Jurnal
Teknologi

Full Paper

FRAMEWORK DEVELOPMENT OF REAL-TIME LIP SYNC ANIMATION ON VISEME BASED HUMAN SPEECH

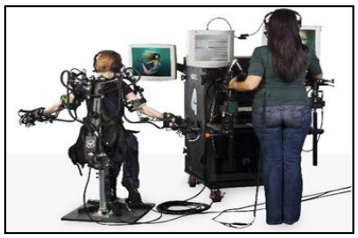
Loh Ngiik Hoon[a], Khairul Aidil Azlin Abd. Rahman[a*], Wang Yin Chai[b]

[a]Faculty of Applied and Creative Arts Universiti Malaysia Sarawak, Kota Samarahan, 94300 Sarawak Malaysia
[b]Faculty of Science Computer and Information Technology Universiti Malaysia Sarawak, Kota Samarahan, 94300 Sarawak Malaysia

Article history
Received
3 December 2013
Received in revised form
2 July 2014
Accepted
25 November 2014

*Corresponding author
azlin@indi.unimas.my

Graphical abstract



Abstract

Performance of real-time lip sync animation is an approach to perform a virtual computer generated character talk, which synchronizes an accurate lip movement and sound in live. Based on the review, the creation of lip sync animation in real-time is particularly challenging in mapping the lip animation movement and sounds that are synchronized. The fluidity and accuracy in natural speech are one of the most difficult things to do convincingly in facial animation. People are very sensitive to this when you get it wrong because we are all focused on faces. Especially in real time application, the visual impact needed is immediate, commanding and convincing to the audience. A research on viseme based human speech was conducted to develop a lip synchronization platform in order to achieve an accurate lip motion with the sounds that are synchronized as well as increase the visual performance of the facial animation. Through this research, a usability automated digital speech system for lip sync animation was developed. Automatic designed with the use of simple synchronization tricks which generally improve accuracy and realistic visual impression and implementation of advanced features into lip synchronization application. This study allows simulation of lip synching in real time and offline application. Hence, it can be applied in various areas such as entertainment, education, tutoring, animation and live performances, such as theater, broadcasting, education and live presentation.

Keywords: Lip synchronization animation, real time, human speech recognition

1.0 INTRODUCTION

Lip sync (short for lip synchronization) in animation is the art of making the animated character appear to speak by matching the mouth-movements to the phonemes from an audio track. The lip sync technique continues to this day, with animated films and television broadcasts which are one of an essential stage in the animation production. In real time, lip synchronization is an approach to perform a virtual computer generated character to talk, which synchronizes an accurate lip movement and speech signal in live. According to Huang and Chen (1998), real time lip sync animation is a technique driven by human voice directly for synthesizing the mouth movement from acoustic speech information. It also considers as human-to-computer interaction interfaces. Consequently, there has been a large amount of research on incorporating bimodality of speech into human-computer interaction interfaces. Lip sync is one of the research topics in this area. However, based on the review, identifying each and every sound in a word or phrase in lip sync animation can create very busy and over-animated mouth action (Spevack, 2011). Conversely, the less of consideration, as a result every mouth shape was emphasised equally for each word, which led to frantic that confuses and tires the viewer. Thus, the

animation is failure in matching the naturalistic of facial model (Beiman, 2010). Therefore, the goal of this study is to animate the face of a speaking avatar in such a way that it realistically pronounces the text based on the process called lip synchronization. For a realistic result, lip movements and its pattern must be perfectly synchronized with the audio. Other than the lip sync, realistic face animation includes facial expressions and emotions is not in the scope of this paper.

2.0 LIP SYNC TECHNIQUES

With the vast increase in computing power nowadays, the qualities of animation enable an extra layer of visually convincing realism. According to Hofer, Yamagishi, and Shimodaira (2008), in order to make character animation believable, correct lip sync is essential. However, as claimed by Parent, King and Fujimura (2002), to make an accurate lip sync animation is complex and challenging. The difficult of Lip sync technique involves figuring out the timings of the speech as well as the actual animating of the mouth movement to match the soundtrack. On the other hand, the inaccurate lip sync animation will give an unnatural facial animation and failure to generate realistic looking animations. According to Rathinavelu, Thiagarajan, and Savithri (2006), movement of the lips is one of an important component of facial animation during speech. Lip movements used in speech make the character seem alive, and the dialogue delivery develops the personality of the character. Hence, for convincing animation, the study of the realistic lip sync in animation is needed by adding the quality and believability to generate realistic looking in an animation.

