Voice Animator: Automatic Lip-Synching in Limited Animation by Audio

Shoichi Furukawa[1(✉)], Tsukasa Fukusato[3], Shugo Yamaguchi[1], and Shigeo Morishima[2]

[1] Waseda University, Tokyo, Japan
furukawa7246@ruri.waseda.jp, wasedayshugo@suou.waseda.jp
[2] Waseda Research Institute for Science and Engineering, Tokyo, Japan
shigeo@waseda.jp
[3] The University of Tokyo, Tokyo, Japan
tsukasafukusato@is.s.u-tokyo.ac.jp

Abstract. Limited animation is one of the traditional techniques for producing cartoon animations. Owing to its expressive style, it has been enjoyed around the world. However, producing high quality animations using this limited style is time-consuming and costly for animators. Furthermore, proper synchronization between the voice-actor's voice and the character's mouth and lip motion requires well-experienced animators. This is essential because viewers are very sensitive to audio-lip discrepancies. In this paper, we propose a method that automatically creates high-quality limited-style lip-synched animations using audio tracks. Our system can be applied for creating not only the original animations but also dubbed ones independently of languages. Because our approach follows the standard workflow employed in cartoon animation production, our system can successfully assist animators. In addition, users can implement our system as a plug-in of a standard tool for creating animations (Adobe After Effects) and can easily arrange character lip motion to suit their own style. We visually evaluate our results both absolutely and relatively by comparing them with those of previous works. From the user evaluations, we confirm that our algorithms is able to successfully generate more natural audio-mouth synchronizations in limited-style lip-synched animations than previous algorithms.

Keywords: Lip-synching · Limited animations · Animation filtering

1 Introduction

Limited animation (LA) is a traditional hand-drawing technique used to create cartoon animations that reuses some frames instead of redrawing entire frames. The LA technique has the advantage of animators being able to create much more expressive and stylized animations as compared to those generated from 3DCG. This is the reason LA is still popular globally despite the growth of 3D animation. However, it is both time-consuming and costly to produce high quality LA-style cartoons. This is especially true because drawing animated frames

A. D. Cheok et al. (Eds.): ACE 2017, LNCS 10714, pp. 153–171, 2018.
https://doi.org/10.1007/978-3-319-76270-8_12

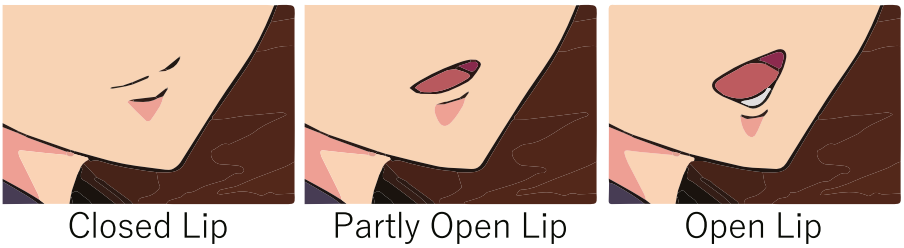

Fig. 1. Examples of key images.

while maintaining lip-audio synchronization is very laborious and requires well-experienced animators.

In general, character lip motion in LA is created by replacing a few lip images (referred to as "key images") that usually represent the closed, partly open and open lip positions (Fig. 1). This approach is similar to that used in flipbooks. Two approaches have been employed to produce speech animations: the "after-recording" and the "pre-scoring" approaches.

The after-recording process basically requires (1) placing key images that follow character script, (2) recording the voice of a voice-actor performing while watching the animated frames, and (3) putting the recorded audio track over the animated frames. However, step (1) is tedious for animators and requires well-trained animators since animators have to imagine suitable lip motions based only on the script. Furthermore, because the speed of voice-actors varies among one another, the created lip motion does not synchronize perfectly with the audio in step (3). As a result, animators are required to repeatedly revise the frames for better lip-audio synchronization. In animation, refining components to address not only lip-audio mismatching but also drawing errors, among others, is called a "retake". According to our interview of the staff of an animation studio, a 30-min animated film generally includes about 200 retakes with 30% of the retakes being related to lip-synching. Furthermore, because audiences are very sensitive to lip-audio de-synchronization, lip-synch retakes are essential for the airing of the film.

On the other hand, pre-scoring is a technique in which animators produce the lip motion based on a pre-recorded voice-acting audio track. In this process, lip-sync retakes could actually be reduced, but the fact remains that animators are still required to repeatedly and laboriously place key images. Furthermore, for dubbing of different languages, the original audio track is just replaced by new translated version resulting in even greater lip-audio de-synchronization.

In this paper, to address these problems, we propose a method that automatically produces lip-synched LA-style animations. Our system does not require any other additional items outside of typical workflows employed in cartoon animation. Our system uses only the voice-actor's voice and a few key images as the input. First, we estimate the motion of the voice-actor's lip from the audio track. We employ a state-of-the-art formant-based method that works in

real-time. Since it is language independent, we can create not only the original animations but also dubbed films. After obtaining the lip motion, we convert it to character lip motion represented by the key images. For this last part, there are two challenging obstacles to producing natural character lip transitions. The first obstacle is how to take correspondences between continuous lip motion and a few discrete key images (we refer to this as the "spatial alignment problem (SAP)"). The second obstacle is in the determination of a suitable timing for the replacement of the key images so that the animation maintains a natural lip movement appearance (we refer to this as the "temporal alignment problem (TAP)"). In this paper, we propose novel methods based on adaptive thresholding and afterimage effects to solve SAP and TAP. In order to faithfully adhere to typical workflows in animation production, we demonstrate our system in Adobe After Effects, which is a standard tool for animators to create animations. In After Effects, animators can easily tune the generated lip transition to suit their personal expressive style. Furthermore, based on the lip transition generated by our system, animators can learn how to draw a natural transition. In summary, our contributions are

- An efficient language-independent system for creating high quality LA-style lip-synch animations using only voice audio and a few key images. This means our system can work for creating both the original animations and dubbed films.
- Algorithms for solving SAP and TAP based on adaptive thresholding and afterimage effects that achieve a natural lip transitions using only a few lip images.
- The possibility of various application both in entertainment production, such as cartoon animations and games, and in the training of animators based on a natural lip motions automatically generated from audio.

In this paper, we evaluate our results both absolutely and relatively by visually comparing them with those of previous works that generate LA-style motion. These previous algorithms can include motionless or flickering mouth motions. From user evaluations, we confirm that animations generated using our proposed techniques results in more natural audio-mouth synchronization.

