Journal of Information Processing Vol.22 No.2 401–409 (Apr. 2014)

[DOI: 10.2197/ipsjjip.22.401]

Recommended Paper

Data-Driven Speech Animation Synthesis Focusing on Realistic Inside of the Mouth

Masahide Kawai[1,a)] Tomoyori Iwao[1,b)] Daisuke Mima[1,c)]
Akinobu Maejima[1,d)] Shigeo Morishima[1,e)]

Received: October 2, 2013, Accepted: December 4, 2013

Abstract: Speech animation synthesis is still a challenging topic in the field of computer graphics. Despite many challenges, representing detailed appearance of inner mouth such as nipping tongue's tip with teeth and tongue's back hasn't been achieved in the resulting animation. To solve this problem, we propose a method of data-driven speech animation synthesis especially when focusing on the inside of the mouth. First, we classify inner mouth into teeth labeling opening distance of the teeth and a tongue according to phoneme information. We then insert them into existing speech animation based on opening distance of the teeth and phoneme information. Finally, we apply patch-based texture synthesis technique with a 2,213 images database created from 7 subjects to the resulting animation. By using the proposed method, we can automatically generate a speech animation with the realistic inner mouth from the existing speech animation created by previous methods.

Keywords: Detai-lization, inner mouth, skull bone, phoneme combination, speech animation

1. Introduction

Creating realistic and convincing speech animations is still one of the most important issues in movie and video game productions [1]. In general, high quality speech animations require professional skills, as well as considerable time and effort, because of the highly complex appearance changes that must be achieved in and around the mouth, particularly for photorealistic human characters.

To solve this problem, some researchers have proposed speech animation synthesis techniques. Some representative methods include the three-dimension model-based methods (3D method), and two-dimension image-based methods (2D method). The 3D methods blendshape with several shape models [2], [3] or retarget real expressions with a motion capturing system [4], [5], [6], [7]. 2D methods synthesize mouth motions using a prepared video corpus [8], [9], [10], [11], [12]. Both methods create speech animations with realistic lip movements, however, the resulting detail of the inner mouth is inadequate due to the complex changes in appearance that must be accommodated. In other words, despite the generation of high-quality lip movements, traditional methods still cannot provide realistic inner mouth animations. The quality of original animations could be significantly improved by improving the quality of the inner mouth animations through post-processing.

Therefore, we propose a method to automatically synthesize a novel speech animation by embedding a photorealistic inner mouth into an original animation. Although the concept is straightforward, mismatches between the internal and external mouth images have so far precluded the use of our method in movie and video game productions. The mismatches are manifested in two primary forms: a sharp boundary between the inner and outer mouth, and a difference in luminance. These mismatches have been addressed here through the combined use of *Detai-lization* (advanced Visio-lization [13]) and seamless transitions [14]. Detai-lization is a method that generates novel images that can be applied to detailed areas like uneven teeth, and sequential images like animations. This approach is an improvement over Visio-lization, which was only applicable to still image, not detailed areas or sequential images. In this paper, we demonstrate that a photorealistic inner mouth animation can be generated by combining the benefits of Detai-lization and the seamless transition method.

Our main contribution is a post effect filter that improves the quality and reality of original animations. As shown in **Fig. 1**, our method can be applied to a wide range of speech animations, including real people or computer generated (CG) characters.

2. Related Work

As previously mentioned, there are several methods for creating speech animations. Chang et al. [12] developed a system that can generate speech animations that transfer the speaking style

[1] Waseda University, Shinjuku, Tokyo 169–8555, Japan
[a)] doara-waseda@toki.waseda.jp
[b)] sazabi@akane.waseda.jp
[c)] ai-zumi@ruri.waseda.jp
[d)] akinobu@mlab.phys.waseda.ac.jp
[e)] shigeo@waseda.jp

The initial version of this paper was presented at the Visual Computing/Graphics and CAD joint symposium held on June 22-23, 2013, which was sponsored by SIGCG. This paper was recommended to be submitted to Journal of Information Processing (JIP) by chairman of SIGCG.

