Evaluation of a computer aided 3D lip sync instructional model using virtual reality objects

A Rathinavelu[1], H Thiagarajan[2] and S R Savithri[3]

[1]Dr Mahalingam College of Engineering and Technology,
Pollachi-642003, TamilNadu, India,

[2]National Institute of Technology,
Trichy, TamilNadu, India

[3]All India Institute of Speech and Hearing,
Mysore, Karnataka, India.

[1]*starvee@yahoo.com*

ABSTRACT

Lip sync model is one of the aspects in computer facial animation. To create realistic lip sync model, a facial animation system needs extremely smooth lip motion with the deformation of the lips synchronized with the audio portion of speech. A deformable, parametric model of the lips was developed to achieve the desired effect. In order to create realistic speech animation, the articulatory modeling of the lip alone is insufficient. The other major articulators such as the tongue and jaw must also be considered. This lip sync model was initially developed by using polygonal model and blended key shape techniques and then parameterized by using 36 control points. The data for lip sync model was collected from video image and magnetic resonance imaging (MRI) techniques. The articulatory movements of our lip sync model were presented along with virtual reality (VR) objects in an interactive multimedia (IMM) interface. This IMM interface was used to teach small vocabulary of hearing impaired (HI) children. Virtual reality objects used to increase the cognitive process within HI children. The bilabial speech sounds were differentiated by using appropriate visual cues. Control panel was developed to present articulatory movements at different speed. For this study, six hearing impaired children were selected between the ages 4 and 7 and they were trained for 10 hours across 2 weeks on 18 meaningful words. The intelligibility of hearing impaired children was experimented to find out their performance in articulation and in memory retention. The results indicated that 65-75% of given words were articulated well and 75-85% of words were identified by all children.

1. INTRODUCTION

Once a child has a reasonable command of language and his phonetic and phonologic skills enable him to produce most speech patterns (including some consonant blends), it becomes possible to measure the intelligibility of his spontaneous speech [Ling, 1976]. An articulation disorder of hearing impaired children may be defined as incorrect production of speech sounds due to faculty placement, timing, direction, pressure, speed or integration of the movements of lips, tongue, velum or pharynx. A teacher/clinician should be able to evaluate the child's performance in articulation of speech sounds by employing suitable computer aided speech training system. It is assumed that the child's vocabulary and his knowledge of the lexical, morphological, syntactic and semantic rules which are essential to meaningful speech communication [Ling, 1976]. To keep track of the variety of individual children's specific speech training needs, teacher must spend lot of time to prepare the log file in the form of chart (or) progress card. In order to minimize the time involvement of teacher, we suggest computer aided sub skills for speech training needs of hearing impaired (HI) children. The speech of hearing impaired children differs from speech of Normal Hearing (NH) children in all aspects [Ling, 1989]. According to deaf researchers [Lundy, 2002], hearing children have consistently demonstrated the ability to perform such tasks between the ages of 4 and 5 years. But hearing impaired children are delayed by approx. 3 years in this cognitive developed milestone. It is also found that hearing impaired children are significantly delayed in the development of language skills. It has been known for some time that hearing impaired children can make use of speech reading for acquisition of words. To speech

read, the children must observe the teachers articulatory movements of the lips, jaw and the tongue. But many children fail to observe the articulatory movements of a teacher [Rathinavelu, 2003].

Due to auditory degradation of HI children, perceiving visual info is most useful informative for them. Computer Aided Articulation Tutor (CAAT) may be an alternative one for the hearing impaired children to acquire speech sounds and syllables by perceiving visual info without teacher's assistance [Rathinavelu, 2003]. Incorporating text and visual images of the vocabulary to be learned along with the actual definitions and spoken words, facilities learning and improves memory for the target vocabulary [Massaro & Light, 2004A]. According to empirical findings, children are good at producing spoken language if they do better at speech perception. When language and articulation disorders coexist, treating only one of them may produce some effect on the other, but it is likely that the effects will not be substantial; more research is needed [Pena-Brooks,2000]. Children who misarticulate may have additional problems in overall language skills. Our approach is a fairly new concept of using computer aided lip sync model using 3D VR objects to present articulatory position of complex speech sounds of Tamil language.