In order to achieve frame-accurate synchronization of sound and mouth-movements in animation, lip sync have traditionally been handled in several ways. According to Lewis (1991), the rotoscoping approach is one of the techniques used in animation to obtain the realistic movement. With this technique, mouth motion and general character movement are obtained by using the live-action footage of actors performing the desired motion. The frames of this footage provide a guide to the animators for the corresponding frames of an animation. Apart from that, another method is the animator trying to mime the animated mouth position accurately synchronized to the soundtrack. Based on Blair (1994), the animator creates an illusion of speech or believable image that is based on reality. An animator analysis real mouth action from his own mouth action, phonetic science and the pronunciation guides in the dictionary. Animation handbooks often have tables illustrating the mouth positions corresponding to a small number of key sounds. The dialogue chart of various mouth shapes is shown as Figure 1:

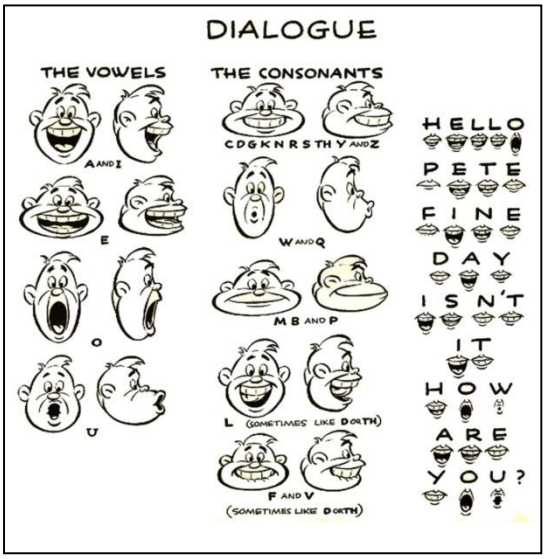


Figure 1 Dialogue chart by Blair (1994)

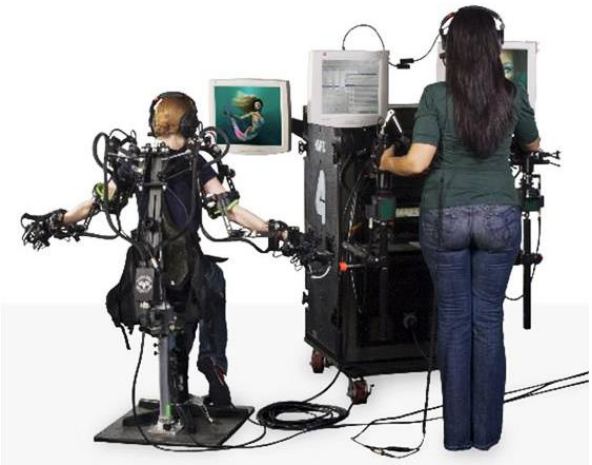


Figure 2 Motion capture system by The Jim Henson Company (2008)

The animated segment the pronunciation guide to create visually convincing mouth movement. This approach neglects non-vowel sounds because vowel sounds correspond to visually distinctive mouth positions and are typically of greater duration than non-vowel sounds. Consequently, the lip sync produced using this approach is often satisfactory but is generally not realistic. Besides, motion capture is also one of the advance technology for real-time capture system that make applying motion data to characters to talk with the use of markers in the real world. It is a method very widely used in the professional world but carries a high price tag (Scott, 2010). Based on (Dyer, Martin & Zulauf, 1995), motion capture builds and use various devices for detecting input data for recording the movement of humans or animals and reproducing it on a virtual character model.

Budiman, Bennamoun, & Huynh (2005) have mentioned that despite the advantage of the technology, optical motion capture technologies have been known to be expensive. The high cost is mainly contributed by the cost of hardware components such as high speed cameras. This is available only for high profile events, television broadcasts and films. Therefore, the automated digital speech system is developed in this research for performing lip sync animation in real time application, which the animation can be addressed to save more time, reduce the workloads and costing as well as enable output to produce faster and ensure a realistic result.