2 Related Work

2.1 Lip-Sync Animation

Various 3D facial models for cartoon characters have been used in the field of computer graphics. The blendshape technique, which is a common method employed to animate the facial expression of 3D characters, is based on the linear combination of base poses (referred to as "key shapes") such as happiness, sadness, laughter, anger, and more. In this technique, manually controlling the linear parameters is very laborious for creators. To address this, Weise et al. [16] proposed a system that transfers the facial expressions of the actor to various

characters using RGB-D data obtained from professional equipment. Weise et al. [15] demonstrated a real-time system that captures facial expressions by utilizing a commodity depth sensor, and achieved more interactive control of character facial expressions. Moreover, Cao et al. [3] captured facial expressions with high accuracy by combining facial depth data with sparse facial landmarks. These methods work successfully when character key shapes are ready to be used, but the problem remains that key shapes are created for each character in a time-consuming process, which greatly hinders its practical adoption in, for example, film and gaming applications. In addition, although these methods can successfully produce realistic and lip-synched full animations, they are difficult to apply directly in LA-style animation production.

In general, LA-style films are recorded at 24 frames per second, and each drawn image is displayed three times (referred to "on threes") resulting in eight drawings per second. Kawamoto et al. [9] mainly focused on the laborious task in creating character key shapes and proposed lip-synched-animation system, which included a function that converts original animations to LA-style shot on threes. While this converting function can produce LA-like animations, their method cannot be directly applied to produce traditional LA-style lip-synched animations, which consist of a few key images, such as those shown in Fig. 1. This is due to the fact that it is difficult to take correspondences between various visemes, which are the visual units to distinguish sounds, and only a few key images to represent natural character's lip motion.

Other audio-based approaches have also been proposed. Bregler et al. [2] segmented new audio into three sequent phonemes (referred to "triphones") and created realistic speech animations by selecting video frames that correspond to the triphones and stitching them together. Ezzat et al. [7] and Chang et al. [4] demonstrated an alternative approach that constructs a generative model to produce mouth images corresponding to input phonemes. These methods succeeded in creating realistic speech animation, but it remains problematic to create LA-style animations because of the small number of key images to take the correspondences to various phonemes.

2.2 Stylized Cartoon Animation

Approaches to stylize 2D/3D animations have been discussed for a long time. For instance, procedural rules defined by a physical simulation (e.g. squash-and-stretch and temporal effect) have been used for stylization. Kazi et al. [10] developed a 2D sketching system that simplifies the creation of dynamic illustration with motion principals. By using these amplifiers, the user can control the movements and deformations of an underlying background grid. In addition, Dvorovzvnak et al. [6] generated stylized 2D animations by transferring exemplars drawn by artists. These approaches allow the animators to reflect their personal styles by user interaction. On the other hand, some approaches focus on converting any motion, such as 3D skeletal motion, to LA-like style by using a temporal filter [11,12,14]. For example, using motion capture data (MoCap), Kitamura et al. [11] demonstrated a LA-like converter by omitting frames that

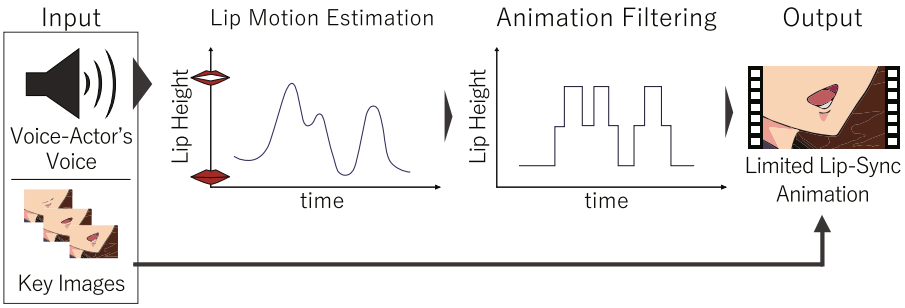

Fig. 2. System overview.

included relatively fast or slow motion. Although these methods can allow for expressive appearance, it is difficult to maintain global timing of the character motion. Therefore, it is hard to achieve suitable audio-mouth synchronization when applying these techniques.

2.3 Lip Motion Capture and Estimation

There are a lot of approaches to obtaining the lip motion of the voice actor. For example, image-processing-based approaches (e.g. facial landmarks detection) is a basic way to capture lip movements. However, it is difficult to produce robust results in environments with unstable or irregular lighting. Moreover, additional efforts are required on the part of the production staff in order to prepare for additional equipments (e.g. video cameras) for capturing in a recording studio. In addition, visual recording risks disturbing the usual performance style of the voice-actor. Phoneme-based methods such as HMM are also common approaches to computing lip motion. However, to use them successfully for a variety of languages, users are required to switch between language-specific phoneme-models. In addition, the reduction in work speed results in increased stress on the animators and therefore working rapidly is essential for software to assist animators. Therefore, in this paper, we employ a state-of-the-art formants-based method that works in real-time to obtain lip motions.

3 System Overview

Our proposed system utilizes only pre-recorded a voice-actor's voice track and three key images ("closed," "partly open," and "open" lip) as inputs, and automatically creates a LA-style lip-synched animation. Our system mainly consists of two steps (refer to Fig. 2): "Lip Motion Estimation" (refer to Sect. 3.1) and "Animation Filtering" (refer to Sect. 3.2). Firstly, in "Lip Motion Estimation", we compute the lip motion of the voice-actor from the input audio track. By using a formant-based method, various languages (e.g. English, Chinese or Japanese)

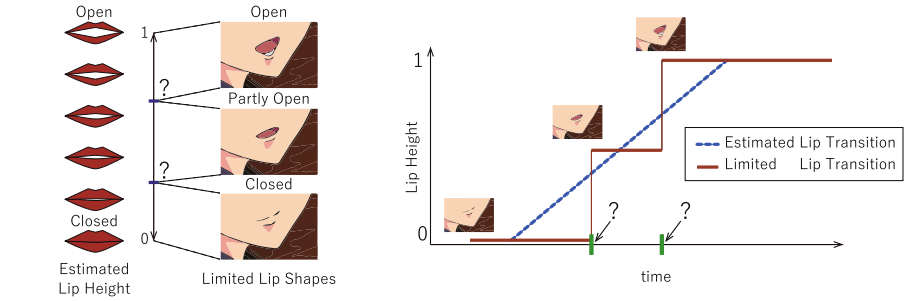

Fig. 3. Overview of SAP and TAP.