The lip sync model visualizes the articulator movements as described by the articulation parameters like lips, jaw and tongue. If gestures and speech express the same meaning, then gestures and speech should function as two sources of info to be integrated by the perceiver [Massaro, 1998]. Both visual and auditory cues are very important for children to acquire speech sounds. Previous studies [Massaro, 1998] indicated that visual stimuli are integrated into perception of speech. In Tamil, there is no exclusive study conducted in the construction of an articulatory model and there is no development of 3D articulatory-animated model for bilabial speech sounds and syllables. Few multimedia researchers [Massaro, 1998; Rathinavelu, 2001 and 2003; Barker, 2003] suggested about presenting new words graphically for hearing impaired children. So they are able to acquire much faster than conventional book form.

2. DESIGN OF LIP SYNC ARTICULATORY MODEL

While the auditory signal alone is adequate for communication, visual information from movements of the lips, tongue and jaws enhance intelligibility of the acoustic stimulus. Adding visible speech can often double the number of recognized words from a degraded auditory message [Massaro & Light, 2004B]. Visual pattern is particularly more effective in recognition of articulatory position of speech sounds for Hearing Impaired (HI) children. Articulatory process helps to convert the linguistic messages into sound. The most fundamental research issue of articulatory process is how to train the speech segments of a language. Polygonal model was chosen here to develop 3D lip sync model and animation so naturalistic.

Lips and tongue movements are the most powerful objects with several skeletons created for the realistic animation. Video clips of speakers articulating words were shot and the frames of this footage provided a guide for the corresponding frames of an animation [Lewis, 1991]. The lip sync model was constructed by digitizing the video image using Polygonal modeling. A 2D surface based coordinate grid was mapped onto the front and side images of a face. Point correspondences were established between the two images and the grid was reconstructed in 3D space [Parent, 2002]. 3D Animation involves three steps: Modeling, animation and rendering. An articulated model is a collection of many deformable objects connected together. The 3D lip sync model has been created with correspond to the visemes that constitute a word (refer figure.1).

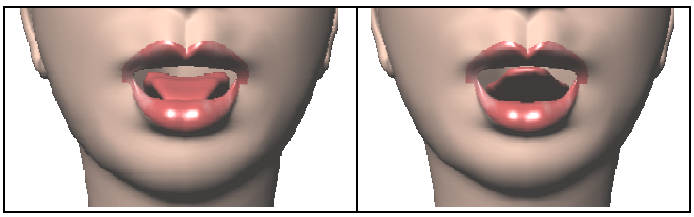

Figure 1. *Lip Sync model.*

These models serve as key frames for subsequent animation. Key framing allows fine control to ensure naturalistic facial movements of lip sync model. Interpolation between a finite set of visual targets was used to achieve speech articulation [Bailly, 2003]. Natural speech of speaker was then synchronized as realistic audio visual sequence. The articulation of voiced speech sounds were indicated by coloring the vocal cord portion of model. Table 1 shows about the parameters used to develop our lip sync model.

Table 1. *Parameters of lip sync model*

No	Articulatory Features	Remarks
1	Voiced or unvoiced	Graphically Highlighted
2	Opening and Closing of lips	Video image Data
3	Rounding of the lips	Video image Data
4	Tongue Position	MRI Data
5	Jaw Movement	Video + MRI Data

The tongue movements were built by animating tongue raise, tongue contact (with palate) and tongue curved. The data for tongue modeling was retrieved from MRI of natural talker [Rathinavelu]. The corresponding phonemes were matched with movements of mouth. In general, articulation of each given word is built from moving articulator's lips, jaw and tongue. In our model, lips and tongue articulators (including jaw) were integrated together to work as a single speech production model in an IMM interface (refer figure 3). The lip was viewed as being made up of 24 polygons each consisting of 5 vertices. The 36 control points, 12 from each of the outer, middle and inner lip contours constitute the vertices of the polygon. The specified polygons were triangulated and then rendered with either Material or Wire frame appearance. Suitable lighting and shading effects provided a realistic appearance of the lip. Our parameterized lip model was obtained by exporting the data from the Image based 3D model. Data collection, labeling and storing was a challenging task to develop this articulatory model (refer Table 2).

