Computer and Information Science; Vol. 8, No. 3; 2015
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Realistic Lip Syncing for Virtual Character Using Common Viseme Set

Itimad Raheem Ali[1], Ghazali Sulong[1] & Hoshang Kolivand[1]

[1] MaGIC-X (Media and Games Innovation Centre of Excellence, UTM-IRDA Digital Media Centre Universiti Teknologi Malaysia, 81310 Skudai, Johor Bahru, Malysia

Correspondence: Itimad Raheem Ali, Faculty of computing, University technology Malaysia, Johor Bahru. Tel: 006-0111-760-7616. E-mail: weffee@yahoo.com.

Received: May 15, 2015 Accepted: May 23, 2015 Online Published: July 31, 2015

doi:10.5539/cis.v8n3p71 URL: http://dx.doi.org/10.5539/cis.v8n3p71

The research is financed by (Sponsoring information)

Abstract

Speech is one of the most important interaction methods between the humans. Therefore, most of avatar researches focus on this area with significant attention. Creating animated speech requires a facial model capable of representing the myriad shapes the human face expressions during speech. Moreover, a method to produce the correct shape at the correct time is also in order. One of the main challenges is to create precise lip movements of the avatar and synchronize it with a recorded audio. This paper proposes a new lip synchronization algorithm for realistic applications, which can be employed to generate synchronized facial movements among the audio generated from natural speech or through a text-to-speech engine. This method requires an animator to construct animations using a canonical set of visemes for all pair wise combination of a reduced phoneme set. These animations are then stitched together smoothly to construct the final animation.

Keywords: Facial animation, lip syncing, phoneme, viseme, speech signal

1. Introduction

Communication is one of the most important aspects of human evolution. All other capabilities, developments, technologies created by our civilization were put together by this ability. In the past decades radio, television and more recently the internet technology influenced human communication and gave it new dimensions in all walks of life such as in education, transportation(Qureshi, K. N., & Abdullah, A. H., 2013), industries (Qureshi, K.N. and A.H. Abdullah 2014), etc. However, in the near future it is predicted that human life will be defined by virtual character communication. In human communication various organs of human body play a key role such as face, voice, body, and the social states. Among all of these communication gifts, human face and voice are the most fundamental ones. Essentially, the face is the most important part of the body through human's recognition possible. On the basis of recognizing a face humans are able to differentiate different faces, and are able to detect nonverbal facial expression in the face. These skills are developed in early age as the major channels of communication.

That is the main reason that virtual character makes attention towards the face. Synchronizing the lip and mouth movements naturally along with animation is an important part of convincing 3D character performance (Chuensaichol et al., 2011). This paper presents a simple method lip-synchronization method. The goal of this paper is to analyze the methodology used in lip syncing, to find an algorithm in order to automate this process and to implement it obtaining an animation tool, as part of the 3D Alias Wavefront Maya 4.0 package, useful to achieve faster and easier a good result in lip syncing. The target of this method is to avoid the animator, who normally can't waste time in understanding concepts such as phonemes, translating words, trying lip movements on a side mirror. This approach would make easy any step of the lip sync animation, leaving to the animator only to listen to the audio speech and enter the time of pronunciation and spoken words. The software section deals with the analysis of the lip syncing methodology following which correct lip synchronization can be reached. The study section covers the method implementation and its technical aspects, while the discussion section describes the user interface and explains how to use the approach and it collects the conclusions about the work

and obtained results.

2. Problem Background

For decades, the major problem in research has been synthesized of realistic visual speech animations corresponding to text or prerecorded acoustic speech input. A common method is to map one or more individual phonemes to corresponding viseme and generate the animation by interpolating the visemes given phoneme sequences. The demonstration of a lip sync algorithm for real-time applications along with animation is an important part of convincing 3D character performance. The simplest strategy is to make human stand in front of a mirror and analyze the lip positions while speaking in order to be able to generate the complex mouth movements. This sort of method to the lip syncing issue pretends a great work by animators, achieving results strongly affected by observation capabilities of the artist and losing much time. Because of these reasons it was necessary to focus on a precise methodology of lip syncing. Thus, a rigorous research study investigated into how written words correspond to a sequence of phonemes and how these are expressed with the same mouth shape and tongue position, finding in this way some fundamental steps of the lip syncing process.

3. Related Work

Facial animation can be done through 2D image based methods and 3D geometry-based methods. However, the focus of this paper is on lip syncing with audio, which is summarized below.