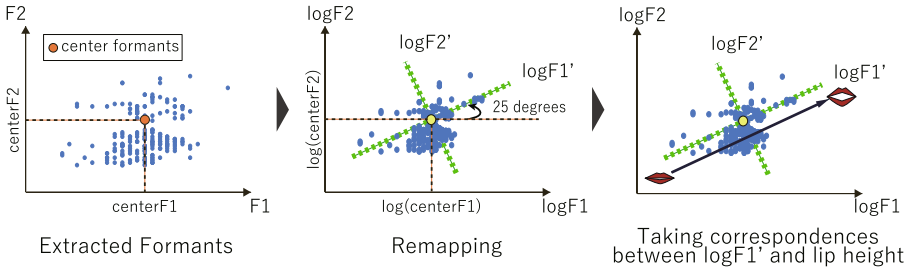

Fig. 4. Overview of lip motion estimation.

can be used here. In "Animation Filtering", our system converts the lip transitions to the LA style. Here, we tackle two challenging problems (Fig. 3). Firstly, for SAP, the difficulty lies in how to take correspondences between continuous distance from upper to lower lip (referred to as "lip height") and a few discrete key images. Secondly, for TAP, a timing has to be determined in which the replacement of the key images results in natural lip motions. Our key idea to solve the SAP is to take correspondences between one peak of the continuous lip transition and one movement of the character's lip by adaptive thresholds. By doing so, the converted lip transition can simulate the voice-actor's lip motion well. To address the TAP, the key idea is to sample lip transitions and perform interpolation while taking into account afterimage effects. By employing these methods, our proposed system can automatically create high quality of lip-synched animations that are comfortable for viewers to watch.

3.1 Lip Motion Estimation

First, we estimate the lip motion from the input voice track. In regard to lip motion capturing, we employ the state-of-the-art formant-based method proposed by Ishi et al. [8]. The original paper aimed at real-time control of the lips of a humanoid-robot. For this paper, however, we modify some parameters of the

method for generating character lip motion at the later steps. Here, we briefly explain its principle concept (refer to Fig. 4) and highlight the modifications that we make.

First, input audio data is pre-emphasized by $1 - 0.97z^{-1}$ to enhance its high frequency and framed by using 32[ms] of hamming window while shifting it by 10[ms]. Then, based on linear predictive coding (LPC), the first and the second formant (denoted as F1 and F2) are extracted. In phonetics, a vowel space, which is the set of (F1, F2) obtained from speaker's voice, differs from speaker to speaker. In order to normalize the difference, the set of (F1, F2) computed from the input voice track is remapped into a log space (logF1 vs. logF2). The origin is moved to speaker-specific center formants, which are the center coordinates of the speaker's vowel space (denoted as (centerF1, centerF2)) resulting in a new origin at (log centerF1, log centerF2). Next, the axes are rotated around the new origin counterclockwise by 25 degrees and the new axes are noted as logF1' and logF2'. As a result, logF1' corresponds to lip height one-to-one. Note that some restrictions are prescribed in the original paper to distinguish between vowels and consonants. Moreover a particular lip height value is used during consonant periods. Furthermore, when a period greater than or equal to 0.2[s] exhibits low power of the input audio signal, the lip height during the period is decayed by multiplying it by 0.9.

In our proposed system, we use a monophonic audio track that is captured at more than 44.1[kHz] and 16[bit]. We set the LPC order as 64. Since the average outline of the vowel spaces is dramatically different depending on the gender, we define the average center formants as specific values in terms of gender: (centerF1, centerF2) = (500 Hz, 1450 Hz) for male and (centerF1, centerF2) = (500 Hz, 1600 Hz) for female. This achieves simplified normalization of individual utterance while preserving fine accuracy by only selecting the gender of the input voice track. This approach is very practical for animators as compared to the original GUI-based normalization techniques. Next we calculate lip height using Eq. (1) with a height scale of 0.5.

$$\text{lip\_height} = 0.5 + \text{height\_scale} * \text{logF1}' \tag{1}$$

We distinguish vowels, consonants, and voiceless periods by simply using the power of the signal (the mean square of signal amplitude) as Table 1. Note that the threshold can vary depending on the recording environment. Next, during periods of uttering consonants, we determine the lip height by taking the product of the previous lip height of a vowel period with a weight parameter α. In this paper, we use $\alpha = 0.5$. Additionally, while human lips would be gradually close after utterance, characters in LA open and close theirs quickly. Considering this, we therefore set the decaying parameter as 0.5 instead of 0.9. Then we normalize the lip height so that it ranges from 0 to 1 and apply a 9-frame sized smoothing filter.

Because the analysis window for LPC is shifted by 10[ms] resulting in the sampling frequency of 100[Hz], we are required to downsample the lip height to 24[fps] for LA-style. Then, we take the $floor(100 * \frac{1}{24} * i)$-th value of the calculated lip height as the i-th value of downsampled one and denote it as an "estimated lip transition (ELT)."

Table 1. Distinction of vowels, consonants and soundless periods.

Periods where power of signal $\geq$ a threshold	$\Rightarrow$ vowels
Periods where power of signal $<$ a threshold for less than 0.2[s]	$\Rightarrow$ Consonants
Periods where power of signal $<$ a threshold for not less than 0.2[s]	$\Rightarrow$ Voiceless

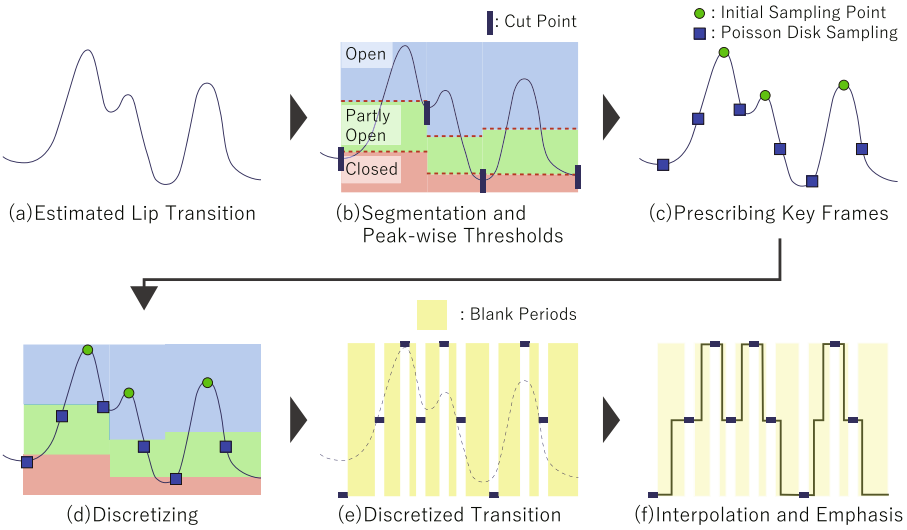

Fig. 5. Processes of animation filtering.