3.0 SPEECH IN ANIMATION

A character's speech in an animation is not only depends on acoustic cues, but also on visual cues such as lip movement and facial expressions. According to Zoric (2005), the key feature points of facial components which will influence the facial animation are eyes, eyebrow, cheeks, lip, nose, ears, chin and jaw. According to Zoric and Pandzic (2005), position of the mouth and tongue must be related to characteristics of the speech signal in order to make lip synchronization possible. It is necessary to determine the human speech behavior for mapping a natural speech to the lip shape animation in the real time. For the lip movement, they involve the consideration on jaw, chin, inner lower lip, cornerlips, midlip. However, the primary facial expressions are joy, sadness, anger, fear, disgust and surprise. The combination of both auditory and visual speech recognition is more accurate than auditory or visual only.

Besides, Hofer *et al*. (2008) has stated that, in animation, movement of the mouth during speech is one of an essential key component of facial animation. A character with the ability to portray correct lip sync will be resulting a sense of natural and convincing animations. Besides, in order to generate speech animation in real time, the determination of the motion of the mouth and tongue during speech is important. Based on Rathinavelu, Thiagarajan and Savithri (2006), the articulatory modeling of the lip alone is insufficient to create realistic speech animation. Tongue and jaw also need to be considered. The movements of lips and tongue are the most powerful objects created for the realistic animation with skeletons. On the other hand, according to Tang, Liew and Yan (2004), simulating a talking animation has to control the proper muscles from different phonemes. The muscles are conducted to generate more detailed mouth shapes to use for extracting lip parameter. Based on the review, the human speaks behavior has to consider several factors, they include lip pattern, facial expressions, jaw, chin, tongue, face's muscle and speech signal. These factors are essential for the stage in visual mapping when the speech is analyzed and classified into viseme categories. The calculated visemes are used for animation of virtual avatar's face.

4.0 SPEECH, PHONEME AND VISEME

Speech is the vocalized form of human communication to express thoughts and feelings by articulate sounds. A different position of the mouth or lip pattern and tongue will give the difference of intonation characteristics of speech and determine the phoneme. Independently of intonation characteristics of speech and determine the phoneme. A phoneme is the perceptually distinct units of sound in a specified language that distinguish one word from another. For example, when we speak the word '/bed/' for the vowel 'e' sound (Figure 3), our mouth seems to be slightly open (Figure 4).

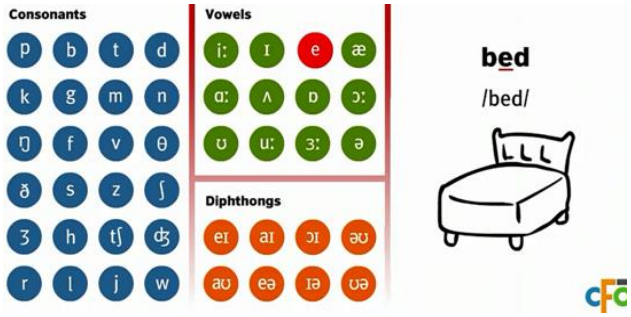


Figure 3 The Word '/bed/' For the Vowel 'e' Sound. http://www.cambridgeenglishonline.com

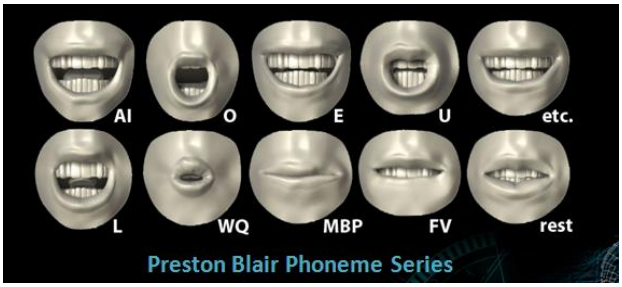


Figure 4 Preston Blair Phoneme Series (Gary C. Martin CGI) http://www.garycmartin.com/mouth_shapes.html

Phoneme also can be described as a basic acoustic unit of speech perception and the visual representation of phoneme is called viseme. There are many sounds that are visually ambiguous when pronounced. Therefore, there is a many-to-one mapping between phonemes and visemes.

Apart from that, based on Frank, Hoch and Trogemann (1997), visemes is an approach to the basic animation parameters in estimating visual similarities between different phonemes. Hence, for different phonemes, visually equal mouth positions are collected into classes of visual phonemes, and used as animation parameters to get the keyframes

for all possible mouth positions. Therefore, it is important to classify them into viseme categories and used for virtual avatar's face in the animation.

