Abdul Rafay Abbasi and Naveed Ahmad

Urdu Viseme Identification

***Abstract**: We have identified parameters of lip shapes for Urdu phonemic strings using tri-viseme samples (VcV). These parameters were used for animation of human lips. In order to perform this animation, a framework for simulation of human lips was developed in Open GL. These simulations have validated data values measured for the identified parameters.*

***Keywords:** Viseme, Visual Speech Synthesis*

1. INTRODUCTION

1.1 What are visemes?

Speech comprises of a mixture of audio frequencies, and every speech sound belongs to one of the two main classes known as vowels and consonants. Vowels and consonants belong to the basic linguistic units known as phonemes, which can be mapped to visible mouth shapes called visemes. Visemes and phonemes can be used as the basic units of visible articulatory mouth shapes (Waters, Levergood, 1993). A phoneme is an abstract representation of a sound, and the set of phonemes in a language is defined as the minimum number of symbols required to represent every word in that language (Breen, Bowers and Welsh). In Urdu there are 60 such symbols. A phoneme is not a sound, it may be viewed as the label for a set of sounds which when spoken as part of a word do not change the meaning of the word. So, we as humans do not speak in discrete units, speech is simply produced as a continuous flow of articulatory movements, which upon inspection can be 'written' as discrete set of symbols called phonemes.

The set of visemes in a language is often defined as the number of visibly different phonemes in that language. A simple definition of viseme would be that a viseme could be generated from a set of archetypal sounds in a language based on the phonemes of that language (Breen, Bowers and Welsh). Möttönen, Olivés et.al. defines a viseme set as phoneme realizations that are visually inextricable from each other. In lip reading studies viseme categories can be defined as clustering response distributions to observed phoneme articulations. These clusters are then used to find those phoneme articulations, which are perceived to be similar. Typically a cluster is considered as a viseme category if the proportion of within cluster responses is at least 70% of the total amount of responses to the phoneme articulations that are included in the cluster (Möttönen, Olivés et.al.).

Though several standards for visemes (Tiddeman, Perret), (Ezzat, Poggio) exist but they were developed for English language. Since each language comprises of a different phonetic set (sounds) therefore, visemes have to be identified separately for each language. The purpose of this research was to identify the visemes of Urdu so that applications of this technology can be implemented. Möttönen, Olivés et.al. developed a Finnish talking head, which had to identify Finnish visemes for each Finnish phoneme. Moreover, since a variety of coarticulation strategies are possible and different strategies may be needed for different languages (Cohen, Beskow and Massaro), so it was necessary to identify Urdu visemes in order to study coarticulation effects in Urdu. One major advantage of identifying visemes for Urdu is that once the viseme table has been constructed it can be correlated to specific Urdu text to speech synthesizer to construct Urdu speech applications for e.g., Urdu talking head, Urdu Visual Email and Urdu Visual Chat etc.

1.2 Applications of viseme identification

Identifying the viseme or viseme categories is the basic problem faced while designing realistic computer facial animations. In order to lip-synch the mouth shape to different sounds in a language, the mouth shapes i.e., visemes corresponding to each phoneme have to be calculated. Computer facial animation has found wide applications in the fields of visual speech synthesis, deaf children education and other kinds of language training techniques.

Traditional facial animation relies on hand drawn images of the lips. A sound track is first recorded, and then an exposure sheet is marked with timing and phonetic information. The animator then draws a mouth shape corresponding to a precise time frame. Early work on automating this process focused on the animation of geometrical facial models, which can be developed using pre-defined expressions and mouth shape components. The focus then shifted to physics-based anatomical face models, with animation performed using numerical simulation of muscle actions. More recently an alternative approach has emerged which takes photographs of a real individual enunciating the different phonemes and concatenates or interpolates between them (Tiddeman, Perret). Another approach to computer facial animation is to use parametric facial models that can be generated by specifying a set of parameter value sequence. The parameters control facial features such as the mouth opening, height, width and protrusion. Parametric models allow us to generate a wide range of required mouth patterns (Waters, Levergood). All these advances in computer facial animation used in combination with text to speech systems have created a new field of technology known as visual speech synthesis.