3.2 Animation Filtering

Figure 5 shows the outline of the animation filtering process. In this part, our system converts the ELT obtained in Sect. 3.1 into a natural LA-style lip transition suitable for characters.

Peak Segmentation and Peak-Wise Thresholds. While ELT includes continuous lip shapes, the input key images only represent discrete shapes. Therefore, we need to take correspondences between these continuous and discrete values. One simple solution is to set constant thresholds for entire frames. For example, the quantized lip height at the t-th frame ($\mathrm{L}(t)$) is defined as follows.

$$\mathrm{L}(t) = \begin{cases} 0 & (0 \leq \mathrm{ELT}[t] < \frac{1}{3}) \\ 1 & (\frac{1}{3} \leq \mathrm{ELT}[t] < \frac{2}{3}) \\ 2 & (\frac{2}{3} \leq \mathrm{ELT}[t] \leq 1) \end{cases} \tag{2}$$

where $\mathrm{L}(t)$ corresponds to the key image number (0: closed, 1: partly open and 2: open), and $\mathrm{ELT}[t]$ is the estimated lip height at the t-th frame. However, these thresholds can cause a motionless or flickering appearance resulting in the audience easily identifies the lip-audio de-synchronization. The reason is that the audiences are very sensitive not to the discrepancy between local instant lip shapes and audio, but to that between the lip motion and the audio. Therefore, to approximate the motion represented by ELT, we propose an adaptive method. Inspired by adaptive thresholding for binary images, our key idea is to take correspondences between a local peak of ELT and one movement of character's lip from open to close using locally adaptive thresholds. We first detect abrupt transitions of ELT by placing cut points at i-th frame where $(\mathrm{ELT}[i+1] - \mathrm{ELT}[i]) > 0$ and $\mathrm{ELT}[i] - \mathrm{ELT}[i-1] \leq 0)$. Here, the first frame is regarded as an initial cut point and the last frame is considered to indicate the end of the final segment. Then, we set peak-wise thresholds based on the maximum and the minimum values in each segment as below.

$$\mathrm{threshold1}[n] = \min[n] + \tau_1 \tag{3}$$

$$\mathrm{threshold2}[n] = \frac{(\max[n] - \min[n])}{2} + \min[n] + \tau_2 \tag{4}$$

where $\mathrm{threshold1}[n]$ and $\mathrm{threshold2}[n]$ means lower and upper thresholds from the n-th to the $(n+1)$-th cut points respectively. Note that the last terms τ_1 and τ_2 are the factors to avoid the overlap of threshold1 and threshold2 that results when the maximum and the minimum are equivalent (in this paper, we set $\tau_1 = 0.01$ and $\tau_2 = 0.02$). Lastly, using Eq. (5), ELT is quantized to the number of key images (as shown in Fig. 5(b)).

$$\mathrm{L}(t) = \begin{cases} 0 & (0 \leq \mathrm{ELT}[t] \leq \mathrm{threshold1}[n]) \\ 1 & (\mathrm{threshold1}[n] < \mathrm{ELT}[t] \leq \mathrm{threshold2}[n]) \\ 2 & (\mathrm{threshold2}[n] < \mathrm{ELT}[t] \leq 1) \end{cases} \tag{5}$$

when the n-th cut point $\leq t <$ the $(n+1)$-th cut point.

Key Framing and Discretization. We determine key frames, where the ELT's values are discretized based on the peak-wise thresholds defined at the previous step. We consider frames with local maxima as the crucial frames to achieve lip-audio synchronization. These frames are initially taken as key frames (as shown in Fig. 5(c): "Initial Sampling"). Next, as inspired by Dunbar et al. [5], we apply Poisson Disk Sampling [5] to ELT in order to pick up other key frames (as shown in Fig. 5(c): "Poisson Disk Sampling"). From our experience, a disk size of five frames can provide the resolution that is similar to LA-style animation. Finally, at the prescribed key frames, our system quantizes ELT by taking the thresholds as reference (Fig. 5(d) and (e)).

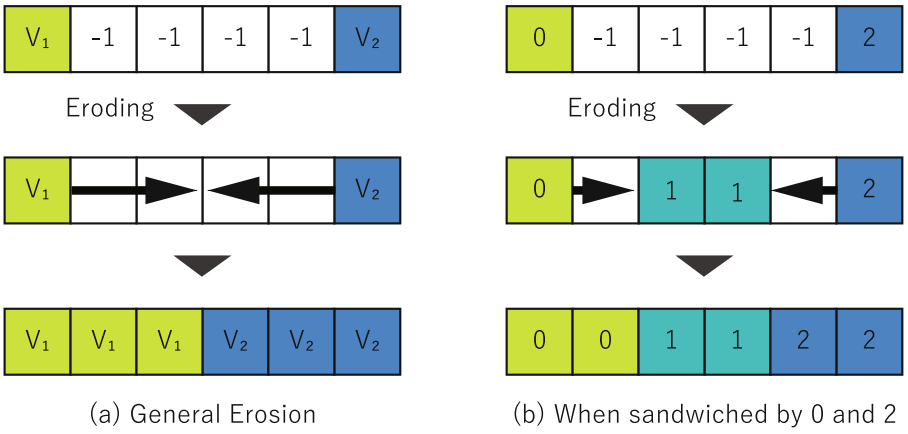

Fig. 6. Example of interpolation.

Interpolation and Motion Emphasis. In limited animations, it is because of the afterimage effect that results by replacing frames (as is done for a flipbook) that viewers have the impression that the character's lip are moving. In addition, the afterimage effect causes a viewer perception delay. That is to say, when image "A" switches to image "B" at time t, viewers generally feel the change a little later, at time $t + \delta$. Considering this effect, we interpolate the frames between key frames. These interpolated frames are referred to as "blank frames".

We first dilate the values at key frames like "image dilation". Here, the values of the blank frames are represented as -1, and the size of the dilation window is 2 frames, which means the dilated value at the i-th frame is the maximum value from the i-th to the $(i+1)$-th frame. Furthermore, each of remaining blank periods is interpolated by eroding it using its head and tail values (denoted as V_1 and V_2 in Fig. 6). Note that when the blanks are sandwiched by the values 0 and 2, the center frame(s) is(are) interpolated by a value of 1 for smooth transitions. As a result of these processes, the interpolated transition precedes ELT a little resulting in a natural LA-style lip motion.