3.1 Visual Motion Speech Synthesis

The process of synchronizing input speech with synthesized facial motion is known as lip-syncing or speech motion synthesis. In speech synthesis input speech is normally represented by a standard speech unit called phoneme. The conversion from speech to phoneme can be done manually (Parke, F.I., 1982) or automatically (Yang et al., 2012). Phonemes are mapped to a set of lip motions, called Visemes. Hence visemes are closely correlated to phonemes. This relationship generates a look up table to produce smooth facial animation (Kalberer et al., 2002). Instead of using phonemes, Egges et al. (2003) used visyllables which are image capture of the mouth when each syllables are pronounced. The use of the sound and the mouth picture of a syllable are called co-articulation because it is the basis for synchronization of the lips and speech. Ten neighboring phonemes are simultaneously included in co-articulation (Cohen & Massaro, 1993).

English language contains 46 phonemes. Hence, implementation of face and lip movements animation require rapid conversion from phonemes to lip movement and vice versa, making simple lookup table not a practical solution. For better animation, co-articulation using a set of rules or by establishing an order of visual importance over the phonemes is implemented (Taylor, S.L. et al., 2012). According to Cao et al. (2005) animating 3D face model requires organization of data structure so that efficient location of appropriate movement is determined. To execute this, appropriate search algorithm such as the Support Vector Machine (SVM) for an automated detection of emotion of arbitrary input utterances is required. Conveying speech alongside gesture or expression is difficult, the question however is how can a realistic virtual human that incorporates emotions as well as gaze and speech behaviors be achieved. Describe a real time voice driven method using which a speaker's lip shape is synchronized with the corresponding speech signal for a low bandwidth mobile devices using sum of absolute difference (SAD) as vowel lip shape likelihood to cluster into categories then adjust the source and destination pictures of lip shape in the transparent level using alpha blending (Shih et al., 2010). The physics-based method uses the laws of physics and muscle forces to drive the motion of the face. Although, it is computationally expensive, however, it has been shown to be quite effective. Data-driven methods such as video rewrite (Bregler, C. et al., 1997). Use a phoneme segmented input speech signal to search within large databases of recorded motion and audio data for the closest matches.

Recorded closest match of motion with audio data are called audiovisual basis units (triphones). A complete record of triphones to simulate a real human being requires big database storage memory. Hence, in practice, animation audiovisual sequence is constructed by shortening the appropriate triphones from the database. In order to be useful, the method requires a large database of triphones, which leads to a scaling problem. Therefore, it is useful to develop compact statistical models for face motion. These models include Hidden-Markov Model (HMM) and Gaussian Mixture Models (GMM) (Taylor et al. 2012).

Vahid Asadpour et al. (2011) proposed a model-based feature extraction method which employs physiological characteristics of facial muscles producing lip movements. This proposal took into consideration the intrinsic properties of muscles such as viscosity, elasticity, and mass which are extracted from the dynamic lip model. The combination of audio and video features has been employed by adopting a multi-stream pseudo- synchronized HMM training method. Vahid Asadpour's model which uses physiological characteristics of facial muscles is

based on HMM and maps voice on to the face. The learning algorithm takes advantage of entropy minimization requiring that the system allocates high ranking to the velocity of facial features which have higher probability of happening and creates appropriate probability distribution over the different facial configurations (Qureshi, K.N. and A.H. Abdullah, 2013), as shown the result below in Figure 1.

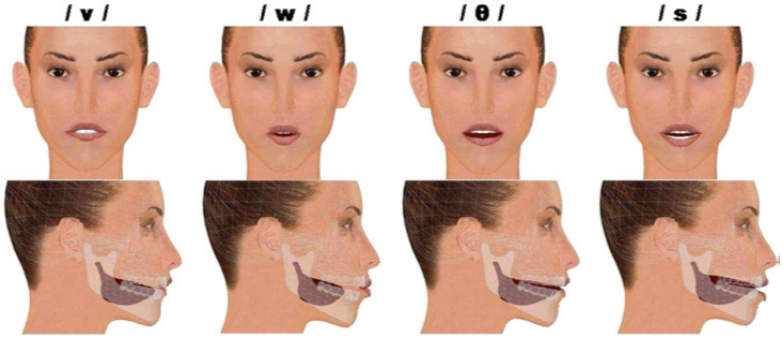

Figure 1. The 3D talking head presents for individual with facial view (Wang, L., et al., 2012)

Stevens et al. (2013) suggested that observation of the primary and secondary reaction times (RT) associated with manipulated audio- visual (AV) and auditory (A) is an effective evaluation to assess operational environments and ascertain cognitive load where multiple tasks are carried out simultaneously. This evaluation process was carried out by making adjustments to AV and A produced by text to speech (TTS) synthesis and by comparing the performance of a talking head and a human male. Xu et al. (2013) demonstrated a lip syncing algorithm for real time applications that can be used to generate synchronized facial movements with audio generated from natural speech or TTS engine. In another paper author Ma, X. and Z. Deng (2012) proposed a novel statistical model SAQP for automatically predict the quality of on-the-fly synthesized speech animations through many data-driven approaches.

