Practical Lip-synch Tools for 3D Cartoon Animation

Shin'ichi Kawamoto[*,**], Tatsuo Yotsukura[†], and Satoshi Nakamura[¶]
[*]Japan Advanced Institute of Science and Technology, Japan
[**]Advanced Telecommunications Research Institute International, Japan
E-mail: kawamoto@jaist.ac.jp
[†]OLM Digital, Inc., Japan
E-mail: yotsu@olm.co.jp
[¶]Nara Institute of Science and Technology, Japan
E-mail: s-nakamura@is.naist.jp

***Abstract*— This paper gives an overview of our lip-synch animation production framework with practical tools for making 3D animation efficiently based on pre-scoring. Our framework is so simple and easy to use that it can be applied to construct various systems: a management tool for making lip-synch animation, a batch processing tool for mass production, Autodesk Maya plug-in software for practical workplaces, and amusement systems. We also demonstrate practicality of our framework through several practical applications. Our framework worked well in the production at practical workplaces of cartoon animations.**

I. INTRODUCTION

A huge number of 3D cartoon animations have recently been produced in the entertainment industry on a global scale. Commercial 3D cartoon animation, such as for a TV series or games, must then be created efficiently to meet such massive demands. The blendshape technique is one basic method that contributes to such practical demands. This linear interpolation technique has generally been applied to create realistic facial animation, but, of course, it could also be used for making 3D cartoon facial animations. However, various time-consuming processes are needed in making these animations. In particular, when animators have the characters making speech animation, it is important to synchronize mouth movements with speech sounds. For making well synchronized lip-synch animation, many practical workplaces are recording voices before making the character animation, which is called pre-recording or pre-scoring. However, this is a very time-consuming task because animator has to carefully analyze the timing of mouth movements. In this research, we propose an efficient lip-synch workflow focusing on stylized 3D cartoon animations which are created with blendshapes and pre-recorded speech information.

Many facial and speech animation techniques have been studied: key-framing methods [1,2], physics-based methods [3,4] and video-based methods [5,6]. Some of these automatic realistic lip-synch techniques can be applied to set blendshape parameters. However, these lip-synch techniques have focused solely on cloning realistic human motion, which is quite different from the stylized motions in 3D cartoon animation.

Several lip-synch animation systems have been released:

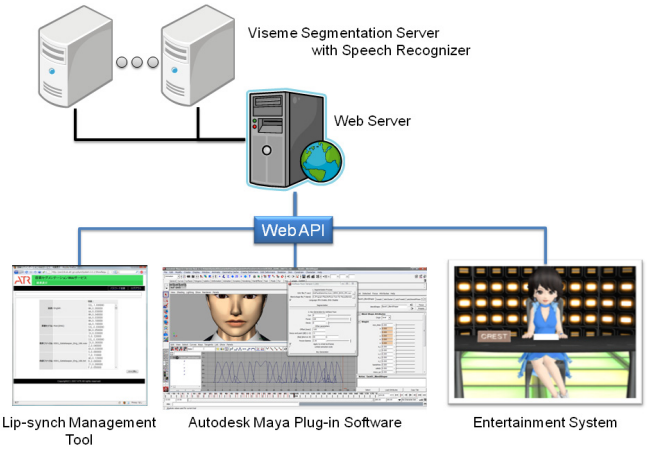

Fig. 1 Overview of our framework.

Autodesk MotionBuilder, and Autodesk Face Robot toolset. Most of them usually need the creation of many target shapes with the corresponding phonemes in order to carefully tune the speech analysis parameters as well as to edit the keyframes independently. We have already proposed a lip-synch system that has the advantage of easily creating cartoon lip-synch animation by using only a few target shapes [7]. In this paper, we propose a novel framework including our lip-synch system for making mass 3D CG animation efficiently. Our system is based on web technology, making it easy to construct various types of applications. We develop following systems based on our framework: a simple management system for making lip-synch animation, a batch processing system for mass production, Autodesk Maya plug-in software for practical workplaces, and interactive lip-synch systems for entertainment.

II. ANIFACE: LIP-SYNCH ANIMATION PRODUCTION FRAMEWORK

Here, we describe our novel lip-synch animation-production framework we call AniFace. As shown in Fig. 1, our system works as an Application Service Provider (ASP) using a Web API based on HTTP protocols, since our technology should be able to apply various types of