In the discretized transition, a value except 0 sometimes lasts for a long periods over several frames due to the interpolation. This is especially true because the dilation can erode the smaller value. Our solution for this has the common idea with a limited-style-converting algorithm proposed by Kawamoto et al. [9], in which, in each constant period, one key frame whose mouth shape is different from that of the previous period is selected. We emphasize the transition by inserting different values in periods where the same value of the quantized lip height lasts (we denote the length of such a period as l), as shown in Table 2. Here, the reason for processing $9 \leq l$ and $6 \leq l < 9$ separately, and for the number of inserted frames is to generate a similar resolution to that of the on-threes method. After this emphasizing process, the final animation is produced by placing key images following the obtained transition (Fig. 5(f)).

Table 2. Emphasis of Lip Transition.

if the value is 2			
when $9 \leq l$	$\Rightarrow$	inserting three frames of "1" centering the middle frame.	(Fig. 7(a))
when $6 \leq l < 9$	$\Rightarrow$	inserting "1" at the half of the period.	(Fig. 7(b))
if the value is 1			
when $9 \leq l$ and the post value is 0	$\Rightarrow$	inserting three frames of "2" centering the middle frame.	(Fig. 7(c))
when $9 \leq l$ and the post value is 2	$\Rightarrow$	inserting three frames of "0" centering the middle frame.	(Fig. 7(d))
when $6 \leq l < 9$	$\Rightarrow$	inserting "0" at the half of the period.	(Fig. 7(e))

4 Results and Discussion

We created 36 limited animation films using our method. In these animations, male voice-actors spoke 18 audio tracks (three scripts in English, Chinese, and Japanese, each). Female voice actors provided 18 as well. Each of the sentences was selected from the 1st to 4th set of "Harvard Sentences [13]" at random and used without change for English or being translated for Chinese or Japanese. The sampling frequency of each audio track was 48[kHz]. The threshold mentioned in Sect. 3.2 was 10^4 and the other used parameters were the same as those referred in Sects. 3.1 and 3.2. Figure 9 describes one of our results. In Fig. 9, the first row shows the input audio track and the script. The second row shows the voice-actor's lip transition estimated by the method described in Sect. 3.1 and the third is the obtained limited transition. Comparing the second and the third, although our result is discretized, it simulates the similar motion to the original transition well.

From related works, we select three algorithms that can be expanded in order to produce traditional LA-style lip-synched animations composed by key images. We compare our result with the transitions created by these algorithms. In Fig. 9, the forth row shows the result generated by using an algorithm for converting to LA-style that was proposed by Kawamoto et al. [9]. Here, we treat the frames in vowel periods as key frames and our adaptive thresholding is used for the quantization. After applying the converting algorithm and linearly interpolating between key frames, the result transition is then obtained by simply thresholding the interpolated values ($0 \leq$ the value $v \leq 0.5$ to 0, $0.5 < v \leq 1.75$ to 1 and $1.75 < v$ to 2). Furthermore, the forth and the bottom row in Fig. 9 show the results generated by using the methods of Kitamura et al. [11] and Morishima et al. [12] respectively. Note that each result also employs our adaptive thresholds for the quantization. In addition, for the method proposed by Kitamura et al. [11], as described in the original paper, we set 4 as the number of frame-omitting process and 1/3 of the total number of frames for frame-holding process. In Fig. 9, a still period can be observed in both Kawamoto et al. and Kitamura et al. results. This results in a motionless appearance. In addition, spiky motion exists in Kitamura et al. and Morishima et al. results, which results in a flickering

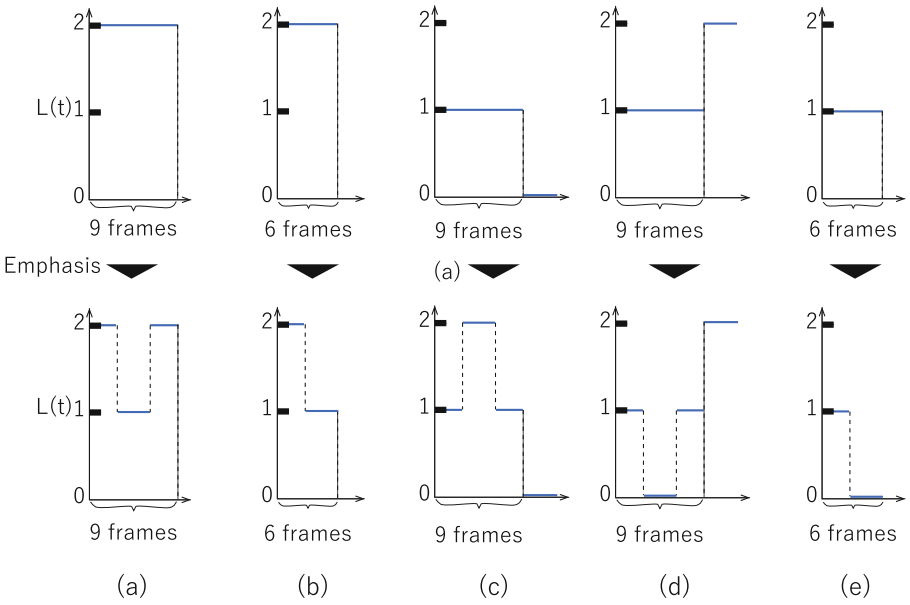

Fig. 7. Emphasis of transition.

appearance. Furthermore, the lip-moving period of the result in Morishima et al. does not match with the original audio track.

In the following parts, we discuss additional experiments that were conducted to visually evaluate our results both absolutely and relatively as compared to previous works.

4.1 User Study A: Naturalness

To assess the naturalness of our results, we performed a crowdsourced experiment. We invited 100 participants and paid a fixed sum for the experiment regardless of the quality, i.e., 0.10 U.S. dollars. The evaluation period lasted approximately 10 min. for each participant. First, we showed the 18 male animations generated by our proposed method. Then, the participants completed a survey regarding the lip-audio sync quality. The answers were scored on a seven-point Likert scale wherein a score of 1 was noted as "highly unnatural" and a score of 7 as "highly natural". We also conducted a scoring experiment for the 18 female animations in the similar manner.

Figures 10 and 11 show the results of the survey. The obtained scores were positive. In fact, all of the average scores except Q8 in Fig. 11 were larger than 4, which was noted as "neither natural nor unnatural". From these quantitative absolute evaluations, we confirmed that the participants were satisfied with the quality of our results. In regard to the female Q8, the animation included a periodic lip motion resulting in the relatively motionless appearance. This can be

because the sampling interval mentioned in Sect. 3.2 was a little longer to maintain the original motion. In this situation, our method can be improved by interactive control of the sampling intervals.

