# Synthesizing Expressive Facial and Speech Animation by Text-to-IPA Translation with Emotion Control

Andreea Stef* , Kaveen Perera† , Hubert P. H. Shum‡ and Edmond S. L. Ho§
*Department of Computer and Information Sciences*
*Northumbria University*
Newcastle upon Tyne, United Kingdom

***Abstract*—Given the complexity of the human facial anatomy, animating facial expressions and lip movements for speech is a very time-consuming and tedious task. In this paper, a new *text-to-animation* framework for facial animation synthesis is proposed. The core idea is to improve the expressiveness of lip-sync animation by incorporating facial expressions in 3D animated characters. This idea is realized as a plug-in in Autodesk Maya, one of the most popular animation platforms in the industry, such that professional animators can effectively apply the method in their existing work. We evaluate the proposed system by conducting two sets of surveys, in which both novice and experienced users participate in the user study to provide feedback and evaluations from different perspectives. The results of the survey highlights the effectiveness of creating realistic facial animations with the use of emotion expressions. Video demos of the synthesized animations are available online at https://git.io/fx5U3**

***Index Terms*—Lip-sync, Facial animation, Facial expressions, Emotion, Character animation**

## I. Introduction

Animated films have evolved rapidly in the past few decades. Today, most of these films are made with CGI (Computer-generated imagery). New technology and tools have been introduced for character modelling, texturing and animation, allowing the creation of natural postures using data-driven Inverse Kinematics (IK) [1], the interaction among multiple characters and/or objects [2], [3], and even realistic micro features to meet with the high expectations of the audiences.

The human face is one of the most important means for human communications. People have the ability to decipher even the slightest changes in others' facial expressions. This can sometimes be identified as an instinct. Experts believe that respective expressions for emotions such as fear, joy, surprise, sadness have taken the same form for thousands of years. In animation, a character with realistic expressions allows the viewer to connect with the scene. Given the complexity of the human facial anatomy, animating facial expressions and lip movements for speech is a very time consuming and a tedious task. Several approaches are proposed to create the facial expressions or speech of a 3D character. However, many of these are limited within major animation studios. 3D animators have limited access to such tools or the available tools are limited in their functionality.

To provide the animators with an intuitive way to edit character animation by emotions, a data-driven approach is proposed in a recent study [4]. However, such a framework only handles full body Motion Capture (MOCAP) data rather than facial animation. In order to create a good implementation of realistic facial expressions with lip-sync, an understanding of the primary expressions produced on the human face is required.

Emotions are the trigger factor of the expressions for communicating with other beings and indicating the traits. Emotional expressions are special because they are involuntary, instead of intentional, and not all signals are the same [5]. Facial expressions are universally understood signals - an indication of how people feel and respond to situations occurred in their day-to-day life, allowing people to identify others feelings. They can be faked and controlled to a certain degree with a lot of practice. An actor's facial expressions can be used to reveal a subtext deeper meaning that contradicts what they say or does in the scene. A character may be greeting another character verbally. However, a slight movement of a different part of the face, such as mocking eyeroll, will show the audience the character's true emotion. Facial expressions and facial muscles do not map one-to-one. Some facial movements involve the contraction of two different parts of the same muscle, while the others involve the contraction of multiple muscles [6].

This study is set to investigate the effectiveness of using IPA based *text-to-animation* framework for speech animation and incorporation of facial expressions in 3D animated characters. In particular, a new speech animation synthesis system is proposed and a prototype of the system is developed as a plug-in in Maya.

The main contributions of this paper can be summarized as follows:

1) Proposed and developed a framework for synthesizing

* email: stef.andreea.sorina@gmail.com
† email: kaveen.perera@northumbria.ac.uk
‡ email: hubert.shum@northumbria.ac.uk
§ email: e.ho@northumbria.ac.uk, *corresponding author*

facial animation in a *text-to-animation* manner.
2) Intuitive control for editing facial expressions to incorporate emotion in lip-sync animation.
3) Conducted a user-study to highlight the effectiveness of improving the expressiveness of lip-sync animation.

## II. Related work

Producing realistic speech and emotion animation is a challenging task and attracted a lot of attention in the computer graphics community in the last few decades. Two decades ago, Bregler et al. [7] proposed a video synthesis system by reusing annotated videos frames to create a new video according to an input audio sequence. Computer vision techniques are used for tracking the movement of the mouth in order to annotate the video frames by matching the Phonemes.

Since then, a lot of interesting work related to audio-controlled animation synthesis were proposed. For example, Voice Puppetry [8] can be used for animating a single input picture with rich facial expressions based on the soundtrack given by the user. Yang et al. [9] proposed a system in which a generic 3D head model is animated according to the input speech first, and then a set of MPEG-4 Facial Definition Parameters (FDPs) will be obtained from the 3D model for synthesizing the results. While the aforementioned approaches are based on machine learning and computer visions techniques to understanding the underlying principles for audio-driven facial animation, Xu et al. [10] proposed a new lip-sync approach based on the expert knowledge from animation artists. In particular, phonemes (diphones) and face poses are associated with animation curves predefined by artists to enable easy control over the animation. Lip-sync animation can be applied to a wide range of applications, such as computer games [11] and animation production [12]–[14]. Also, there is a related work focused on incorporating emotion in speech-driven animation [15]. While we share similar interests on emotion expression, the previous work [15] applied to 2D cartoon animations only, while our proposed framework focuses on synthesizing 3D animations.