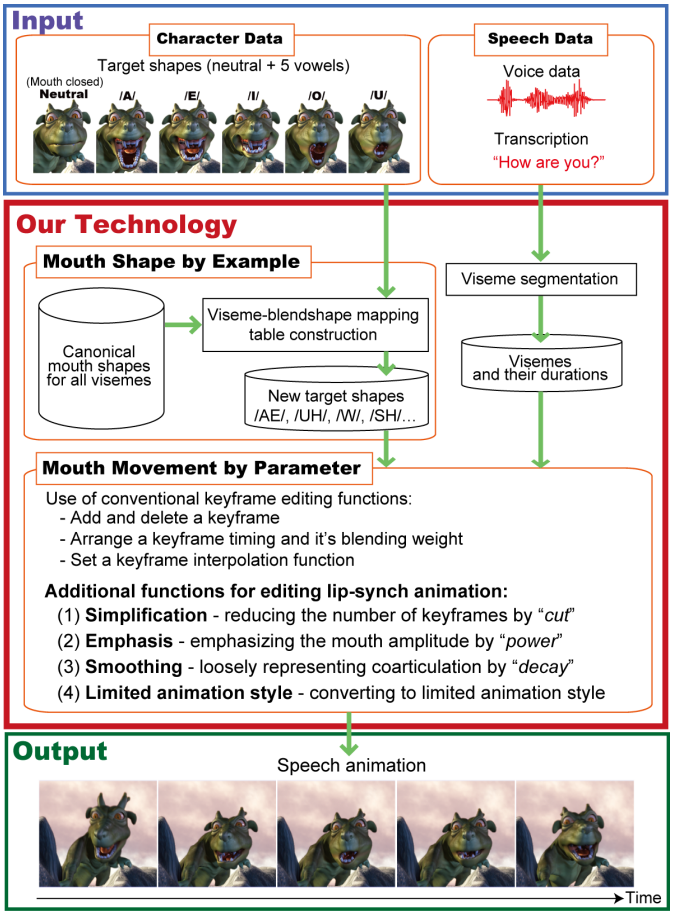

Fig. 2 Processing flow of our lip-synch technology [5].

applications to unify simple and common protocols. In this section, we describe the advantages of our framework.

A. Lip-synch Animation Production Techniques for 3D Cartoon Characters

Stylization in 3D cartoon speech animation should be achieved with mouth shape and movement that are shown in simpler and/or more exaggerated ways than realistic cases. One of the typical stylizations is limited animation using one cell for every two or three frames of film, and/or reducing the amount of drawings necessary to give an illusion of movement. Our system is implemented with editing parameters to realize these functions [7]. As shown in Fig. 2, the designer prepares the character data and the speech data as input. The former data are the base target shapes of the character to be animated, whereas the latter are the character voice data and its transcription. We assume that an arbitrary mouth shape during animation is represented as a linear sum of the input target shapes. This means that the base target shapes are designed to constitute the blendshape basis. We leverage the following system features to achieve a high level of the system's practical efficiency:

(1) Mouth shape by example: Our system constructs the viseme-blendshape mapping table from the character data. This means that all of the character's viseme mouth shapes

Fig. 3 Screenshot of web interface.

are parameterized using the blendshape weights with regard to the base target shapes. This feature is beneficial for quickly making a default mouth shape for a given viseme. The default will thereafter be modified using the blendshape technique, which provides a much easier way of mouth-shape design than making it from scratch.

(2) Mouth movement by parameter: We introduce a few parameters for quickly editing mouth motions in order to control mouth amplitude, movement speed, and shape during a certain period of time. These parameters are introduced to complement blendshape-based keyframing and thus achieve fine yet easy tuning of speech animation.

B. ASP for Viseme Segmentation

Our system sends the speech data to the viseme segmentation server via the Internet using HTTP protocols. For analyzing the speech data, we employ an HMM-based speech recognizer (ATRASR [8] developed by Itoh et al.), which is high-precision speech recognition software for noisy environments [9] to obtain phoneme segmentation. The phoneme segmentation result is further converted to the viseme segmentation using a phoneme–viseme mapping table by the simple table lookup method [10]. The supported languages are Japanese and English. The primary advantage of using this server is that we can easily maintain it independently of our system. For example, we can get the latest segmentation function or replace the acoustic model data soon after they are updated. Furthermore our system can replace and add the viseme segmentation servers easily for system failure maintenance and throughput improvement.

The simple web interface allows us to manage the viseme segmentation processing. Fig. 3 shows a screenshot of the

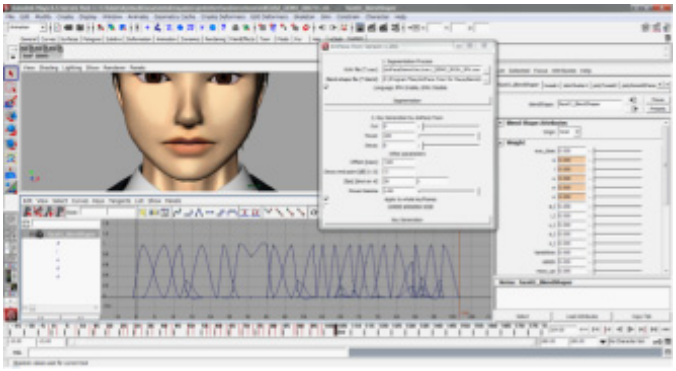

Fig. 4 Screenshot of plug-in software.