3.2 Expressive Speech Synthesis

The eye gaze and lip synchronization is a necessary component in human communication, and thus plays a major role in the production of realistic conversations between human and virtual characters. The features in emotional facial expression are very important to capture the reality of the virtual characters. The voice carries much emotional information and the speech can be represented as a sequence of phones, each phone can be associated with a visual representation of the phoneme viseme. The animated visemes depend on the locality of the lip, jaw and tongue in a given phoneme. These techniques are popular for generating real time speech animation system that is able to project personality and interactive emotions. With the development of communication systems, it is now possible for people to interact directly with real time video devices. Such a communication system allows synchronization of the lip shape character with the corresponding speech in a real time voice driven mobile device designed by Shih et al. (2010).

The challenge is to maintain synchronization between the body movement and the expected normal human behavior. The eye movement is an important part of face to face conversation which carries the nonverbal information and emotional intent. Salvati et al. (2011) improved the animation scripting tool of Tatsuo Yotsukura et al. (2003) when she created (FSM) Face image Synthesis Module which is a general toolkit for building an easily customize embodied agent based on multimodal integrated dialog, speech synthesis, speech recognition and face image synthesis.

Virtual character is achieved by combining the eye gaze, lip shapes, and expressions. The lip movements and voice should be synchronized to provide realistic lip-synchronization animation. Normally higher-level module provides the synchronization between the two modules to obtain the capabilities used in spoken dialog. Most systems use a set of visemes that are activated by a text-to-speech engine (TTS). The TTS engine translates an utterance in text format into a series of phonemes. This technique is used to generate a realistic speech animation without having to manually set the positions for a set of visemes (Lee, C.Y. et al., 2011).

Serra et al. (2012) presented a visual speech animation module aimed at speeding up the dialog of virtual human and he assessed the quality of phonemes- to- viseme mappings devised for the English language. He discovered that the mechanism for lip synchronization could be carried out by decomposing the speech into a set of phonemes. These phonemes could then be represented as a set of visemes. The relationship between the phonemes in the signal and the visemes in the database is used to construct the appropriate lip shape. Cognitive

demand has a significant effect on the human hearing. On the talking head system, Athanasopoulos et al. (2011) proposed a talking head system using facial expressions which were generated from a Partial Differential Equation (PDE).

4. Method

The spoken word can have several meanings depending on the nuances given to the character, such as eye movement and facial expressions. The first step in the lip synch process is to understand the foundation, i.e. Phonemes and Visual Phonemes (Visemes). A closer inspection shows that the tongue is in different positions. The tongue may seem to be an insignificant element in lip synch, but it is very important to create truly realistic dialog. Now if the tongue movement is unnatural because too few visual phonemes were used, then the animation will look unrealistic. Large number of visual phonemes grantees a major detailed lip movement, because more phonemes have their own corresponding mouth position. Thus the same viseme, covering fewer phonemes, less frequently is used more than once within the same word avoiding the lips vibrating effect that this may cause.

Generally, few visual phonemes are preferred when the virtual character does not appear in the foreground during the animation, while accuracy is necessary for important characters making long speeches attracting the viewer attention. Creating more than ten visual phonemes requires more efforts of modeling not only to generate a greater number of mouth positions, but also because of the major request of precision. Visual phonemes are listed in Figure 2, divided into ten schemas.

Phoneme aa Phoneme b, m, p

Phoneme th, dh Phoneme ch, j, sh

Phoneme er, r Phoneme k,g

Phoneme v, f Phoneme n, l, t, d

Phoneme ow, oh, w, uh Phoneme eeh, ih, ey, ae, s, z

Figure 2. Short list of 10-based visual phonemes

In the list with ten elements, several of the visual phonemes with similar exterior appearances have been combined. In fact, once the spoken word is translated into its phoneme sequence it is just a question of selecting the corresponding visual phonemes according to the adopted schema, without particular calculations. The proposed method consist from three basic steps, the first step consists of creating a correspondence between phonemes and visemes, while the second step covers the phase of entering data concerning the speech to be synchronized, and the third step focusing on the execution of the algorithm of translation from text to phoneme and computes to set the blendshapes keys at the right time and value.

4.1 The Lip Sync Process

Using the phoneme chart to translate what is heard in the dialog into their audible phonemes, then translating those phonemes into a predetermined set of visual phonemes are the fundamental steps of lip synch process.

The steps of animating lip syncing are summarized in following.