4.2 User Study B: Our Method Vs. Previous Methods

We randomly selected (2 English + 2 Chinese + 2 Japanese) × 2 gender = 12 films from the 36 results mentioned at the beginning of this section. Then, we invited 15 participants to compare them with animations generated using the (1) Kawamoto's method, (2) Kitamura's method, and (3) Morishima's methods described above. Each question required the participants to watch a video wherein four animations generated by the four different methods were placed in random order. The participants were asked to decide how "natural" they looked. Note that this study was conducted independently of "User Study A" and a seven-point Likert scale was used for each task where a score of 1 is noted as "highly unnatural" and a score of 7 as "highly natural".

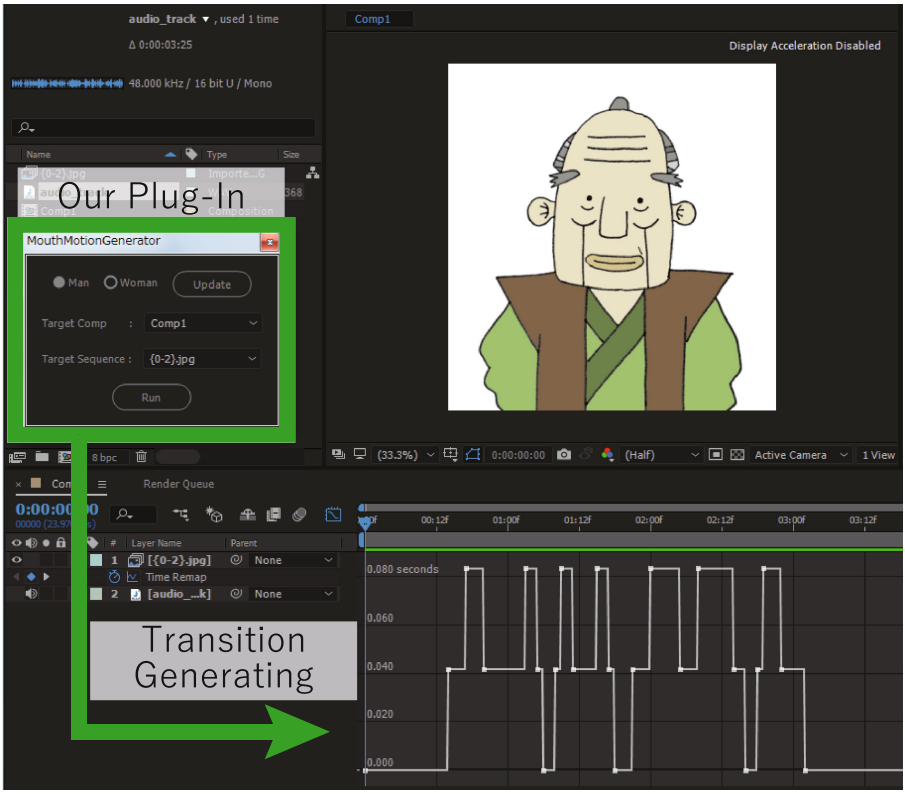

Fig. 8. Our Plug-In.

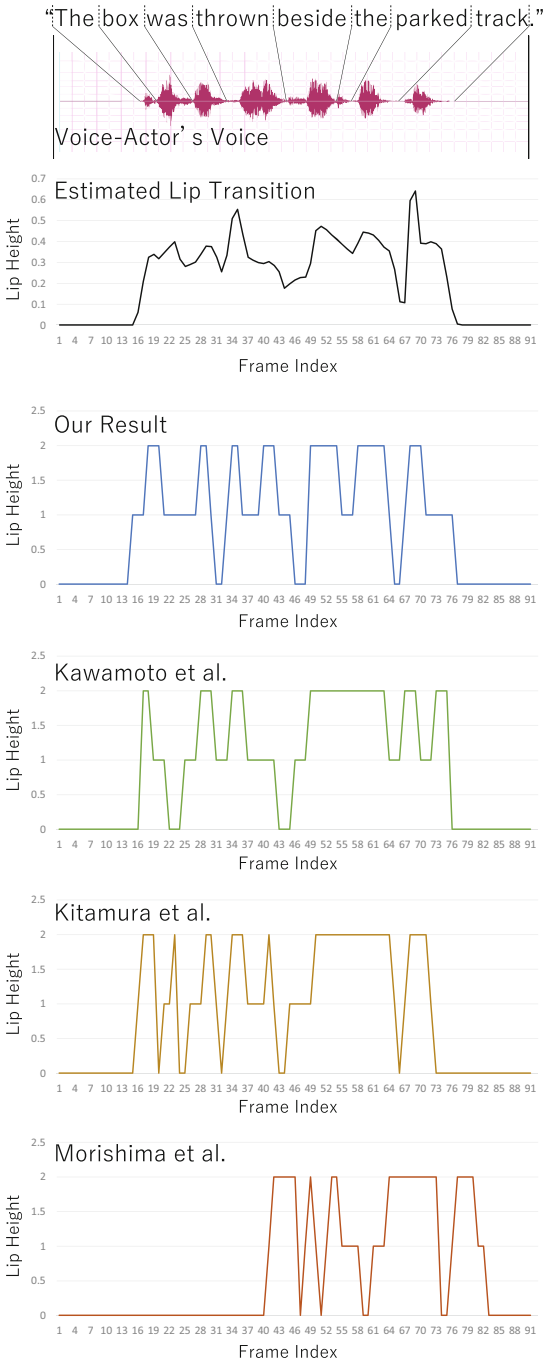

Fig. 9. Our result and previous methods.

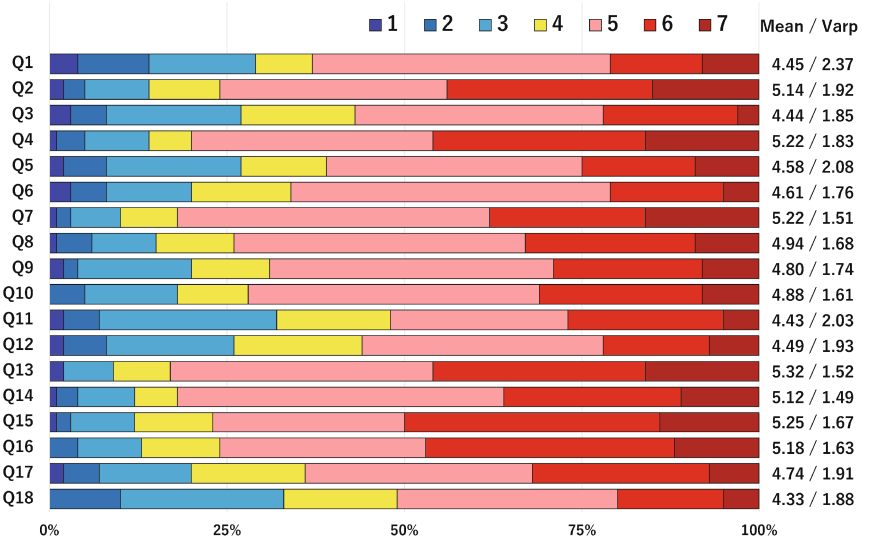

Fig. 10. Scores of user study a: male.