One of the earliest applications of visual speech synthesis was to use it to understand the mechanism of speech perception. It was discovered that there are three mechanisms involved in speech perception: auditory, visual and audio-visual (Waters, Levergood). So perception of speech could only be well studied with the

help of visual speech (visual auditory) experiments. Moreover, improvement in visual speech synthesis can only be brought by studying in detail, how real humans produce speech (Cohen, Beskow and Massaro). By using both auditory and visual modalities scientists were able to better understand the speech than by relying only on audition, especially if the speech was exposed to noise, bandwidth limitations, hearing limitations or other disturbances (Möttönen, Olivés et.al.).

In one of the recent applications, visual speech synthesis is being used to explore new ways of presenting information and enhancing computer user interfaces. Facial animations combined with speech synthesizers to provide a unique form of computer user interaction, which will make computers accessible to a much wider range of people. This application will help realize the dream of a true personable computer (Möttönen, Olivés et.al.).

Visual speech synthesis is also being used to address the problem of transmitting video over limited bandwidth. Applications like IMPersona (www.impersona.com), which provides fast visual chat facilities, have changed the concept of chatting online.

Another application of this technology is in Deaf Children Education. Deaf and hearing-impaired children use auditory visual speech (visual auditory) perception in order to lip-read the words. Children with hearing-impairment require guided instruction in speech perception and production. Some of the distinctions in spoken language cannot be heard with degraded hearing, even when the hearing loss has been compensated by hearing aids or cochlear implants. To overcome this limitation, visual speech can be used as speech targets for the child with hearing loss (Cohen, Beskow and Massaro). It was observed by (Möttönen, Olivés et.al.) that visual information improves the intelligibility of both synthetic and natural acoustic speech by approximately 15% to 50%. Lip reading research has shown that deaf children can receive 70% of the information in speech if the combined approach of sight with sound (visual auditory) is used.

Moreover, human teachers have proved insufficient for teaching deaf children. Due to specific articulatory and phonetic features associated with a human teacher many deaf children complain that they can only understand their teacher. This provides motivation to develop an automated facial animation whose parameters can be changed at runtime to model the viseme shapes of persons of different ages and sex. Lip movements associated with each of the speech sounds under ideal viewing conditions, taken from (Jeffers), is given in Appendix B.

This technology can also be used in general language training, as in the learning of non-native languages and in remedial instruction with language-disabled children. Speech therapy during the recovery from brain trauma could also benefit from this. Children with reading disabilities could profit from interactions with animation of talking head (Cohen, Beskow and Massaro).

2. LITERATURE REVIEW

Different researchers have used different methods to identify visemes. Breen, Bowers and Welsh developed a low computation mouth shape generation algorithm attempting to account for many of the known coarticulation effects observed in continuous speech. They used di-viseme (A di-viseme records the change in articulation produced when moving from one viseme to another (Breen, Bowers and Welsh)) and analyzed only 48 samples of video recordings. Accurate coarticulation effect could not have been measured since a di-viseme does not take into account context effects of the previous phoneme that was spoken. Moreover they stored di-visemes as a set of nonsense syllables increasing the probability that viseme mappings may be inaccurate.

Tiddeman and Perret generated new, photo-realistic visemes from a single neutral face image by transformation using a set of prototype visemes. They used these visemes to generate visual speech from photographs and portraits where a full set of visemes is not available. However, there systems did not cater for coarticulation.

Möttönen, Olivés et.al. used separate visemes for each phoneme of their Finnish talking head. However, they overlooked the effects of phonological context and coarticulation on visemes. They performed an intelligibility test for the viseme parameters identified by mapping them onto a facial model. The test corpus consisted of 39 VcV words (13 consonants in symmetric vowel contexts of /a/, /e/ and /y/). The words were presented with natural, audiovisual, synthetic audiovisual, natural audio only, synthetic audio only, natural audio/synthetic visual and synthetic audio/natural visual conditions and with 0, -6, -12, -18 dB signal-to-noise ratio (SNRs). The subjects were 20 native Finnish speakers 20 to 33 years old.

Researchers such as Cohen and Massaro do not use visemes explicitly for computer facial animation. The method proposed by them attempts to generate realistic movements through an algorithm, which combines basic facial action units to produce an overall effect. The combination process, associates a set of feature based dominance functions with each phoneme in that language, where a phoneme will have been decomposed into a set of basic attributes such as lip rounding (Breen, Bowers and Welsh).