- Determine the speech pattern.
- Analyze the audible dialog and enter the phonemes into the timing chart.
- Use the timing chart to set the key frames in the animation.
- Test the animation for synching and getting the animation where necessary.

4.1.1 Determine the Speech Pattern

The first step in lip synch is to determine the speech pattern of the dialog. Because of the use of abbreviations and dialect influences in current speaking. This is an important element, since it is necessary to assign phonemes to the actual heard sound, not the viewed text. For example, the word “every” can often been spoken in a contract manner as “evry”.

4.1.2 Analyze the Audible Dialog to Determine Phonemes

The first thing to do is to load the audio file and allows to identifying the actual times when sounds occur and relate them in frames. Another important feature is the presence of a scrub tool that lets to drag the audio forward and backward in order to obtain a bigger precision. After loading the audio file, usually the lip sync methodology suggests to hear the dialog and write down the phonemes, but it was understood that was necessary to follow the strategy to write down the real spoken words, with their abbreviations or contractions, and not the phonemes itself because the algorithm will make this translation. After this, scrubbing the sound back and forth it is possible to determine the exact location of each sound in the audio file. Using an audio tool thus is useful because it provides a visual representation of the sound too, which help to determine the point where words, and so phonemes, are recorded.

4.1.3 Use Timing Table to Set Frames

Once the sound has been translated into their visual phonemes and the frames identified for each, it is necessary to morph the visual phonemes from one to another at the appropriate frame in the animation. There are two types of morphing, straight or weighted. Straight morphing simply morphs the object in a linear progression from one object to another. The morph can be any value from 0 and 100%. The only issue is that it is limited to a single morph object. On the other hand, weighted morphing allows blending multiple objects in a single morph. This is very useful when adding facial expressions and emotions to the lip synch animation in the future.

4.1.4 Getting the Finished Animation

Lip-syncing is not an exact science and it will almost always require a little tweaking. This may involve altering

certain poses or tinkering with the timing. In some cases, the lip sync may appear more realistic if it is a frame so that the mouth is moving as the viewer's brain is processing the sound rather than when their ear is receiving it. These are subtle things, but lip syncing as an exercise in subtlety. If the aim is making the audience to suspend disbelief and accept the virtual character as a living being, it is important not to stop until it actually seems as if the virtual character is speaking the dialog. Not all letters are pronounced in normal speaking, particularly if an accent is present.

4.1.5 Phoneme Dropping Guidelines

After obtaining the animation, often is necessary to drop some phonemes, but there are some rules to be observed. Never drop a phoneme in the beginning of a word. It is possible to drop a consonant at the end of words but never at the beginning because it will change the visual phonetic pronunciation for the word. Otherwise the consonant at the end of words can be dropped without having much impact on the word. Such as m, n, p letters have often to be dropped out, but this not necessary means that the artist has to physically delete those keys from the timelines. The letters "t" or "d" are not to be deleted, but that are very little persistent during the pronunciation, so sometimes they can be substituted or erased, the animator can shift keys to assign them fewer frames than that associated by the algorithm.

4.1.6 Synchronization Time and Artistic Freedom

Once the algorithm has generated the phoneme sequences, these are to be set at the right time within the range of the word's pronunciation. This final step in lip synch is important as the previous ones. It remains to be considered that an algorithm based on precise calculation could never reflect the complexity and the random way humans speak. The purpose is to obtain a quite perfect synchronization between lip movement and audio without any other action than entering in the user interface the spoken word and start and end times. This allows to obtain a good result spending a little time and doing a reduced work, but without affecting freedom and creativity of the artist who can later characterize in a more detailed way his speaking creature. Figure 3 shows the methodology of the proposed method.