Figure 12 shows the results of an analysis of the answers. It is confirmed that our results have higher average scores as compared to the previous methods. Furthermore, we calculated p-values by running a Wilcoxon signed-rank test. For this, we used a function implemented in the "coin" package of R language [1].

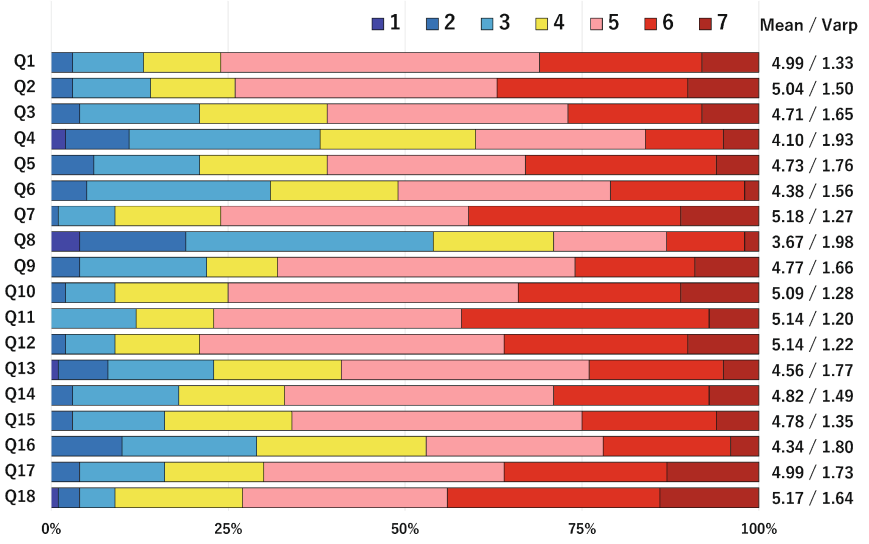

Fig. 11. Scores of user study a: female.

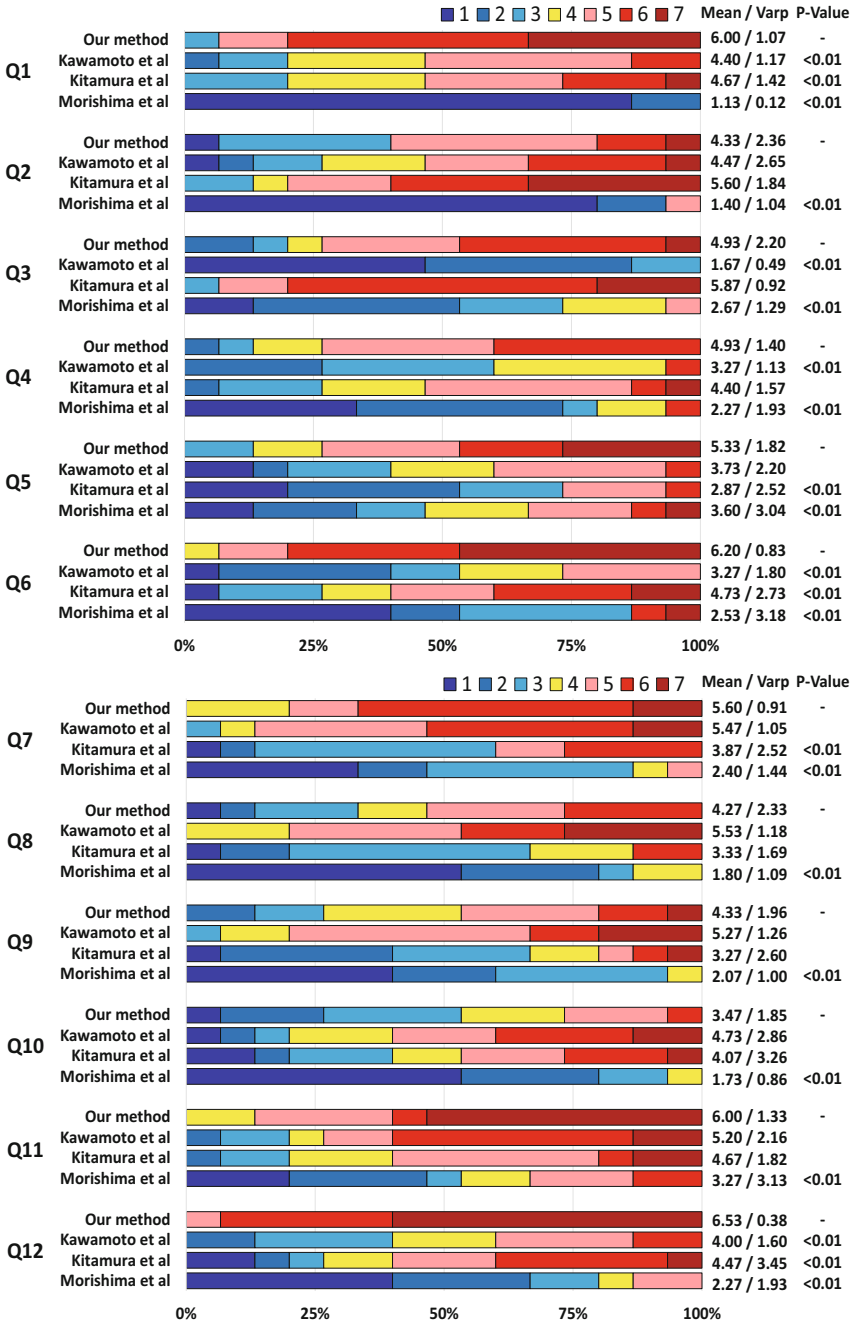

Fig. 12. Comparison between our method and previous methods with a 7-point likert scale.

As the result, 62.5% of questions wherein our results have higher average scores as compared to Kawamoto et al., 55.6% of those to Kitamura et al., and 100% of those to Morishima et al. have p-values that are less than 0.01. We conclude that in these questions, the scores were significantly different. In a few questions, although from the statistical test significant difference was not detected, our results received relatively lower scores than Kawamoto et al. or Kitamura et al. results. This can be because the quantized lip values at some consecutive key frames were the same resulting in the motionless or periodic appearance. To address this situation, our system requires to iteratively apply the motion emphasis process or implement interactive controls of the interval of sampling. As mentioned above, however, in these questions, the statistical test did not show the significant difference. Therefore, we plan to conduct additional evaluations with more samples to verify the dependencies on scripts.