## III. Methodology

In this section, a new facial animation framework will be presented. In particular, the facial animation will be created in a *text-to-animation* manner. The proposed *text-to-animation* framework consists an RP English (Received Pronunciation is an accent of Standard English in the United Kingdom) to IPA translation module, IPA symbol to pose mapping module and a keyframe/animation clip module. An overview of the framework is illustrated in Figure 1. First, an RP English dictionary was created using data harvesting methods from the internet. The IPA symbol translations of these words were obtained from https://tophonetics.com/. Video recordings of reference Lip movements covering all the IPA symbols were obtained from 3 participants. Then after careful study of the references an IPA symbol to pose table was generated. Same references were used to model poses for Maya blendshapes. For the implementation of this research, the fully rigged digital

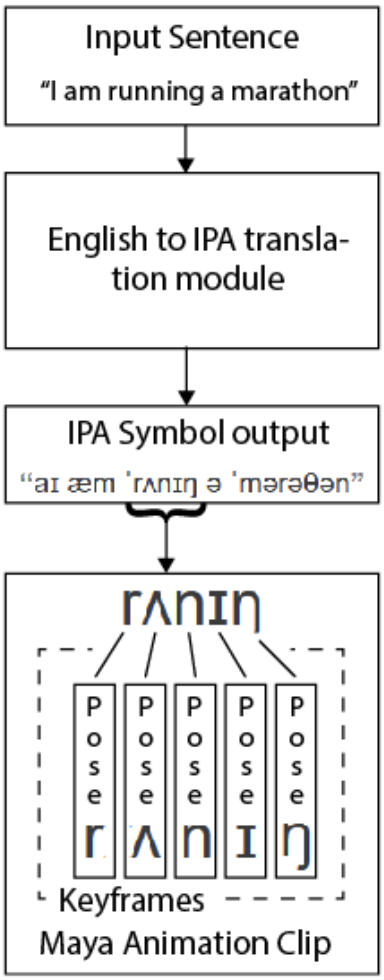



Fig. 1. The overview of our proposed framework.

double head model Louise from Eisko [16] was used. The model is animated using blendshapes.

When an English sentence is submitted as the input, each of the words will be translated into IPA. The system will then create keyframes for matching IPA translation of each IPA symbol with two frames gap between them. Next, using these keyframes the system creates an animation clip for each word. Users can indicate if they wish to blend two words when using the '+'. The period symbol '.' at the end of the sentence signals the system to return to neutral pose after speaking a word. For example, consider the following sentence, 'I am running a marathon' is spoken at a regular speed, 'I', 'am' and 'running', 'a' is spoken without a pose in-between. In this instance, the two animation clips should be blended. Then at the end of the sentence, the character's mouth pose should be returned to neutral pose. To indicate that the user can input the same sentence to the system as the following:

**'I+am running+a marathon.'**

### A. Using IPA in Speech Animation

IPA is designed with a methodical analysis of human speech. IPA defines a specific symbol for each phoneme of many of the modern languages. In English IPA it is very useful to indicate the correct pronunciation of a word independent of the spellings variations and when words are not spelt phonetically. Thus, using IPA to define the poses for speech animation provides a solid foundation for us to start with.

IPA for RP English contains 44 symbols (20 vowels and 24 consonants) [17]. Speech for 20 IPA vowels symbols can be animated with using only 6 or 12 poses. Vowels contain 6 short vowels, 6 respective long vowels and 8 diphthongs. Mouth shapes for long vowel pronunciations extend further than their respective short vowels. For example, the pose for ü takes a pointed lips shape while for the respective short vowel u:, lips only point to a lesser extent. During this research, we modelled 12 different poses for long and short vowels. Diphthongs are formed by combing two vowels and it was observed that mouth shapes can be formed by combining the poses of two respective vowels.

Natural looking speech for the 24 IPA consonant symbols can be animated only using 15 poses. 6 of these poses (visemes) are used by more than one IPA symbol. Some special characteristics were noticed with h (a voiceless glottal fricative) and n (a nasal). When pronouncing words containing IPA symbol h, the mouth takes the shape of the following IPA symbol. For an example, the word overhead ([ˈəʊvəhed]) has 7 IPA symbols, yet this can be animated with only using 6 poses ə, ʊ, v, ə, (h)e, d respectively. When IPA n is presented at the start or the end of a word the lips are slightly opened. However, when n is presented at the middle of a word, the (eg: funny - [ˈfʌnɪ]) the mouth shape for n is only a long transformation between ʌ and ɪ. If the 3D character has visible tongue and teeth, the tongue should be animated to touch the hard palate of the mouth in place of n.

The number of poses can be further reduced by combining poses for a more cartoon style speech animation project.

### B. The Aniamtion Production Pipeline

The production pipeline of our proposed framwork is illustrated in Figure 2. We use Maya's character sets to store animation information. The scene should be prepared first by creating a character set and importing relevant audio recordings. Then, animation clips are created once the text is entered and previewed in IPA to rectify any errors. Then all the animation clips can be selected, scaled and moved in Maya's Trax editor to match timing. This process should take less than a minute. Then the keyframes within the animation clips can be adjusted to fine-tune timing if required.