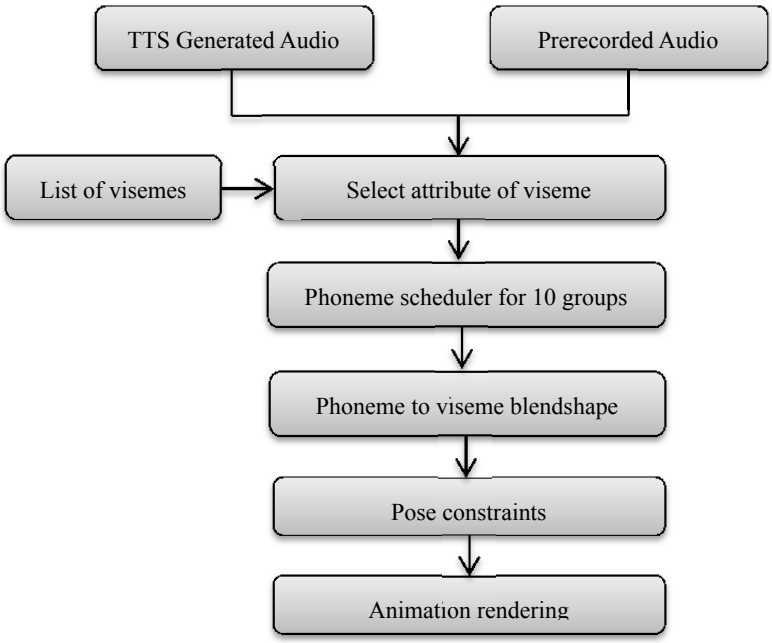

Figure 3. The methodology of the proposed method

5. Result

There is a need to make all the lip sync processes are the easiest one and the algorithm reflects this issue. In fact, with its user interface divided into the five steps, each one covered by a specific tab that contains all controls for that purpose, follows the structure of the workflow.

5.1 Creating Data Structures

The goal of this step is to collect fundamental information on the scene and to initialize variables used by the text to phoneme translation algorithm. Created using MEL, it is used by the real algorithm and its content is

displayed in a scrolling list. Figure 4 illustrates building the data structure of phonemes and visemes.

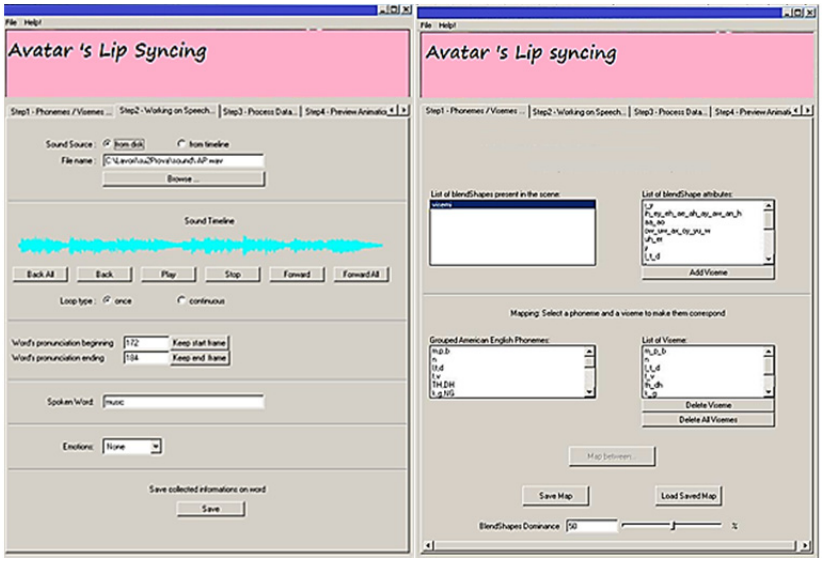

Figure 4. The first and the second steps in the user interface of the proposed method

Figure 4 shows the first and second step toward in the user interface of the proposed method. The Figure shows the parameters buttons and pop up menu as well.

5.2 Collecting Information

The collecting information step focuses on the analysis of the audio track to collect information on speech. In fact, the purpose to write down the phoneme sequence of the speech was considered as an impracticable one by an animator. Understanding phonemes while listening to a recorded audio it is not a simple task to perform. On the basis of these arguments, the idea was not to write down phonemes but the real spoken word that is understandable, as shown in Figure 5.

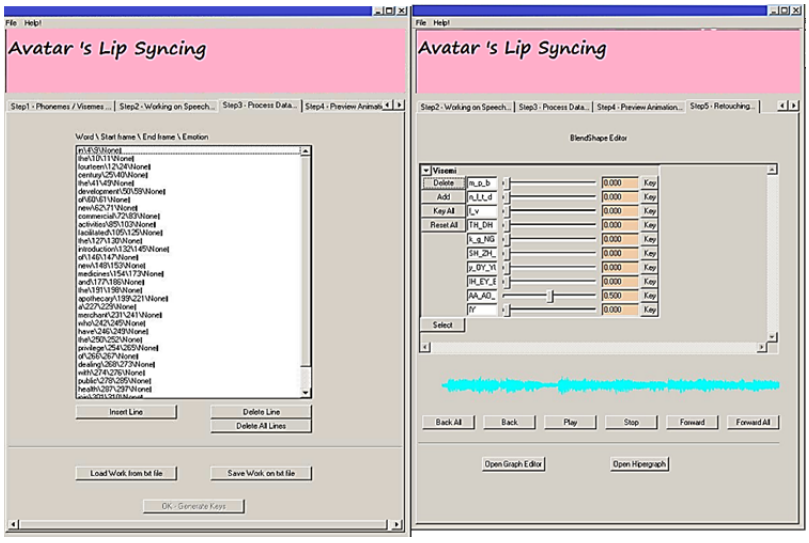

Figure 5. The third and fifth steps in the user interface of the proposed method

Thinking about the syncing of the first word in a speech, it is evident that any visemes value has to be keyed at a zero value because of the interpolation of Maya itself. In fact, Maya automatically interpolates during the animation when the same attribute has a key at different values in different frames. The gap between the two times is covered by Maya itself that interpolates the value of the first key to the second. Thus, for the first word of a speech or a word pronounced after a pause, it is important to set all the keys for the visemes at zero, but for

all other words during the speech it could be very dangerous to perform this action.