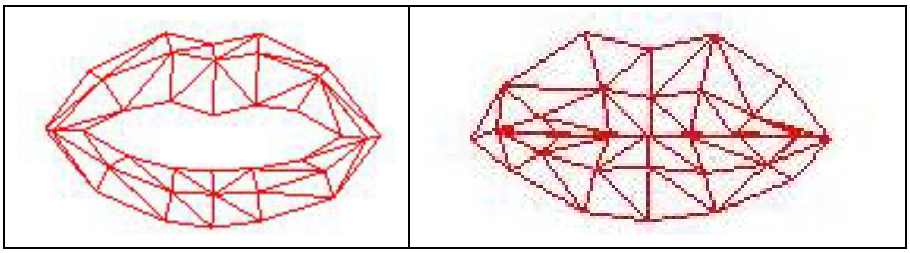

Figure 2. *Parameterized lip model for 'pa' and 'm'.*

The desired animation was obtained by specifying the target control points through a control panel interface. Using this parameterized model, the articulatory movement of the lips was modeled for bilabial Tamil speech sounds such as 'pa', 'mm' etc. as shown in figure 2.

Table 2: *Sample of data for speech sound 'pa' in the word 'palam (fruit)'.*

Point Number	**X Coordinate**	**Y Coordinate**	**Z Coordinate**
1	2.58766	–0.015768	0.04
2	1.93075	0.85171	1.11022e–016
3	1.08925	1.27246	1.11022e–016
4	0.5002	1.13782	1.11022e–016
5	–0.08885	1.28929	1.11022e–016
6	–1.04816	0.80122	1.11022e–016
7	–1.62247	–0.236976	–0.19
8	–0.962	–0.7148	0
9	–0.095	–1.091	0
10	1.095	–1.091	0
11	2.013	–0.794	0
12	1.80066	–0.435607	0.65

In order to create realistic speech animation, the articulatory modeling of the lip alone is insufficient. The other major articulators such as the tongue and jaws are also suitably modeled. The figure number.3 represents the articulation of the Tamil word 'palam'. The phonemes 'pa' and 'm' were modeled using the lips alone. The phoneme 'la' involved the tongue; hence the articulation of 'la' was modeled from MRI data of Speaker [Rathinavelu].

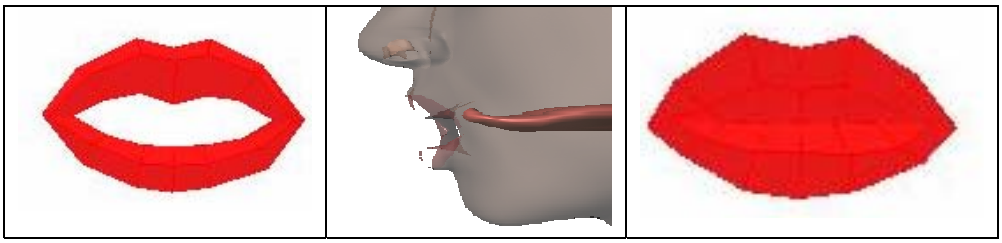

Figure 3. *Parameterized lip model along with tongue portion for the word 'pa-la-m'.*

Virtual reality object gives 360 degree view and more cognitive processes within children to retain newly learned activities [Rathinavelu, 2006]. The making of VR objects can be divided into the following phases: drafting, modeling, animation and exporting into VR Scene. Drafting software was used to draw an object from dimensions of the video image of real-world object and then polygonal modeler was used to model 3D view of the object. VR software helped to view the object three dimensionally after animating them suitably.