5.0 SPEECH SIGNAL

The automatic lip sync can be assumed as a typical speech communications application, which is one kind of the speech recognition. According to Zoric (2005), Speech recognition is a technology that allows the computer to identify and understand words spoken by a person by using a communication device (Figure 5). It deals with analysis of the linguistic content of a speech signal and its conversion into a computer-readable format. Hence, the information in the speech signal is important to categorize the class of the phoneme. Likewise, this is a necessary step in the process of mapping the speech to lip movements.

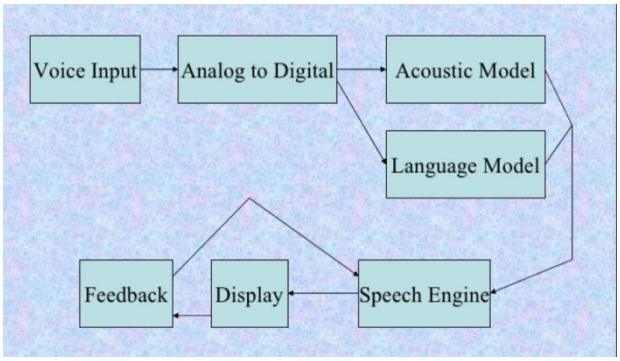


Figure 5 Speech recognition by Joshi (2009)

In order to convert a speech signal to the lip shape information, the position of the mouth and tongue must be related to the characteristics of the speech signal. Based on Lewis and Parke (1990), the vibration of the vocal cords produces the speech sound in the case of voice sounds and air turbulence in the case of whispered sounds. A vocal tract consists of the throat, mouth, tongue, teeth, lips and nasal cavity. However, the relatively free passage of the breath produces the vowels through the larynx and oral cavity, while consonants are created by a partial or complete obstruction of the air stream by any of various constructions of the speech organs. Intonation characteristics are pitch, amplitude and voiced/whispered quality and they are dependent on the sound source, while the vocal tract determines the phoneme. Moreover, Parent, King and Fujimura (2002), explained that the human vocal tract consists a number of deep structure can contribute to producing and modifying the basic sound. According to them, the vocal folds (glottis), velum (soft palate), nasal cavity, oral cavity (surrounded by the palate, teeth, cheeks, and tongue), jaw (mandible) and lips are the main components of the sound production system. Besides, they also described one of the main agents that modifies sound is the tongue. It is extremely deformable and controllable by extrinsic muscles. The understanding of the characteristics of the speech signal is critical developing the lip sync system. A physical relationship between the shape of the vocal tract and the sound that is produced concentrated by speech signal. It is segmented into frames. Acoustic to visual feature mapping is then performed by using calculated algorithms, frame by frame.

6.0 AUTOMATED DIGITAL SPEECH SYSTEM

Automated digital speech system is also known as automatic speech recognition or computer speech recognition, which means understanding the voice of the computer and performing any required task. The system in the context of live performances or recordings, a virtual character is driven by speech in the real time. The speaker is talking while at the same time the lip sync process is performed, the animation is displayed on the screen and interacting with the audience (Figure 6). Likewise, real time animation will present to the audience towards the dynamics and changes of new media within live performances, such as theatre, broadcasting, education and live presentation.

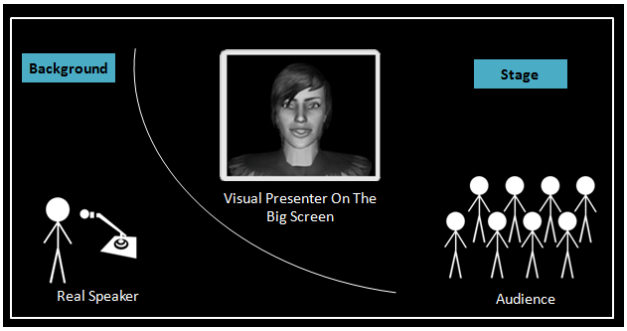


Figure 6 A symbolic view of the virtual presenter by Zorić (2005)

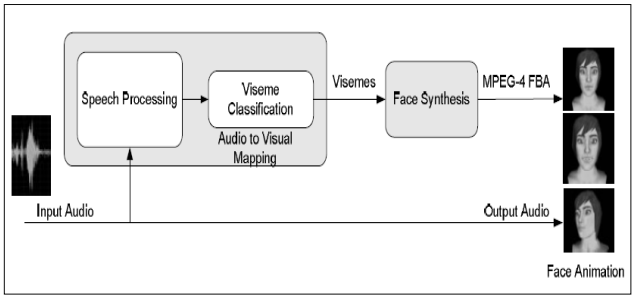


Figure 7 A basic idea of lip synchronization by Zorić (2005)