5.3 Discussion

To obtain lip syncing, it was first written the talking text because this was the most important element in such animation that wants to give prominence to the synchronism of the lips during the speaking. Moreover, beside the flat text, many annotations were added to characterize the manner of speaking, as facial expressions and emotions, or movements of the head and eyes.

With this information, the audio track was recorded in a special structure for professional audio recording, and then the process of the lip syncing was started. Having the written text avoided the phase of writing down the spoken words, and the speech was listened many times to write down the beginning and ending frame of each word moving through the sound waveform. It is possible to see the plain text that was spoken by the speaker, and then the text that was introduced, word by word:

"Hi, my name is Rosa. I can speak every language based on English grammar, thank you and thanks to the lip syncing system used Maya environment"

In order to obtain a correct lip synch, some of these words were linked together or modified according to the pronunciation. Below, there are the entered words:

"Hi myname's Rosa, I canspeak ev'ry 'aguage based o'Eglishgrama thankyou and thanks to the 'ip syncing use' Maya environmen' "

As it can be seen, many words are linked together to create an unique word, while others have less letters or have some apostrophe substituting some letters and moreover some other words have a single letter to create the sound instead of the correct letters, such as the "u" role in the "thank you" sound. All this tricks are really important because they allow with little work and attention to avoid to the animator to make changes and correction after the lip synch process editing keys time or value. In fact, to realize the demo animation, the animator just simply edited the value of the blendshape in the key corresponding to the visual phoneme "Big Oh", in order to obtain a more opened mouth to suggest a more expressive exclamation. No other interventions were necessary to have a realistic lip syncing animation. The evaluation was executed using 40 persons (between 20 and 65 years old) at a multimedia systems lab at Universiti Technologi Malaysia (UTM). All of them did not have any problem in the vision or hearing, but three of them had expertise in this area. The comparison of the preferences for three sentences during the experiment is shown in Figure 6. These sentences as follows:

S1: "Communication is an important to convey the ideas".

S2: "Messi plays football with Barchelona team".

S3: "The Sea today is so stormy".

The evaluation of the proposed method was carried out using these sentences. Each sentence was animated using the phoneme-to-viseme mapping in Figure 2. The animations of the virtual character were shown two times to the persons who then filled out a questionnaire in the appendix A.

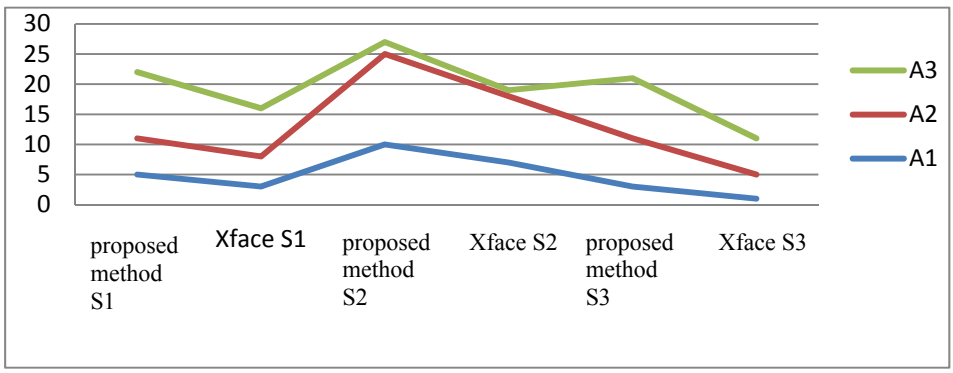

Figure 6. Comparison preferences for three sentences between proposed method and Xface.

As shown in Figure 6, the persons slightly preferred the decreasing in mapping for S1 and S3 but strongly to S2. The proposed method particularly is favored more than Xface (Balci, K., 2007). We can deduce that the differences of visemes classes had little importance for the quality of S1 and S3, but had a positive effect on the quality of S2. That means some sentences have a small animation effect without changing on the quality.

Corresponding to figure 6, the subjects had to choose one of the following alternatives for each sentence:

A1: Strongly the animation is clear.
A2: A bit the animation is clear.
A3: Neutral.