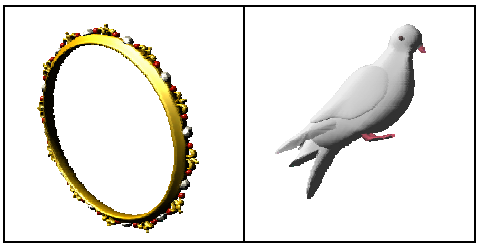

Figure 4. *Three dimensional Virtual Reality Objects.*

The interface of CAAT displays both the articulatory position of each word and 3D view of corresponding object together. Interactive Multimedia (IMM) software helped to reduce stress and strain of teachers who involved in content preparation and replicated teaching for specific speech sounds/syllables [Rathinavelu, 2006].

3. METHOD

The goal of our study was to examine the effectiveness of a Computer Aided Articulation Tutor (CAAT) to teach the articulatory position of speech segments for hearing impaired Children. Prior to scheduled training, teachers were involved to select the unknown but meaningful words under the categories of bilabial speech sound p, b, and m. The combination of consonants [p, b, and m] and vowels [a, e, I, o, and u] helped us to build 18 meaningful words for this computer aided articulation test. As shown in table 3, three sets of 6 words each were created. There were three stages in this study: an initial test, training for 2 weeks and a final test.

The training interface of CAAT was designed to combine lip sync model and Virtual Reality Objects to engage hearing impaired children effectively. Consonants have been traditionally described according to place, manner, and voicing dimensions. Place of articulation indicates where along the vocal tract the consonant is formed; manner of articulation indicates how it is formed; and voice indicates whether the vocal folds are vibrating during its production [Pena-brooks, 2000]. To avoid the confusion in perceiving the bilabial sounds 'p' and 'b', visual cues were used in IMM interface to differentiate 'p' and 'b'. For example, the visual cue 'palam' (fruit) was used for speech sound 'p' and the visual cue 'ball' was used for teaching syllables starting with 'b'. The Lip sync model was built-in with four modules as described below.

- Basic research module: A new concept was initiated to design and develop computer aided lip sync model using VR objects to help HI children in the acquisition of articulatory movements of bilabial speech sounds [p, b, and m] and syllables.
- Development module: Data of video image and MRI was used to develop 3D lip sync model. The lips were modeled using 36 control points distributed equally over three contours – outer, middle and inner. This module involves data collection, labeling and storing.
- Training module: 18 meaningful words were presented to 6 hearing impaired children for 10 hours over 2 weeks. Three boys and three girls were selected as subjects. Every day, they spent 30 min each in

beforenoon and afternoon sessions. Six words were selected in each category of bilabial speech sounds p, b, m.

- Evaluation module: After 2 weeks training, Children were asked to articulate the speech sounds and syllables on 10th day. The results were aimed to discuss about their performance in articulation and memory retention. In IMM interface, VR objects only were presented to identify the suitable/correct meaningful word from the list of words displayed.

[p] In its production the lips are closed and the soft palate is raised to close the nasal passage. When the lips are opened the air suddenly comes out with explosion. There is no vibration in the vocal cords. This sound may be described as a voiceless bilabial stop [Rajaram, 2000].

[b]. The movements of the speech organs are exactly the same as those for its corresponding voiceless variety[p] except for the vibration of the vocal cords. In Tamil it occurs initially in some of the borrowed words and medially after the nasal [m]. This may be described as a voiced bilabial stop [Rajaram, 2000].

[m]. In its production the lips are closed. The soft palate is lowered and the air stream comes freely through the nasal cavity. The vocal cords are vibrated during its production. In short, the articulatory movements for [m] are the same as for [b] except that the soft palate is lowered and the air is emitted through the nasal cavity. So the formation of this sound may be described shortly by defining it as a voiced bilabial nasal [Rajaram, 2000].

Table 3. *Meaningful words chosen for test.*

p	b	m
Pu-li (tiger)	Bim-bam (image)	Ma-ra-m(tree)
Pu-l(grass)	Kam-bam (post)	Mak-kal(people)
Pa-la-m (fruit)	Bak-ti (devotion)	Miin (fish)
Up-pu (salt)	Am-bu (arrow)	buu-mi (earth)
Kap-pal(ship)	Cem-bu (copper)	Paam-bu(snake)
Ap-paa(father)	Ba-la-m (strength)	Am-maa(mother)

4. EXPERIMENTS AND RESULTS

Our evaluation aimed to determine the degree to which the CAAT contributes to the acquisition and retention of meaningful words as well as student's response to our lip sync model. According to training schedule (refer table 4), 18 words were gradually presented to them during 2 weeks period. At that time, participants were asked to perceive the articulatory movements of Lip Sync Model to identify the speech segments of each word presented to them. Teachers and parents were instructed not to teach or use these words. During the training period of 2 weeks; it was still possible that the words could be learned outside of this training environment.