The plugin consists its own preview window, sliders to adjust facial expressions, options to create a character set, import audio, text input field, IPA preview field, the Trax editor to edit animation clips, timeline and time controls. The plugin displays a word for word IPA translation and the interpretation

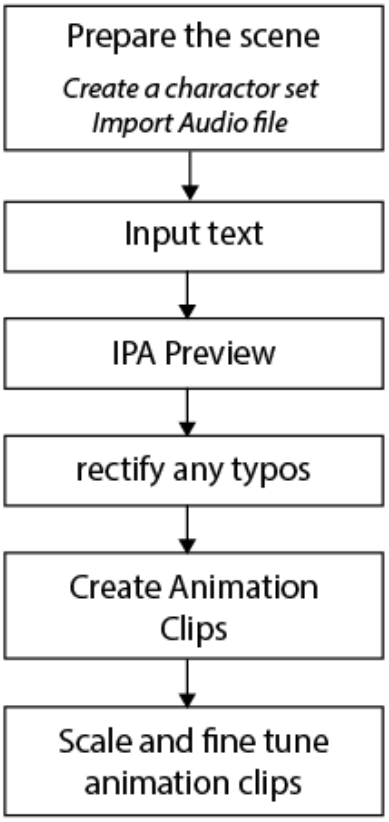



Fig. 2. The workflow when using the proposed framework and the plugin

of the control symbols '+' and '.'. The output of the translation is shown in Figure 4.

## IV. Experimental Results - The User Study

To evaluate the quality of the animation synthesized using our proposed framework and the effectiveness of improving the expressiveness of the facial animation, two surveys were designed.

The first survey was targeted at experienced or expert 3D animators. They were asked to produce speech animations of the pre-recorded audio 'I am running a marathon' using the plugin as well as manual keyframing of the pre-defined poses. Then they were asked to evaluate the effectiveness of improving the expressiveness of the facial animation.

The second survey was focused on evaluating the quality of the facial animations synthesized using our system. Specifically, animations were created by specifying the emotion using our GUI. Then, the participants were asked to watch the animations (3-10 seconds) and to select the most appropriate emotion expression to describe the animation. To evaluate the expressiveness we created 2 sets of facial animations using our proposed *text-to-animation* framework. The first set had lip-sync animations without editing facial expressions. Emotions were added to the second set and the users were asked to select the most expressive animation clips.

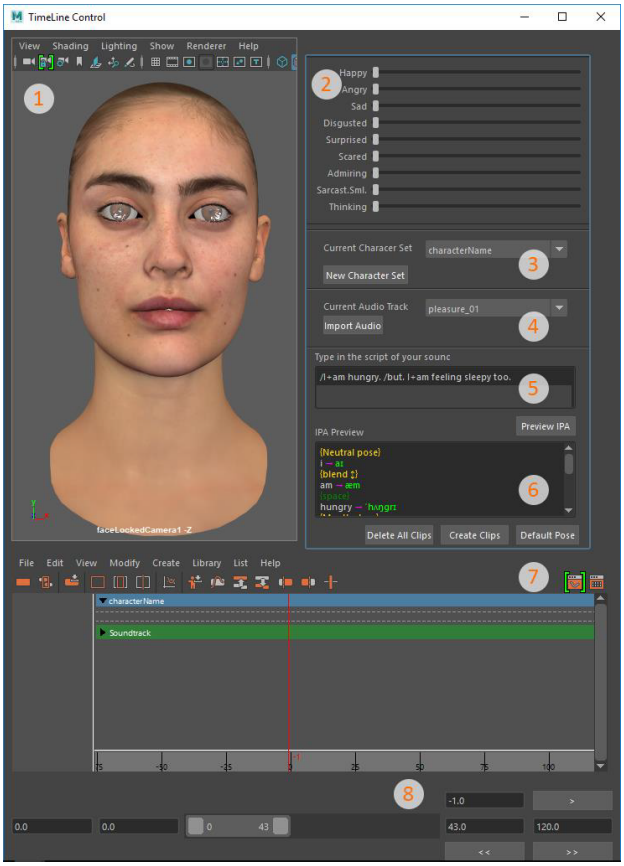



Fig. 3. The user interface of the proposed system developed as a Maya plugin.

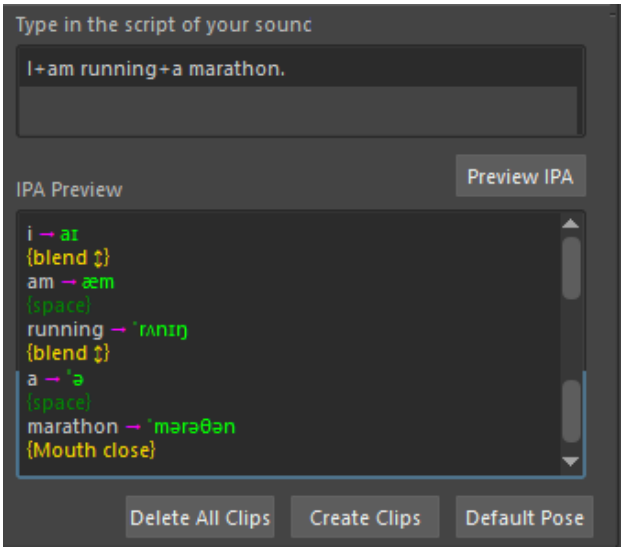



Fig. 4. An example of translating a sentence written in English to IPA.