3. METHODOLOGY

In order to consider the immediate context effects on each viseme, we decided to use tri-viseme video recording instead of di-visemes. This allowed us to cater for effects of coarticualtion without employing any separate algorithms. This decision was based on the observation of (Breen, Bowers and Welsh) that a tri-viseme records the immediate left and right context effects on a center viseme.

Since it would not be feasible to analyze 12427 (17 vowels x 43 consonants x 17 vowels) recordings, we restricted our recording set to all places {p, tt, t and k) and four cardinal vowels {i, o, u, a}. This choice was based on the observation that all visibly different viseme shapes can be catered by this set. A total of 256 samples

were collected with 4 samples for each of the 64 phoneme string combinations. Each of the two subjects uttered 2 samples for each phoneme string.

We selected VcV strings because changes in lip shapes for vowels are quite apparent, however changes in lip shapes for consonants can only be observed in the context of vowels. Moreover (Waters, Levergood) had already used VcV strings to successfully determine mouth shapes.

The most difficult problem was to determine a standard set of parameters, which could be used for different applications, and can be easily converted to other standards. MPEG-4 standard provides 84 feature points located in the face in order to provide a reference for facial animation parameters (Ostermann). We decided to use the feature points for lips provided by MPEG-4, as shown in Figure 1.

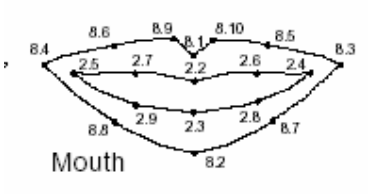

Figure 1 MPEG-4 standard features points for lips

Since MPEG-4 is the first international standard that standardizes multimedia communication including synthetic video and 3D graphics, using its parameters ensured wide and easy use of our results.

3.1 Equipment Used

Equipment included an YHDO CCD Camera with Lens of 2.8 mm, on a Pentium III 500 machine. We used pixel view playTv pro capture card with conexant Bt878 chipset to interface the camera with the computer. Video capture rate was 30 frames per second, size of 160 x 120 pixels. The video standard of PAL –B, D, G, H, I was used. In order to get the side view of the person we placed a mirror, at an angle of 45 ° adjacent to the person.

4. RESULTS

During the course of collecting video samples we came across a very striking phenomenon that lip shapes do not change while we speak short vowels. There is a tendency in humans to make as little movement as possible while producing sounds. Since moving lips require a lot of muscle movement, so for short vowels humans do not move their lips at all. This phenomenon can be described by the principle of least effort (Assimilation).

Summary results for different parameters for the phoneme string /apu/ are shown in Figure 2.

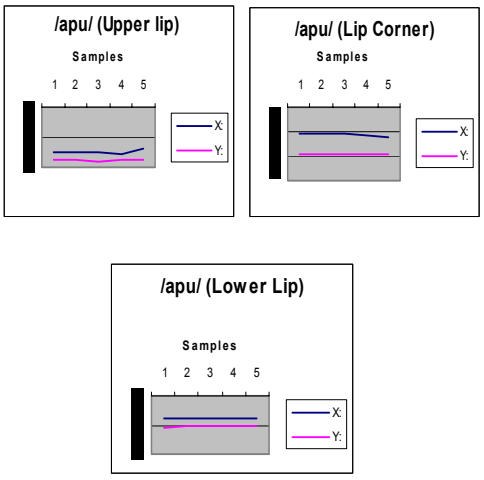

Figure 2 Summary results for Upper lip point, Lip corner point and Lower lip point for / apu /

To view the complete results of the research visit website: http://nu.edu.pk/99j/939/

5. DISCUSSION

We determined parameters for left half of the lip shape and it was assumed that parameters for right half of lip shape would be symmetrical. Parameters based upon other standards can also be determined using simple mathematical operations on MPEG-4 standard parameters (called feature points). This will allow facial animators using different standards to use our data for performing Urdu and other visual speech synthesis.

It was very difficult to judge the authenticity of the measured parameters from the graphs only, so in order to validate the results we developed a simulation of the lip shapes using OPEN GL. Screen shot of that tool is given in Figure 3.

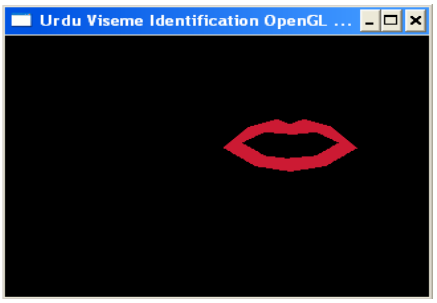

Figure 3 OpenGL simulation for / ɒpu/

The results of this animation were found to be close to the video recordings.