A basic idea of lip synchronization is shown on Figure 7. Based on Zorić (2005), the process of the

automatic lip sync consists of two main parts. The first one, audio to visual mapping, or more specific speech to lip shape mapping, is a key issue in bimodal speech processing. In this first phase speech is analyzed and classified into viseme categories. The mapping of accurate lip patterns and movement to the sound is extremely important for achieving convincing lip sync animation in real time. In the second part, calculated visemes are used for the animation of virtual character's face. Hence, the process and elements on viseme based human speech for real time lip sync animation is clearly explained in the conceptual framework as shown in the Figure 8.

7.0 CONCEPTUAL FRAMEWORK

Based on the review, a conceptual framework for real-time lip sync animation on viseme based human speech is developed. It is an ideal approach to produce realistic lip sync motion in real time animation. Visually equal mouth positions are categorized into classes of visual phonemes and used for virtual avatar's speak in the animation. A conceptual framework (Figure 8) was developed to outline the overall ideas for generating accurate lip sync on viseme based human speech, and assumptions that hold together, comprising a broad concept of performance of lip sync in real time animation and develop an automated digital speech system.

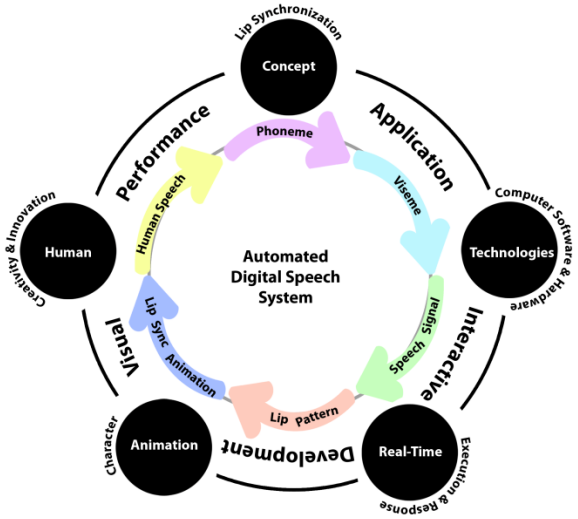


Figure 8 Conceptual framework for real-time lip sync animation on viseme based human speech

The conceptual framework focused on the automated digital speech and corresponding to the main focus, which derived 5 major elements; concept, technologies, real time, animation, and human. These elements are related to each other. From the relation of these elements, it shows that real time lip sync animation is an innovation of interactive application derived from the concept of ordinary lip synchronization animation. The advancement in computer hardware and software in computational technologies enables the interactive application in computer graphics to develop the automated digital speech system in real time application. In real time performance, the visual character is driven by the speech to present to the audience. It involved the interaction between the human and computer to produce the lip sync animation in real time. Apart from that, in real time application, the lip sync animation is derived from the human speech. When the human is speaking, in the mean time, it produced the phonemes, thus, it is classified into the viseme and analyze in computer readable format that is speech signal. The algorithm is used to generate the lip pattern based on speech signal to produce lip sync animation that simulate the human speech. However, there are several factors that influence the human speech. They are human speech behavior, viseme classification and lip pattern. Different human speech behavior will give the different categories of viseme and the lip pattern such as the feature point movement of face components in the generated facial animation. Therefore, the relationship between these factors need to be determined and consider in order to achieve the correct lip sync animation in real time application. Lastly, The conceptual framework can be presented as a guideline to the animator as a tool for generating the automated digital speech on viseme based realistic lip motion.

8.0 CONCLUSION

Lip sync animation is an approach for mapping natural speech to the lip shape animation. In animation, lips remain a crucial part of the outward expression of emotion. In order to make character animation believable, correct lip sync is essential. An automated digital speech system is developed to simulate lip sync in real time, thus ensuring an accurate result through viseme based human speech method. It used to increase the performance of the lip animation. Automatic design with genetic algorithms, use of simple synchronization tricks which generally improve visual impression and implementation of advanced features into lip synchronization application. This makes it possible for a wider range of people and companies to use the application for their needs in real time and offline application. Besides, it's also possible to save more time, reduce the workloads and costing as well as enable output to produce faster. The system directly synthesizing the lip movement with speech from human voice will be resulting a sense of realism and more attractive animation for live performance.

Acknowledgement

The project is supported by the Institute of Design and Innovation at Universiti Malaysia Sarawak.