Actually, most of the preferences are focused on the more neutral alternatives A2 and A3 greatest in case of S1 (1.3), S2 (2.40) and S3 (2) respectively. A realistic estimation of approach performances leads to the following judgment about syncing a 15 second of audio: about 30 minutes are necessary to the artist to recognize spoken words and their start and end times; about 5 minutes of CPU time are required to execute method working on a complex facial model and using the 16 visemes based schema; About 30 minutes to execute again method on different timing chart and word data structure. In a little more than one hour, the animator can synchronize successfully 15 seconds of speech, having the possibility of improving the animation testing it in different configuration, thus he obtains a final result.

This means that in about four hours of work, an entire minute of speech is synchronized and so 2 minutes of animation are finished in a day of work. These are very meaningful values, because they imply the possibility of synching a movie of standard length of 90 minutes in about one month and a half of work of a single animator, while it could be a shorter time if many animators work on it. The speakable virtual character is shown in Figure 7.

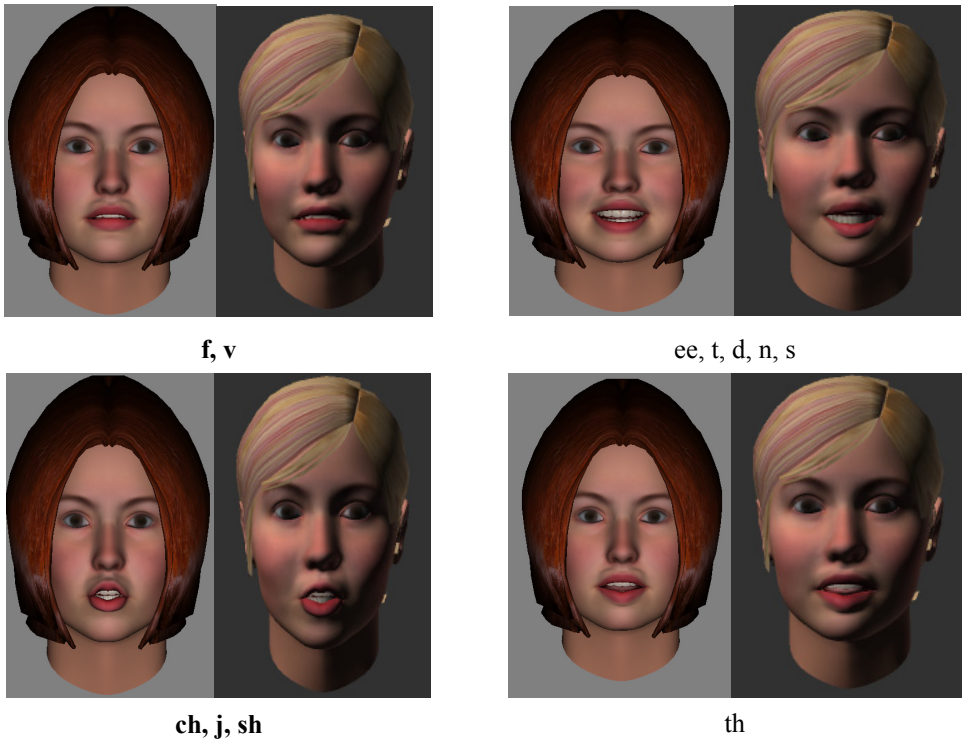

Figure 7. Phonemes synchronization results for the proposed method comparing with Xface (Balci, K., 2007).

6. Conclusion

Creating animated speech requires a facial model capable of representing the myriad shapes the human face expressions during speech. Moreover, a method to produce the correct shape at the correct time is also in order. One of the main challenges is to create precise lip movements of the avatar and synchronize it with a recorded audio. The proposed lip synchronization method can be employed to generate synchronized facial movements among the audio generated through a TTS engine. This method requires constructed animations using a canonical set of visemes for all pair wise combination of a reduced phoneme set. These animations are smoothed stitched together to construct the final animation. This method has achieved good results considered unexpected at the beginning of the work. In fact, the obtained lip synchronization very satisfactory, because it is perfectly realistic and rarely needs an intervention from the animator in order to improve some movements. The evaluation was executed using 40 subjects at a multimedia systems lab at Universiti Technologi Malaysia (UTM) by the questionnaire showed in Appendix A.

A future development foresees surely the inclusion of the option of saving all text files in a path chosen by the user instead of a default one, and the possibility of implying different blendshape dominance for any single word instead of using one dominance for all the speech. Moreover, to allow the characterization of a model with emotions, it would be projected the use of a special field named "emotions", that is already present in the user interface, in order to set keys for other blendshapes corresponding to the positions of the eyebrows, eyes and eyelids.