6. FUTURE ENHANCEMENTS

There are many ways in which we would like to enhance our work. For e.g., currently we have collected only VcV samples. This can be improved by collecting CvC samples and building a complete database of vowel to consonant and consonant to vowel transitions. This database can then be used for developing a facial

animation system for an Urdu text to speech system. Constructing such a database would also be helpful to study the effects of coarticulation in Urdu speech synthesis. Coarticulation is a phenomenon, which deals with the differences in articulation of a phoneme due to its context. Cohen and Massaro define Co articulation as "Changes in the articulation of speech segment depending on preceding and upcoming segments". For e.g., we can easily perceive that in two different words like 'tea' and 'two', the alphabet 't' maps to two different lip shapes. We can validate the measured samples by performing an audio-visual perception experiment. In this experiment deaf children, trained in lip-reading, would try to identify the phoneme strings by the lip shapes simulated on the monitor screen.

7. ACKNOWLEDGMENT

This study was supported by the **C**enter for **R**esearch in **U**rdu **L**anguage **P**rocessing (CRULP), National University. We would like to also like to acknowledge our teacher Dr. Sarmad Hussain and our colleagues Shakil Nasir, Muhammad Awais Anwar, Khwaja Umar Suleman, Sarmad Shabbir, Muhammad Sabtain and Muddassar Hameed who helped us in this research.

REFERENCES

-Breen, Dr. A. P, Bowers, Ms. E. and Welsh, Dr. W. "An Investigation into the generation of mouth shapes for a talking head".

-Cohen, Michael M., Beskow, Jonas, and Massaro, Dominic W. "Recent Developments in facial animation: An inside view". In *proceedings of auditory visual speech perception,* Pages 201—206. Terrigal-Sydney Australia, December, 1998.

-Cohen, Michael M. and Massaro, Dominic W.. "Modeling Coarticulation in synthetic visual speech". Tiddeman, Bernard and Perret, David. "Prototyping and transforming visemes for animated speech"

-Ezzat, Tony and Poggio, Tomaso. "Visual speech synthesis by morphing visemes". In *A.I. Memo No. 1658 C.B.C, L,* Paper No. 173, Artificial Intelligence Laboratory, M.I.T, May 1999.

-Jeffers, Janet. *Speech Reading*. Charles C Thomas Pub Ltd, June 1980.

-Möttönen, Riikka, Olivés, Jean-Luc, Kulja, Janne and Sams, Mikko. "Parameterized visual speech synthesis and its evaluation".

-Ostermann, Jörn. "Animation of Synthetic faces in MPEG-4". Computer Animation, pages. 49-51, Philadelphia, Pennsylvania, June 8-10, 1998.

-Waters, Keith and Levergood, Thomas M.. "DECface: An automatic lip-synchronization algorithm for synthetic faces". In *Technical report series,* CRL 93/4, Digital Equipment Corporation, Cambridge Research Lab, September 23, 1993.

APPENDIX A

Appendix A illustrates the mouth shapes with associated tri-visemes.

/ ▫pu /

/ ipæ /

/ æ pa /

APPENDIX B

Combined consonant and vowel speech reading movements in ideal viewing conditions

Visible	*Obscure*
1. Lower lip to upper teeth /f,v/	10. Lips rounded, moderate opening to lips back, narrow opening, /ɔɪ/
2. Lips relaxed, moderate opening to lips puckered, narrow opening, /ɑʊ/	11. Tongue up or down /t,d,n,l/
3. Lips puckered, narrow opening, /w,hw,r,u,ʊ,o,oʊ,ɝ/	12. Lips relaxed, moderate opening /ɛ,æ,ɑ/
4. Lips together, /p,b,m/	13. Lips relaxed, moderate opening to lips back narrow opening. /ɑɪ/
5. Tongue between teeth /θ,ð/	14. Tongue back and up /k,g,ŋ/
6. Lips forward, /ʃ,ʒ,tʃ,dʒ/	
7. Lips back, narrow opening /i,ɪ,eɪ,e,ʌ,j/	
8. Lips rounded, moderate opening, /ɔ/	
9. Teeth together, /s,z/	

/ æ tæ /

/ ukæ /

/ upi/