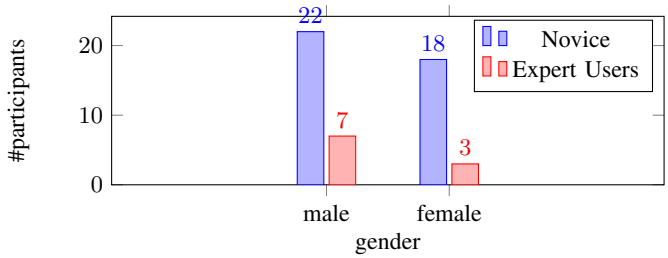



Fig. 5. Gender distributions of the participants

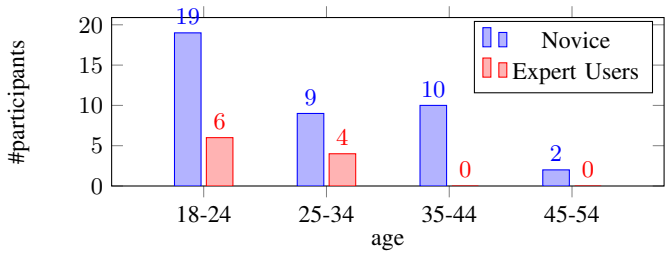



Fig. 6. Age distributions of the participants

Video demos of the synthesized animations are available online at https://git.io/fx5U3

### A. The Participants

41 *novice* and 10 expert Maya users participated in our surveys. *Expert users* had to take part in both surveys while the novice users only completed the second survey. The second survey was made available online to worldwide participants. We will present the results of the survey into 2 groups, namely novice and expert users. The age and gender distributions are illustrated in Figure 6 and 5, respectively. The first survey was targeted at expert users with additional questions related to their experience in using Maya (Figure 7) for animation production. The majority of the expert users had more than 2 years experience in using Maya.

### B. Evaluating the quality of the synthesized facial animations

*1) First Survey:* The text-to-animation framework allows the instant creation of animation clips to begin with regardless of the length of the sentence. Timing, fine-tuning and animating of the 5 letter words sentence, I am running a marathon' *only took an average of 5 minutes among the expert users*

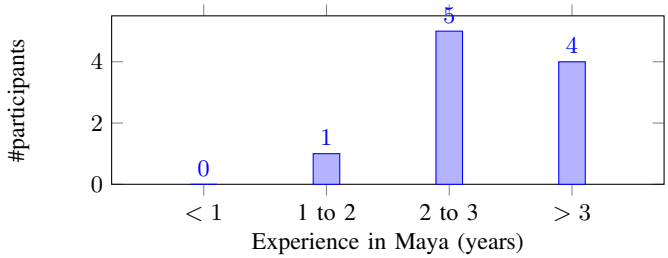



Fig. 7. Experience in Maya (years) of the expert users

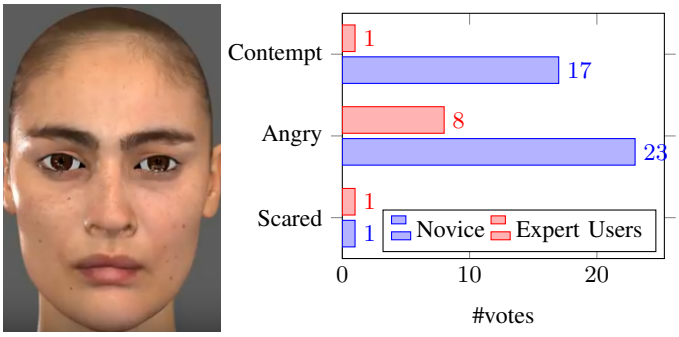



Fig. 8. The screenshot of a synthesized *angry* facial animation (left) and the results of the online survey (right)

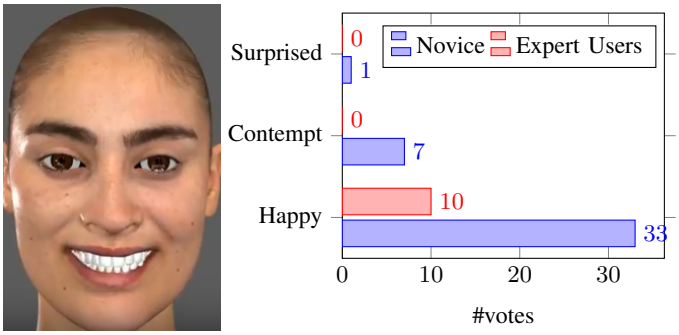



Fig. 9. The screenshot of a synthesized *happy* facial animation (left) and the results of the online survey (right)

*when using the plugin*. The same task recorded an average of 20 minutes with manual keyframing of blendshpes. Add more details about the second survey.

*2) Second Survey:* The first part of the second survey tested the participants' ability to recognize the facial expressions. Following a thorough analysis of the results, it was observed that expert users performed better than the novice group (see Figures 8 to 14). In some situations, happiness was misinterpreted with sadness or anger. Contempt is a complicated expression which even has created confusion among psychologists who have studied facial expressions. The animators are more experienced in observing these expressions and (in general with every human movement) therefore they were able to recognize expressions more accurately. Nevertheless, the majority of the participants, in both the novice and expert user groups, perceived the correct emotion from the synthesized facial animations. This highlights the effectiveness of our proposed framework for facial expression editing.