References

Asadpour, V., Homayounpour, M. M., & Towhidkhah, F. (2011). Audio–visual speaker identification using dynamic facial movements and utterance phonetic content. *Applied Soft Computing, 11*(2), 2083-2093.

Athanasopoulos, M., Ugail, H., & Castro, G. G. (2011). On the development of a talking head system based on the use of PDE-based parametic surfaces. InTransactions on computational science XII (pp. 56-77). Springer Berlin Heidelberg.

Balci, K. (2007). Xface: MPEG-4 based open source toolkit for 3d facial animation. in Proceedings of the working conference on Advanced visual interfaces. ACM.

Bregler, C., M. Covell, & M. Slaney. (1997). Video rewrite: Driving visual speech with audio. in Proceedings of

the 24th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co.

Cao, Y., et al. (2005). Expressive speech-driven facial animation. ACM Transactions on Graphics (TOG). 24(4): p. 1283-1302.

Chuensaichol, T., P. Kanongchaiyos, & Wutiwiwatchai, C. (2011). Lip synchronization from Thai speech. in Proceedings of the 10th International Conference on Virtual Reality Continuum and Its Applications in Industry.ACM.

Cohen, M. M., & Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech, in Models and techniques in computer animation, *Springer*. p. 139-156.

Egges, A., S. Kshirsagar, and N. Magnenat-Thalmann. (2003). A model for personality and emotion simulation. in Knowledge-based intelligent information and engineering systems. Springer.

Kalberer, G. A., Müller, P., & L. J. Van Gool, (2002). Speech Animation Using Viseme Space. in VMV.

Lee, C. Y., Lee, S., & Chin, S. (2011). Multilayer structural wound synthesis on 3D face. Computer Animation and Virtual Worlds, 22(2‐3): p. 177-185.

Ma, X., & Deng, Z. (2012). A statistical quality model for data-driven speech animation. *Visualization and Computer Graphics,* IEEE Transactions on, *18*(11), 1915-1927.

Parke, F. I. (1982), Parameterized models for facial animation. *Computer Graphics and Applications,* IEEE, 1982, *2*(9), 61-68.

Qureshi, K. N., & Abdullah, A. H. (2013). A survey on intelligent transportation systems. *Middle-East Journal of Scientific Research, 15*(5), 629-642.

Qureshi, K. N. & Abdullah, A. H. (2014). Adaptation of Wireless Sensor Network in Industries and Their Architecture, Standards and Applications. *World Applied Sciences Journal, 30*(10), 1218-1223.

Salvati, M., et al. (2011). Developing tools for 2D/3D conversion of Japanese animations. in ACM SIGGRAPH 2011 Talks. ACM.

Serra, J., Ribeiro, M., Freitas, J., Orvalho, V., & Dias, M. S. (2012). A proposal for a visual speech animation system for European Portuguese. In Advances in Speech and Language Technologies for Iberian Languages (pp. 267-276). Springer Berlin Heidelberg.

Shih, P. Y., Wang, J. F., & Chen, Z. Y. (2010). Kernel-Based lip shape clustering with phoneme recognition for real-time voice driven talking face, in Advances in Neural Networks-ISNN 2010, Springer. p. 516-523.

Stevens, C. J., et al. (2013). Evaluating a synthetic talking head using a dual task: Modality effects on speech understanding and cognitive load. International Journal of Human-Computer Studies,71(4): p. 440-454.

Taylor, S. L., et al. (2012). Dynamic units of visual speech. in Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation. Eurographics Association

Wang, L., et al. (2012). Phoneme-level articulatory animation in pronunciation training. *Speech Communication, 54*(7), 845-856.

Xu, Y., et al. (2013). A practical and configurable lip sync method for games. in Proceedings of the Motion on Games. ACM.

Yang, M., et al. (2012). A multimodal approach of generating 3D human-like talking agent. *Journal on Multimodal User Interfaces, 5*(1-2), 61-68.

Yotsukura, T., Morishima, S., & Nakamura, S. (2003). Model-based talking face synthesis for anthropomorphic spoken dialog agent system. InProceedings of the eleventh ACM international conference on Multimedia (pp. 351-354). ACM.

Appendix A

The follow-up questionnaire with the testers' opinions.

Personal information.

Sex: Male / Female

Age:

Questions around the test that which was carried out:

1. How natural did you feel the integration of the virtual character was?

Very unnatural				Very natural
1	2	3	4	5

2. Did you feel freer to interact with the virtual character that is fixed on a desktop computer?

Not at all				A lot more free
1	2	3	4	5

3. Did you feel the lip syncing is conformable to the entered sentence?

Not at all				A lot more
1	2	3	4	5

4. Does the lip syncing look realistic?

Not at all				A lot more
1	2	3	4	5