Table 4. *Training schedule.*

Day	**Weekly Schedule**	
	First week— Set of words	**Second week — Set of words**
1	1	3, 1
2	2	1, 2, 3
3	3	2, 3, 1
4	1, 2	3, 1, 2
5	2, 3	Test

During two weeks period of training, every child was trained by until they achieve maximum accuracy in articulation of each word and then only they were instructed to move to next word. No feedback was given to them during training period. Testing and training were carried out in a quiet room at oral deaf school, for about 60 min each day of a week for two weeks. After training period, students were instructed to articulate the words on 10th day. At that time, the lip sync face was removed from the interface of CAAT. Instead, 3D VR object and few similar titles of object were presented for each word. The children were instructed to

recognize, identify and articulate the words one by one in front of CAAT without any guidance from teachers. The performance of each participant was rated as good and poor (Ref. figure 5).

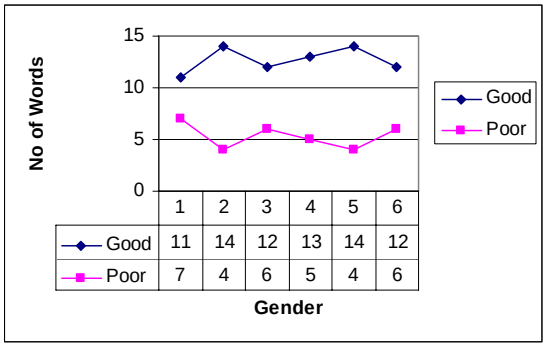

Figure 5. *Performance of student's articulation.*

Even after two weeks training, few words were not articulated properly by some of the hearing impaired children in the final test. For example, 'uppu' (salt) was substituted with 'o' and articulated as 'oppu'. This was due to similar kind of lip protrusion perceived by most of the children. Few of them didn't have any control over short vowels. For example, 'pu-li' (Tiger) was articulated as 'puu-li' and 'pu-l' (grass) was articulated as 'puu-l'. The Common articulatory errors were noted as detailed below.

Table 5. *Substituted speech sounds by few HI children.*

Substituted Speech sounds	Target Speech Sounds	Articulated Speech Sounds
/bi/-/pi/	Bi-m-bam	Pi-m-bam
/pa/-/ba/	Pa-lam	Ba-lam
/ba/-/pa/	Ba-k-ti	Pa-k-ti
/ra/-/la/	Ma-ra-m	Ma-la-m

Apart from these substitution errors, most of the children faced difficulty to articulate the words like 'ma-k-kal' (people) and 'buu-mi' (earth). It was commented by teachers that those words are meaningful but not familiar for hearing impaired children to use. This kind of incorrect articulation was considered as poor performance. This investigation also helped us in obtaining information about articulatory-acoustic performance of bilabial speech sounds and syllables.

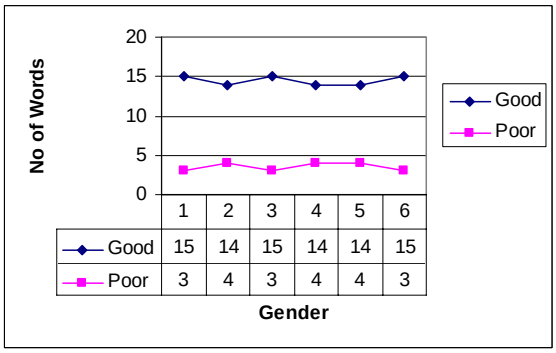

Figure 6. *Performance of student's memory retention.*

In memory retention test, all the children did well in identifying the VR object with suitable title of the particular VR object in the given list. As indicated earlier, some of the unfamiliar words were not retained well by most of the children. Those words were like 'mak-kal' (people), 'bim-bam' (image), 'pus-bam' (flower), 'Buu-mi' (earth). Rest of the words was identified by almost all the children. The individual articulatory performance of hearing impaired children is shown in figure 5 and their performance in memory retention of words is presented in figure. 6.

