1	92		2.2	5.53						
2		82.26			17.78					
3	.11		88.93	10.005						
4	15.56		17.79	67.78						
5		46.7			53.6					
6						80.04	20.01			
7						.67	93.38			
8								94.49	4.45	.11
9								33.35	67.78	
10								53.34		46.67

Table 5 is the Intraclass confusion matrix showing the percentage of the phonemes in their viseme classes for digits1-10 vertically being identified as phonemes within the viseme classes for digits 1-10 on the horizontal. Table 5 shows results obtained by training neural network for 75 files of phonemes of each digit of 5 different speakers.

Testing is done same network with samples taken from 5 unknown speakers. Table shows result of percentage of correct recognition of each digit which is averaged from tested results. We can see from table that on an average we are getting good efficiency between 80 – 90 percent. For digits five and ten we are not getting good result at present. One of the reasons for not getting good results for digits two and five is we have taken less number of input to neural network to train and test as well as phonemes extracted using phoneme extractor is not satisfactory for them.

6. Conclusion

We are getting good results for phoneme based lip synchronization as compared to only LPC or other features based lip synchronization where speech is analysed to get such features by working on signal level. We are not getting good results for only two digit classes which can obviously improve by recording sounds in noiseless environment with more number of speakers pronouncing those digits accurately to generate good phonemes. We have planned to work on neural network configurations to get better results.

References

- [1] T. Chen and R. Rao, "Audio-visual integration in multimodal communication", Proceedings of IEEE, Special Issue on Multimedia Signal Processing, (998) May, pp. 837-852.
- [2] P. Vanroose, G. A. Kalberer, P. Wambacq and L. Gool, "From speech to 3D face animation", Proceedings of the Benelux Symposium on Information Theory, (2002).
- [3] D. F. McAllister, R. D. Rodman, D. L. Bitzer and A. S. Freeman, "Lip synchronization for Animation", Proceedings of SIGGRAPH 97, Los Angeles, CA, (1997).
- [4] J. P. Lewis and F. I. Parke, "Automated lip-synch and speech synthesis for character animation", Proceedings of SIGGRAPH, (1990).
- [5] S. Kshirsagar and N. Magnenat-Thalmann, "Lip synchronization using linear predictive analysis", Proceedings of IEEE International Conference on Multimedia and Expo, New York, (2000).
- [6] F. J. Huang, T. Chen, "Real-time lip-synch face animation driven by human voice", Proceedings of IEEE Multimedia Signal Processing Workshop, Los Angeles, California, (1998).
- [7] Y. Li, F. Yu, Y. Xu, E. Chang and H. Shum, "Speech-driven cartoon animation with emotions", Proceedings of the ninth ACM international conference on Multimedia, Ottawa, Canada, (2001).
- [8] M. Brand, "Voice Puppetry", Proceedings of SIGGRAPH'99, (1999).
- [9] Y. Huang, X. Ding, B. Guo and H. Shum, "Real-time face synthesis driven by voice", Proceedings of Computer-Aided Design and Computer Graphics, Kunming, PRC, (2001).
- [10] D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry and T. Rodriguez, "Picture my voice: Audio to visual speech synthesis using artificial neural networks", Proceedings of AVSP'99, Santa Cruz, California, (1999).
- [11] I. S. Pandžić and R. Forchheimer, "MPEG-4 Facial Animation - The Standard, Implementation and Applications", John Wiley & Sons Ltd, (2002).
- [12] P. Hong, Z. Wen and T. S. Huang, "Real-time speech driven avatar with constant short time delay", Proceedings of International Conference on Augmented, Virtual Environments and 3D Imaging, Greece, (2001).
- [13] Ph.D. Thesis, "Animating Faces from Speech", By Gwenn Englebienne, Faculty of Engineering and Physical Sciences, (2008).
- [14] G. Mahesh, N. Dave and N. M. Patel, "Performance Analysis of Lip Synchronization Using LPC, MFCC and PLP Speech Parameters", Computational Intelligence and Communication Networks (CICN), 2010 International Conference on. IEEE, (2010).
- [15] N. Dave, "Real-Time Lip Movements Driven by Human Voice", International Journal For Advance Research In Engineering and Technology, (2013).
- [16] N. Dave, "A Survey on Feature Extraction Methods LPC, PLP and MFCC in Speech Recognition", International Journal For Advance Research In Engineering And Technology, (2012).
- [17] N. Dave, "Performance Analysis of LPC, PLP and MFCC parameters in Speech Recognition", ADIT Journal of Engineering, (2009).
- [18] A. M. Mustapha, Master Thesis, "A Multimodal Sensor Fusion Architecture for Audio-Visual Speech Recognition", Electrical and Computer Engineering, Waterloo, Ontario, Canada, (2007).

Authors



Namrata Dave
Assistant Professor,
Computer Engineering Department,
G H Patel College of Engg and Tech,
Gujarat, India.



Dr Narendra M. Patel
Associate Professor,
Computer Engineering Department,
Birla Vishvakarma Mahavidyalaya, Gujarat, India.