5 Implementation

In addition, we implemented our method as a plug-in for Adobe After Effects, which is a standard tool following the workflow of creating cartoon animations (shown in Fig. 8). In regards to user interaction, the user first loads the key images and a recorded voice track and simply selects the gender through a button click. The changing of the gender just switches center formants, as described in Sect. 3.1. Then, by clicking the "run" button, our system automatically calculates the natural lip transition of the character, and the key images are placed following the transition. Modifications of the generated transition is intuitive and is performed interactively. Because our method is quick and because it fits into existing workflows, animators can find use of the proposed system without much disturbance. In fact, we received positive feedback from users of the system prototype. For example, an amateur animator thought that the rapid performance contributes to its usability and that it is especially convenient when generating long speech animations.

6 Conclusions and Future Work

In this paper, we proposed a new method to automatically generate limited animation (LA) style lip-synched animations using only a voice-actor's audio track and a few images that represent closed, partly open, and open lip. Our key ideas to generate natural lip transition are based on (1) taking correspondences between the continuously changing lip height of the voice-actor and the discretized key images using adaptive thresholds and on (2) sampling and interpolating lip transitions while taking into account afterimage effects. Our method (1) solves the spatial alignment problem and (2) solves the temporal alignment problem. We obtained highly positive scores about the appearance of our results. Additionally, highly positive scores were obtained when our results were compared to those using previous methods. These scores mean that our results can

achieve high-quality audio-lip synchronization in LA-style animations. In addition, the results, which took into account various languages (English, Chinese, and Japanese), strongly suggest that our system can work for expanded multi-language productions. This means that animators can produce animations with less concern for the target language since the proposed method automatically addresses this aspect. We also demonstrated that our method can be successfully implemented as a plug-in in Adobe After Effects, which is a widely used tool in the field of animators. As a result, we propose that our system faithfully adheres to existing workflows in animation.

One of the limitations of our method is that it is difficult to directly apply the method for generating speech animations that include head rotations since lip shapes differ from frame to frame. However, as our method mainly focuses on generating a natural character lip transitions, animators can use the generated transition as a drawing guide. In practical scenes, the number of speech frames in which the character's head rotates is much less than that without head rotation because drawing frame by frame is time-consuming and costly. In addition, such frames require well-trained animators and they allow for personal influences. Practically, our method would be used as follows. For scenes without character head rotation, animators can efficiently produce high-quality lip motions using our method. For scenes with character head rotation, the system can serve as a guide for animators. As a part of future work, and with the aim of improving our system, image interpolation methods can be implemented to automatically generate speech animations with head rotations.

In addition, LA-style animations sometimes have speech scenes using multiple (more than three) key images to exaggerate character lip motions. In order to be able to apply our method to these types of scenes, we plan to (1) distinguish vowels and consonants more accurately by taking into account formant bandwidth, to (2) separate consonants into sub-categories (e.g. fricatives, laterals, or plosives) and to (3) take fine correspondences between the mouth shapes of the voice-actor and the key images.

There is demand for online lip-synching of LA characters. For example, in the testing of character lip motion for animation directors, in the development of remote avatars for communication entertainment and more. In this paper, we proposed a quick method for LA-style lip-synched animations. The system can be expanded for the use in real-time applications. In future works, we aim to optimize our algorithm for online applications.

Acknowledements. This work was supported in part by the Japanese Information-Technology Promotion Agency (IPA), JST ACCEL Grant No. JPMJAC 1602, and JSPS Grant No. 17H06101, Japan.

References

1. coin: conditional inference procedures in a permutation test framework. https://cran.r-project.org/web/packages/coin/index.html. Accessed 22 Oct 2017
2. Bregler, C., Covell, M., Slaney, M.: Video rewrite: driving visual speech with audio. In: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, pp. 353–360. ACM Press/Addison-Wesley Publishing Co. (1997)
3. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. (TOG) **33**(4), 43 (2014)
4. Chang, Y.J., Ezzat, T.: Transferable videorealistic speech animation. In: Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 143–151. ACM (2005)
5. Dunbar, D., Humphreys, G.: A spatial data structure for fast poisson-disk sample generation. ACM Trans. Graph. (TOG) **25**(3), 503–508 (2006)
6. Dvorožňák, M., Bénard, P., Barla, P., Wang, O., Sỳkora, D.: Example-based expressive animation of 2D rigid bodies. ACM Trans. Graph **36**(4), 10 (2017)
7. Ezzat, T., Geiger, G., Poggio, T.: Trainable videorealistic speech animation. ACM Trans. Graph. (TOG) **21**(3), 388–398 (2002)
8. Ishi, C.T., Liu, C., Ishiguro, H., Hagita, N.: Speech-driven lip motion generation for tele-operated humanoid robots. In: Auditory-Visual Speech Processing 2011 (2011)
9. Kawamoto, S.I., Yotsukura, T., Anjyo, K., Nakamura, S.: Efficient lip-synch tool for 3D cartoon animation. Comput. Anim. Virtual Worlds **19**(34), 247–257 (2008)
10. Kazi, R.H., Grossman, T., Umetani, N., Fitzmaurice, G.: Motion amplifiers: sketching dynamic illustrations using the principles of 2D animation. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (2016)
11. Kitamura, M., Kanamori, Y., Mitani, J., Fukui, Y., Tsuruno, R.: Motion frame omission for cartoon-like effects. In: Proceedings of International Workshop on Advanced Image Technology (IWAIT), pp. 148–152. KSBE (2014)
12. Morishima, S., Kuriyama, S., Kawamoto, S., Suzuki, T., Taira, M., Yotsukura, T., Nakamura, S.: Data-driven efficient production of cartoon character animation. In: ACM SIGGRAPH 2007 Sketches, p. 76. ACM (2007)
13. Rothauser, E.: IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. **17**, 225–246 (1969)
14. Wang, J., Drucker, S.M., Agrawala, M., Cohen, M.F.: The cartoon animation filter. ACM Trans. Graph. (TOG) **25**, 1169–1173 (2006)
15. Weise, T., Bouaziz, S., Li, H., Pauly, M.: Realtime performance-based facial animation (TOG). ACM Trans. Graph. **30**, 77 (2011)
16. Weise, T., Li, H., Van Gool, L., Pauly, M.: Face/off: live facial puppetry. In: Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pp. 7–16. ACM (2009)