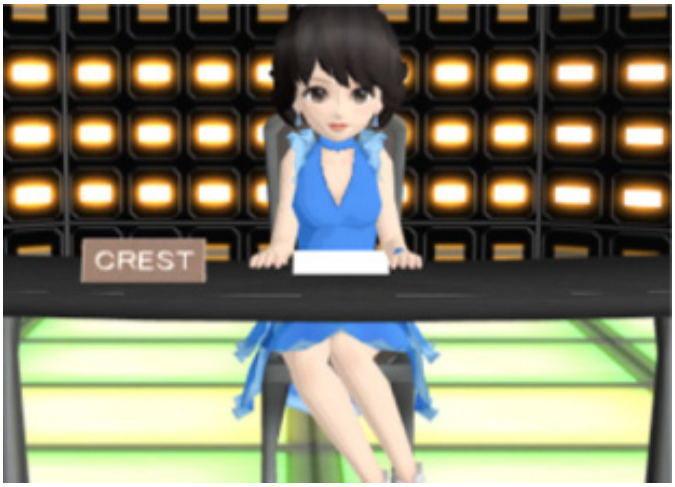

Fig. 5 Screenshot of iFACe.

web interface. We also provide a simple CUI system for mass production that applies batch processing to multiple speeches.

C. *Interface Design for Plug-in Software*

One of the time-consuming tasks for making lip-synch animation is placing keyframes of mouth shapes by analyzing input speech manually. Tuning the keyframe timing and its blending weight is also time-consuming. Therefore, we unify two functions in a GUI for making lip-synch animation: the access function to the viseme server and a quick keyframe-editing function.

The current GUI provides the least number of parameters to be tuned, compared to previous versions of this system, after repeating several system revisions based on designers' critiques. Since we refined the GUI to make it fully adaptive to actual workplaces, the current GUI has become very useful for designers. The current system provides several useful functions such as keyframe timing shift function and selection of start and end frames for editing locally. The GUI of our system is implemented as Autodesk Maya plug-in software (Fig. 4).

D. *Application of Entertainment System*

AniFace is applied for realizing an Interactive Facial Animation system (iFACe) [11] that lets players easily experience being a professional voice actor. In current Japanese anime production of voice-over acting, producers

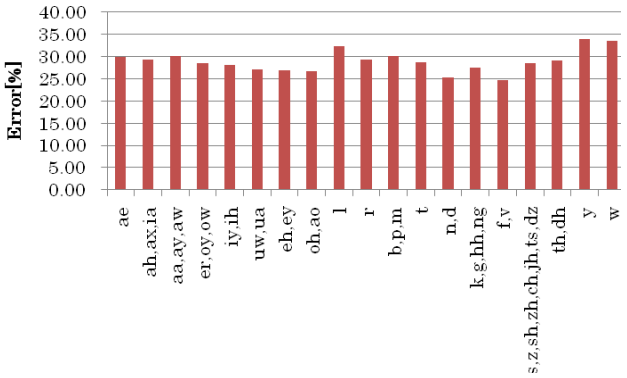

Fig. 6 Facial expression discrimination error for each viseme class.

commonly employ "post-recording," which requires specialized skill in the exact synchronization of voices with the lip-movements of pre-created characters. iFACe is a blendshape-based lip-synch animation system whose only input is the player's spontaneous speech with or without a transcript. First, iFACe estimates visemes and their durations from speech, even in such noise environments as amusement facilities. Second, iFACe automatically represents suitable target-shaped keys for cartoon CG characters using viseme information. Furthermore, the system generates real-time cartoon-character speech animation from output keys on graphics hardware. AniFace generates the keyframes of lip-synch animation from the player's speech. Fig. 5 shows a screenshot of iFACe.

III. Results and Discussion

Our approach to creating lip-synch animation has been tested in close collaboration with many designers in digital production companies. They have made several short films with our systems. All of the shots shown here were taken from the two short pieces in English: "Iron Wand Princess" (2007) and "Parallel World Bus Tour" (2008). Our system worked on various characters, including a non-realistic/non-human character in these short films. In making these pieces, the base target shapes, i.e., the character data, for five vowels were assigned to each character. Using our system, it took only 3 days for the designers to create the lip-synch animation part in the 18 shots of "Iron Wand Princess." This short film was created with a process about three times as efficient as a conventional creation process, according to the designers' comments.