References

[1] Rathinavelu, A., H. Thiagarajan and S.R. Savithri. 2006. Evaluation of A Computer Aided 3D Lip Sync Instructional Model using Virtual Reality Objects. *Proceeding of 6th International Conference on Disability, Virtual Reality & AssociationTechnology*. Esbjerg, Denmark. 18-20 Sept. 2006. 67-73.

[2] Joshi, C. 2009. Speech Recognition. [Online]. From: http://www.slideshare.net/charujoshi/speech-recognition.

[3] Gary, C. M. 2007. Preston Blair Phoneme Series. [Online]. From: http://www.garycmartin.com/mouth_shapes.html.

[4] Cambridge English Online: Phonetics Focus. [Online]. From: http://cambridgeenglishonline.com/Phonetics_Focus/.

[5] Huang, F. J. and T. Chen. 1998. Real-Time Lip-Sync Face Animation Driven by Human Voice. In *IEEE Workshop on Multimedia Signal Processing*. Los Angeles, California. 7-9 Dec. 1998. 1-6.

[6] Hofer, G., J. Yamagishi and H. Shimodaira. 2008. Speech-Driven Lip Motion Generation with A Trajectory HMM. In *Proc. Interspeech 2008*. Brisbane, Australia. 22-26 Sept. 2008. 2314-2317.

[7] Zoric, G. 2005. *Automatic Lip Synchronization by Speech Signal Analysis*. Master Thesis. Faculty of Electrical Engineering and Computing. University of Zagreb.

[8] Zoric, G. and I. S. Pandzic. 2005. A Real-Time Lip Sync System using A Genetic Algorithm for Automatic Neural Network Configuration. *Proc. IEEE International Conference on Multimedia & Expo ICME*. Amsterdam, The Netherlands. 6 July 2005. 1366-1369.

[9] Lewis, J.P. 1991. Automated Lip-Sync: Background and Techniques. *Journal of Visualization and Computer Animation*. 118-122.

[10] Lewis, J. P. and F. I. Parke. 1990. Automated Lip-Synch and Speech Synthesis for Character Animation. *SIGGRAPH 1990 17th International Conference on Computer Graphics and Interactive Techniques*. Dallas, TX, USA. 6-10 Aug. 1990. 83-87.

[11] Spevack, J. 2008. 2D Animation. [Online]. From: http:// http://profspevack.com/archive/animation/course_cal/week12/week12.html.

[12] Berger, M. 2012. Move Over Motionscan; New Lip Synch Tech Aims to Revolutionize Motion Capture. [Online]. From: http://venturebeat.com/2012/02/09/lip-synch-tech-to-revolutionize-motion-capture/.

[13] Scott, M. J. 2010. *Digital Puppetry*. Interviewed by Kenny, H.S.H. Universiti Malaysia Sarawak.

[14] Beiman, N. 2010. Animated Performance: Bringing Imaginary Animal, Human and Fantasy Characters to Life. Switzerland: AVA.

[15] Blair, P. 1994. Cartoon Animation. California: Walter T. Foster Publishing.

[16] Budiman, R., M. Bennamoun and D.Q. Huynh. 2005. Low Cost Motion Capture. The University of Western Australia, Australia.

[17] Parent, R., S. King and O. Fujimura. 2002. Issues with Lip Synch Animation: Can You Read My Lips?. *The 15th International Conference on Computer Animation*. Geneva, Switzerland. 19-21 June 2002. 3-10.

[18] Clara, S. 2009. The Henson Digital Puppetry Studio Revolutionizes Television Production using NVIDIA Quadro Processors. [Online]. From: http://www.renderosity.com/nvidia-congratulates-jim-henson-s-creature-shop-on-winning-primetime-emmy-engineering-award-cms-14712.

[19] Dyer, S., J. Martin and J. Zulauf. 1995. Motion Capture White Paper. [Online]. From: http://reality.sgi.com/jamsb/mocap/MoCapWP_v2.0.html-HDR0.

[20] Tang, S. S., A. W. C. Liew and H. Yan. 2004. Lip-Sync in Human Face Animation based on Video Analysis and Spline Models. *Proceedings of the 10th International conference on Multimedia Modeling*. Brisbane, Australia. 5-7 Jan. 2004. 102-108.

[21] Frank, T., M. Hoch and G. Trogemann. 1997. Automated Lip-Sync for 3D-Character Animation. In *15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics*.