### *C. Evaluating the effectiveness of improving the expressiveness of the facial animation*

The results of the surveys are illustrated in Figures 15 to 18. In both the novice and expert user groups, the participants marked the animations that included facial expression with speech as being the best to match to the given scenario. From the results, 86.25% and 83% of participants choose the videos created using our facial expressions editing interface as the better ones from the novice and expert user groups,

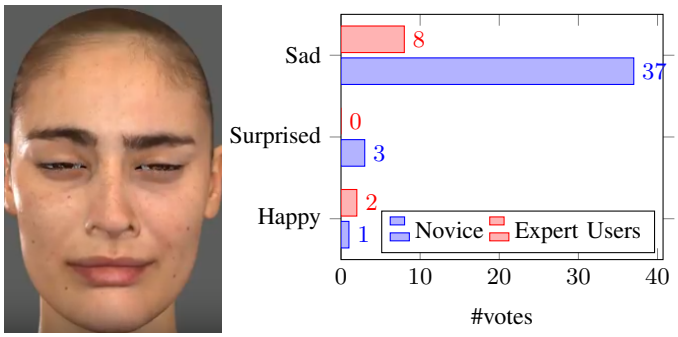



Fig. 10. The screenshot of a synthesized *sad* facial animation (left) and the results of the online survey (right)

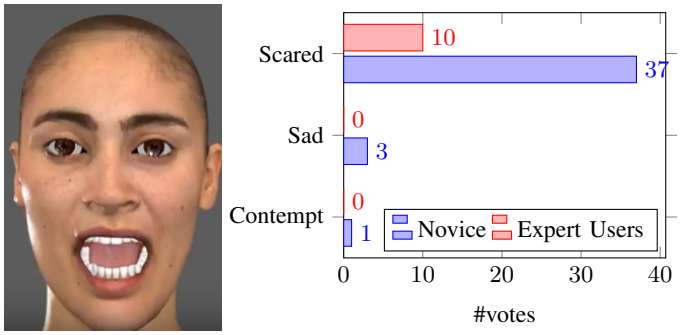



Fig. 11. The screenshot of a synthesized *scared* facial animation (left) and the results of the online survey (right)

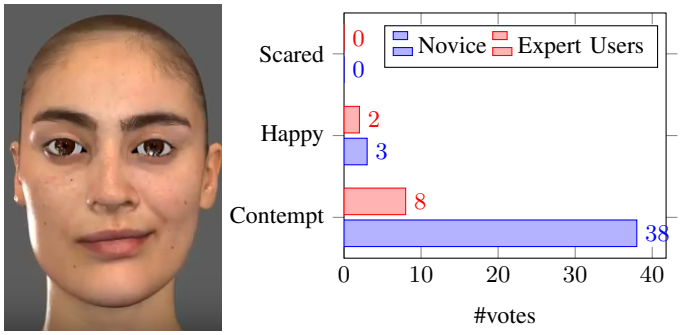



Fig. 12. The screenshot of a synthesized *contempt* facial animation (left) and the results of the online survey (right)

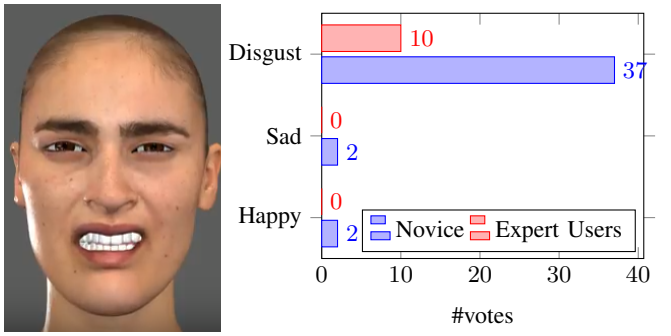



Fig. 13. The screenshot of a synthesized *disgust* facial animation (left) and the results of the online survey (right)

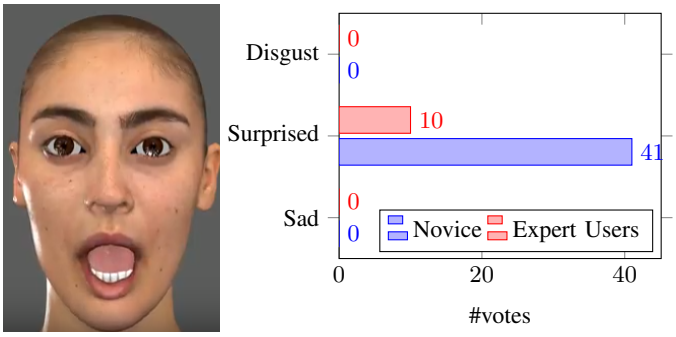



Fig. 14. The screenshot of a synthesized *surprised* facial animation (left) and the results of the online survey (right)

respectively. The high percentage of choosing the animations with facial expressions confirms that the effectiveness of improving the expressiveness of our proposed framework.