A. *Interference between mouth and facial expression*

Target shapes for lip-synch and facial expressions are sometimes designed independently. In the creation of expressive speech animation, a simple combination of target shapes sometimes produces undesired shapes that conflict with the target shapes. Therefore, we investigated the interference between the mouth shapes and the facial expressions by a pattern recognition technique for discriminating each facial expression. The mouth shape and

facial expression data are collected by using a motion capture system. The speech data is recorded at the same time. The subject, who is a female native-English speaker, uttered 60 sentences for each facial expression. The varieties of facial expression are neutral and angry. The mouth-shape data was sampled from the start of the viseme based on the viseme recognition results, since the keyframes in our method are placed at the start of the viseme. We assumed that a small recognition error means the facial expression affects the target mouth shape, since the mouth shapes of the same viseme are significantly different depending on the facial expression. The feature vector is the 3D positions of 12 markers placed around the mouth. The facial expressions discriminate by using Bayesian Logistic Regression [12], which is a simple and robust pattern classifier. Fig. 6 shows the facial expression discrimination errors for each viseme class. As shown in Fig. 6, the discrimination errors depend on the viseme. This result indicates that the amount of interference depends on the visemes. If we extend our technology [7] to handle lip-synch and facial expression at the same time, we should consider interference between the visemes and the facial expressions.

IV. Conclusion and Future Work

We described a practical lip-synch animation production framework and tools for 3D characters. Our framework provides web-based viseme segmentation service for making lip-synch animation. The core technology of our lip-synch method is implemented for various systems. Our approach to creating lip-synch animation has been tested in close collaboration with many designers in digital production companies. They have also made several short films with our system. Our approach was applied to create lip-synch animation in Capcom's video game "Sengoku BASARA 3." [13]

Important future work includes developing a more comprehensive way of efficiently creating expressive speech animation from a few target shapes. These shapes express not only mouth geometry but also the entire face with emotions. In this case, we must consider how to manage the interference among target shapes.

Acknowledgment

The authors would like to thank Ken Anjyo for helpful discussion. The authors would also like to thank Shanghai Benson Animation, Total Planning Office, and Shanghai Jishi Business Consulting for many useful comments on the system's evaluation and improvement. Authors would also like to thank Ming C. Lin for valuable discussions on the interference between mouth shapes and facial expressions. This work was supported in part by the Japan Science and Technology Agency (JST), CREST Project.

References

[1] M.M. Cohen, and D.W. Massaro, "Modeling coarticulation in visual speech," in *Models and Techniques in Computer Animation*, Thalmann N. M., Thalmann D. (eds). Springer-Verlag: Tokyo, pp.139-156, 1993.

[2] P. Joshi, W.C. Tien, M. Desbrun, and F. Pighin, "Learning controls for blend shape based realistic facial animation," In Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation (SCA '03). pp.187-192, 2003.

[3] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, pp.55-62, 1995.

[4] E. Sifakis, I. Neverov, and R. Fedkiw, "Automatic determination of facial muscle activations from sparse motion capture marker data," ACM Trans. Graph. Vol. 24 Issue 3, pp.417-425, 2005.

[5] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, pp.388-398, 2002.

[6] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, pp.353-360, 1997.

[7] S. Kawamoto, T. Yotsukura, K. Anjyo, and S. Nakamura, "Efficient lip-synch tool for 3D cartoon animation," *The journal of Computer Animation and Virtual Worlds*, vol.19, issues.3-4, pp.247-257, 2008.

[8] G. Itoh, Y. Ashikari, T. Jitsuhiro, and S. Nakamura, "Summary and evaluation of speech recognition integrated environment ATRASR," in *Autumn Meeting of the Acoustical Society of Japan*, pp. 221–222, 2004.

[9] M. Fujimoto, and S. Nakamura, "A non-stationary noise suppression method based on particle filtering and Polyak averaging," *IEICE—Transaction on Information Systems*, E89-D(3): pp. 922–930, 2006.

[10] T. Yotsukura, S. Morishima, and S. Nakamura, "Model-based talking face synthesis for anthropomorphic spoken dialog agent system," in *Proceedings of the 11th ACM International Conference on Multimedia*, pp. 351–354, 2003.

[11] T. Yotsukura, S. Kawamoto, S. Matsuda, and S. Nakamura, "iFACe: Interactive Facial speech Animation Control systEm for 3D Cartoon Characters," *IPSJ Journal*, Vol. 49, No. 12, pp. 3847-3858, 2008.

[12] A. Genkin, D.D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, Vol. 49, No. 3, pp. 291-304, 2007.

[13] S. Kawamoto, T. Yotsukura, S. Nakamura, J. Yamamoto, T. Shirahama, and H. Yamamoto, "Integrating lip-synch into game production workflow: Sengoku BASARA 3," in *ACM SIGGRAPH ASIA 2010 Sketches*, Article 2, p.1, 2010.