5. CONCLUSION

Our goal was to develop computer aided articulatory model to help Hearing impaired children in acquisition of speech sounds and syllable. Instead of teacher's role, computer can do tireless job by repeating the articulatory movements any number of times. It has been proved in our computer aided lip sync model that the visual information is most useful for the children with hearing loss in acquisition of speech sounds and syllables. Previous research findings suggested [Massaro, 1998] that hearing impaired children were able to response well when the appropriate visual cues (text, image and graphics) are presented to them. Our lip sync model integrates lips and tongue movements to articulate the given words. This model helped hearing impaired children to perceive the position and manner of each speech sounds of the word. Since co-articulation is an important issue with perception of speech segments with hearing impaired children, the control panel was used to show articulatory movements at different speed. Substitution and addition disorders were noted as most common errors during articulation of hearing impaired children. It was also noted that the software interface had really motivated and helped most of the children to correct their articulatory errors by them. Every child was given two weeks training and then tested with our lip sync model to identify their performance in articulation and memory retention. The results indicated that 65-75% of given words were articulated well and 75-85% of words were identified well by all children.

Acknowledgements: This research work is financially supported by Dr Mahalingam College of Engineering and Technology, Pollachi, South India. I wish to thank members of college management and faculty collegues of CSE and IT for their extended support.

6. REFERENCES

G Bailly et al (2003) , Audio Visual Speech Synthesis, International Journal of Speech Technology, Kluwer Academic Publishers, Netherlands, pp.331–346.

L J Barker (2003), Computer-Assisted Vocabulary Acquisition: The CSLU Vocabulary Tutor in Oral-Deaf education, Journal of Deaf Studies and Deaf Education, 8(2), spring 2003, pp.187–198.

J Lewis (1991), Automated Lip-Sync: Background and Techniques, The Journal of Visualization and Computer Animation, 2:118–122.

D Ling (1976), Speech and the hearing impaired child: Theory and Practice, Alexander Graham Bell Association for the deaf, Inc, U.S.A.

D Ling (1989), Foundations of spoken language for hearing impaired child, Alexander Graham Bell Association for the deaf, Inc, U.S.A.

J E B Lundy (2002), Age of language skills of deaf children in relation to theory mind development, Journal of deaf studies and deaf education 7:1, PP.41–56

D W Massaro (1998), Perceiving Talking Faces: From speech Perception to a Behavioral Principle, MIT Press, U.S.A

D W Massaro & J Light (2004A), Using visible speech for training perception and production of speech for hard of hearing individuals, Journal of Speech, Language, and Hearing Research, 47(2), pp.304–320.

D W Massaro & J Light (2004B), Improving the Vocabulary of children with hearing loss. Volta Review, 104(3), pp 141–174.

R Parent (2002), Computer Animation Algorithms and Techniques, Morgan Kaufmann Publishers, San Francisco.

A Pena-Brooks and M N Hegde (2000), Assessment and treatment of articulation and phonological disorders in children, Pro-Ed Publishers.

S Rajaram (2000) Tamil Phonetic Reader, CIIL, Mysore, India.

A Rathinavelu (2003), Early language learning for elementary students with hearing impairment in India, Masters' Thesis, School of Communications and Multimedia, Edith Cowan University, Australia.

A Rathinavelu, A Gowrishankar (2001), e-Learning for hearing impaired (p21.1–21.6), Proceedings of the Apple University Consortium Conference 2001 (ISBN 0-947209-33-6), James Cook University Townsville, Queensland Australia, Sept 23–26.

A Rathinavelu et al ,Interactive multimedia tool to help vocabulary learning of hearing impaired children by using 3D VR objects as visual cues, National Journal of Technology, (Accepted for publication).