### *D. Overall satisfaction on the proposed facial animation synthesis framework*

Research findings of lip shapes and mouth movements with IPA for speech animation were implemented in production. This showed exceptional results with 51% of the participants rating 'Extremely satisfied' and 28% rating 'Very satisfied' (see Figure 20). Regarding the overall satisfaction on combining the speech and facial expression for 3D animation production, 53% of the participants rating 'Extremely satisfied' and 22% rating 'Very satisfied' (see Figure 19)

The participants provided very positive feedback with the following quotes:

**" This was a very useful and easy program to use, especially if an animator needs to lip sync audio."**

**" An interesting and complex plug-in which could be very useful for animators."**

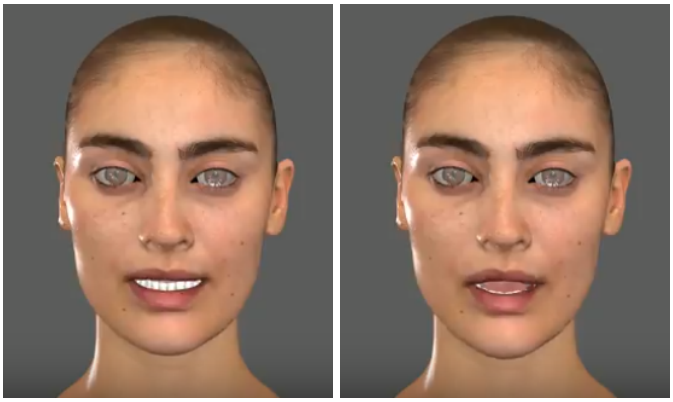
**" Excellent plug in, interaction between various face shapes need work, could be an extremely effective animation plug-in."**

The users were able to understand the concept of using IPA very quickly with the information presented on the IPA preview panel. Further, they relished the design of the plugin, verbose IPA preview panel, the inclusion of a viewport and the Trax Editor within the interface. Above feedback and the 98% satisfaction from the end users demonstrates the great success of the framework.

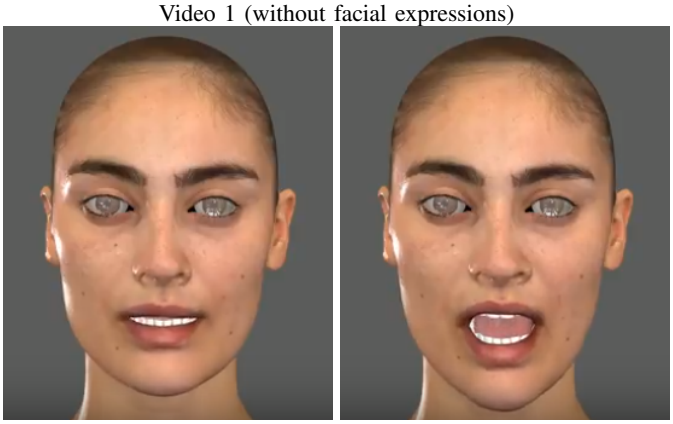
## V. Discussions

The research on facial expressions dives deeper into analysing facial movements. However, for the actual implementation of the expressions on the 3D character.

The text-to animation framework allows producing fast speech animation, which could be very useful in fast phased animation productions, games and VR developments. This



Video 1 (without facial expressions)



Video 2 (with facial expressions)

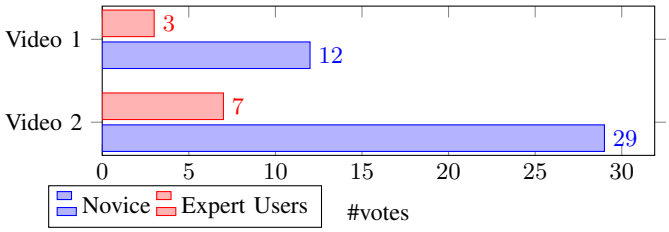



Fig. 15. Animations created from the sentence *"A spider!"* and online survey results

could also provide a foundation for speech animation in feature film productions with the addition of finer controls.

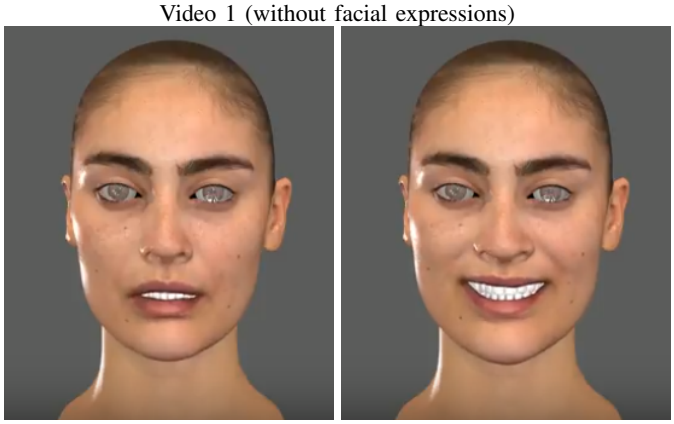
The inclusion of extra tools in the Maya plugin such as the trax editor, range and time controls, facilities to import audio, set or create character sets and a dedicated viewport provided a fast and convenient workflow to the users. These tools implemented together massively increase the productivity. The users rated the plugin with 98% satisfaction, which indicates the success of its design and implementation.

## VI. Conclusions and Future Work

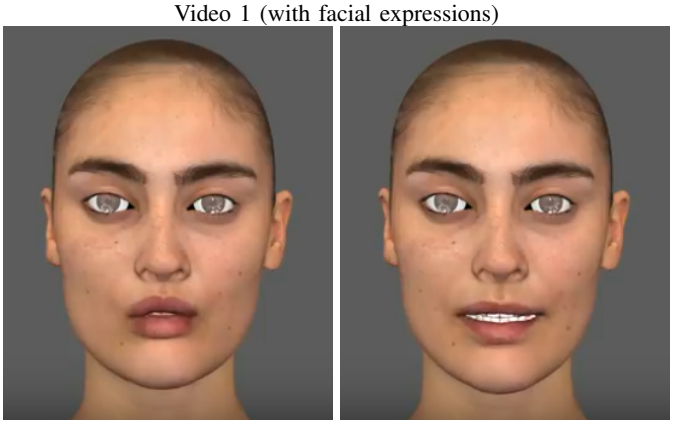
In this paper, a new *text-to-animation* framework for speech animation synthesis is proposed and developed as a Maya plug-in. Further, we investigated the effectiveness of improving the expressiveness of facial animation. The new framework is intuitive and easy to use. An extensive user study is conducted to evaluate the proposed system. The results of

Video 1 (without facial expressions)

Video 1 (with facial expressions)



Video 2 (with facial expressions)



Video 2 (without facial expressions)

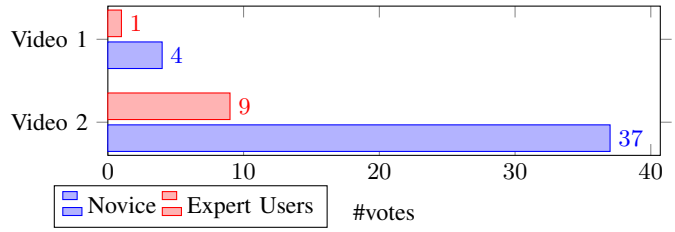



Fig. 16. Animations created from the sentence *"It's sunny outside!"* and online survey results

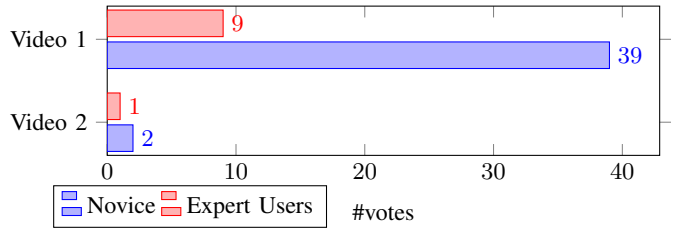



Fig. 17. Animations created from the sentence *"Who stole my cup?"* and online survey results

the survey revealed that the framework is able to 1) Rapidly synthesize facial animation with emotion expressions correctly and 2) improve the expressiveness of lip-sync animations by incorporating facial expressions. In terms of user experience, the developed plugin received 100% user satisfaction indicating its success. In the future, we are interested in expanding the support to other languages.

## Acknowledgment

This work was supported in part by the Royal Society (Ref: IE160609).

## References

[1] E. S. L. Ho, H. P. H. Shum, Y.-m. Cheung, and P. C. Yuen, "Topology aware data-driven inverse kinematics," *Computer Graphics Forum*, vol. 32, no. 7, pp. 61–70, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12212

[2] E. S. L. Ho and T. Komura, "A finite state machine based on topology coordinates for wrestling games," *Computer Animation and Virtual Worlds*, vol. 22, no. 5, pp. 435–443.

[3] ——, "Wrestle alone : Creating tangled motions of multiple avatars from individually captured motions," in *15th Pacific Conference on Computer Graphics and Applications (PG'07)*, Oct 2007, pp. 427–430.

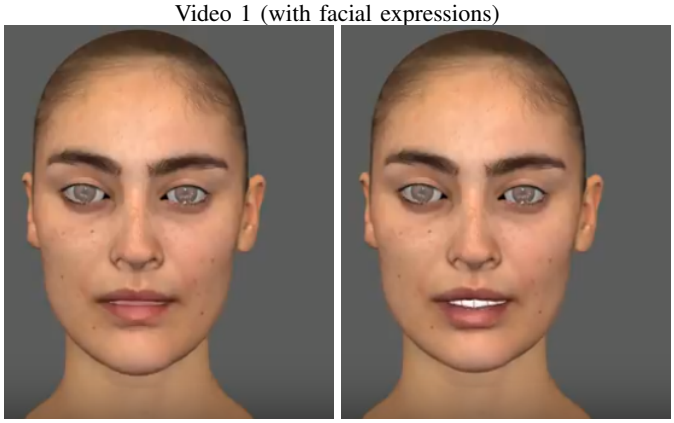
[4] E. S. L. Ho, H. P. H. Shum, H. Wang, and L. Yi, "Synthesizing motion with relative emotion strength," in *ACM SIGGRAPH ASIA Workshop: Data-Driven Animation Techniques (D2AT)*, 2017. [Online]. Available: http://eprints.whiterose.ac.uk/121250/

[5] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969. [Online]. Available: http://science.sciencemag.org/content/164/3875/86

[6] D. Matsumoto and H. S. Hwang, "Evidence for training the ability to read microexpressions of emotion," *Motivation and Emotion*, vol. 35, no. 2, pp. 181–191, Jun 2011. [Online]. Available: https://doi.org/10.1007/s11031-011-9212-2

[7] 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*, ser. SIGGRAPH '97. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1997, pp. 353–360.

Video 1 (with facial expressions)



Video 2 (without facial expressions)

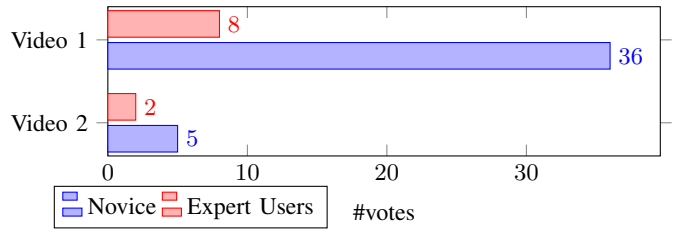



Fig. 18. Animations created from the sentence *"My dog died..."* and online survey results

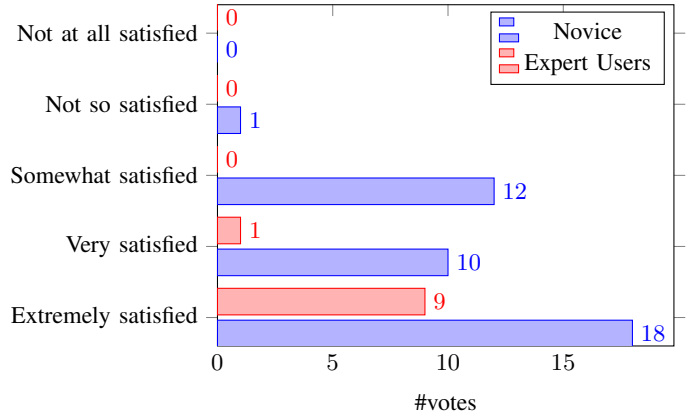



Fig. 19. Satisfaction on combining speech and facial expression for 3D animation production

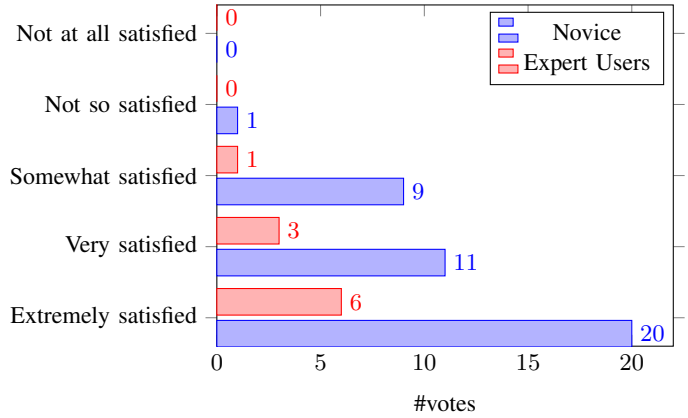



Fig. 20. Overall satisfaction on proposed facial animation synthesis framework


[8] M. Brand, "Voice puppetry," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99. New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., 1999, pp. 21–28.

[9] T.-J. Yang, I.-C. Lin, C.-S. Hung, C.-F. Huang, and M. Ouhyoung, "Speech driven facial animation," in *Computer Animation and Simulation '99*, N. Magnenat-Thalmann and D. Thalmann, Eds. Vienna: Springer Vienna, 1999, pp. 99–108.

[10] Y. Xu, A. W. Feng, and A. Shapiro, "A simple method for high quality artist-driven lip syncing," in *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, ser. I3D '13. New York, NY, USA: ACM, 2013, pp. 181–181. [Online]. Available: http://doi.acm.org/10.1145/2448196.2448229

[11] Y. Xu, A. W. Feng, S. Marsella, and A. Shapiro, "A practical and configurable lip sync method for games," in *Proceedings of Motion on Games*, ser. MIG '13. New York, NY, USA: ACM, 2013, pp. 109:131–109:140. [Online]. Available: http://doi.acm.org/10.1145/2522628.2522904

[12] Y. Chen, F. Huang, S. Guan, and B. Chen, "Animating lip-sync characters with dominated anime models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 9, pp. 1344–1353, Sept 2012.

[13] F.-C. Huang, Y.-M. Chen, T.-H. Wang, B.-Y. Chen, and S.-H. Guan, "Animating lip-sync speech faces by dominated anime models," in *SIGGRAPH '09: Posters*, ser. SIGGRAPH '09. New York, NY, USA: ACM, 2009, pp. 2:1–2:1. [Online]. Available: http://doi.acm.org/10.1145/1599301.1599303

[14] S.-H. Guan, Y.-M. Chen, F.-C. Huang, and B.-Y. Chen, "Lip-synced character speech animation with dominated anime models," in *SIGGRAPH Asia 2012 Technical Briefs*, ser. SA '12. New York, NY, USA: ACM, 2012, pp. 26:1–26:4. [Online]. Available: http://doi.acm.org/10.1145/2407746.2407772

[15] Y. Li, F. Yu, Y.-Q. Xu, E. Chang, and H.-Y. Shum, "Speech-driven cartoon animation with emotions," in *Proceedings of the Ninth ACM International Conference on Multimedia*, ser. MULTIMEDIA '01. New York, NY, USA: ACM, 2001, pp. 365–371. [Online]. Available: http://doi.acm.org/10.1145/500141.500196

[16] EISKO, "Eisko: Digital doubles for entertainment," https://www.eisko.com/, 2018, accessed: 2018-10-30.

[17] Aston University, "Sounds of english," https://www2.aston.ac.uk/lss/research/lss-research/ccisc/discourse-and-culture/west-midlands-english-speech-and-society/sounds-of-english, 2018, accessed: 2018-10-30.