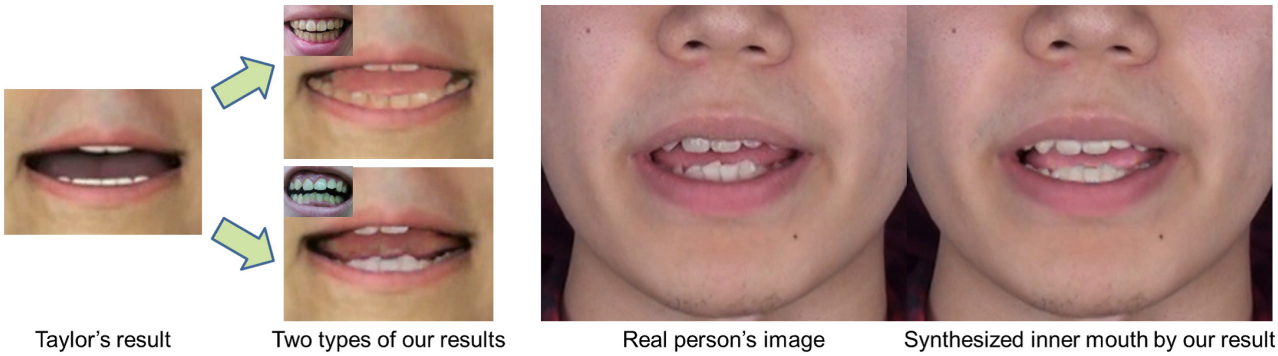

Fig. 1 Example shot of speech animation. Left: Comparison with the result from "Dynamic Units of Visual Speech" [15]. Realistic inner mouth synthesis achieved using small database. Right: Comparison with actual image. Photorealistic inner mouth expression is achieved, closely matching the real person's image and validating the proposed approach.

of one person to another using a multidimensional morphable model. Unfortunately, inner mouth appearances in the resulting animations are expanded and contracted because the inner mouth morphs along with the lip movements.

Taylor et al. [15] proposed a method for lip synchronization that achieves realistic lip movements by connecting sequences of active appearance model (AAM) parameters based on phonetic information. Though this method can generate natural lip movements, the inner mouth quality depends entirely on that of target shapes made by skilled artists using AAM parameters. Li et al. [16] developed a system to generate facial blendshape rigs with a small number of training poses. As with Taylor et al. [15]'s method, speech animation quality also depends on target shapes. For both of the above works, complex inner mouth appearances could not be generated from the resulting videos.

Motion capture systems are often used to create facial animations. For example, Huang et al. [17] and Bickel et al. [18] both used motion capture systems to create facial animations. However, the inner mouth was blank in all of the facial animations because the systems are unable to capture inner mouth data.

As a whole, methods available to create speech animations are unable to represent the inner mouth areas adequately (e.g., blurred or blank inner mouth appearances). To represent inner mouth animation, King et al. [19] used the tongue model, which is composed of a B-spline surface with 60 control points. However, tongue movements could not be represented accurately. To represent tongue movement accurately, a tongue simulation was proposed using the 3D finite element method (FEM) [20], [21]. However, the computational cost associated with tongue simulations is relatively large and the 3D-based method does not produce photorealistic tongue appearances. Therefore, an actual tongue database is used here. Rather than proposing a method to create speech animations, this paper describes a method to improve the quality of inner mouth animations.

3. Data Acquirement

This section describes the three inputs and two databases required to perform the proposed method.

Inputs:

(1) An original animation
(2) A frontal image of the teeth
(3) A syllabic decomposition of the speech

Databases:

(1) Sets of tongues-database
(2) Mouth-database

In general, teeth, tongues, and lips move independently of one another. For example, you can move your tongue without moving your teeth or lips. To represent these complex inner mouth movements with a small number of inputs and databases, the inner mouth is treated as dissociated teeth and a tongue. In fact, it would be extremely difficult to construct a complete inner-mouth database that handles the teeth and tongue simultaneously. Since one's teeth shape is uniform during speech, teeth movement can be generated using a single image. Because the tongue moves intricately during speech, tongue movements are determined through the use of a syllabic decomposition of the speech and a tongue database. By combining these two movements, a wide range of English inner mouth movements can be represented. Since the tongue moves in a specific pattern to annunciate vowels and consonants, tongue movements are treated as a series of movements. Consequently, the sets of tongues-database are composed of many consecutive images. Detai-lization (discussed in a later section) uses a mouth-database to minimize discontinuities between independently inserted teeth and tongue images. In Section 3.1, we provide details about each of the inputs. The two databases are then described in Sections 3.2 and 3.3.

3.1 Input Data

As indicated above, three inputs are needed for operation.

To begin, an original animation is required. Our method is intended to serve as an upgrade for original animations by improving the appearance of the inner mouth, thereby providing a realistic inner mouth animation while maintaining the benefits of the original animation, such as realistic lip movement, detailed wrinkles, etc. The inner mouth region is extracted from the original animation manually. Also,the proposed method is capable of supporting 2D translation and angular rotation. The technique is applicable to any size animation; for the examples used in this paper, the input animation was 512×512 pixels.

Second, an image of the speaker's teeth is required (*frontal*

Table 1 Classification of the tongue appearances for vowels.

Name	Tongue's appearance	Class	Ex.
Front vowel	Forward of inner mouth	1	/e/
		0	/i/
Back vowel	Backward of inner mouth	0	/a/

Table 2 Classification of the tongue appearances for consonants.

Name	Tongue's appearance	Class	Ex.
Bi-labial	Between upper and lower lip	0	/p/
Labio-dental	Between upper teeth and lower lip	0	/f/
Dental	Between upper teeth and apex linguae	1	/θ/
Alveolar	Between upper alveolar arch and apex linguae	1	/t/
Palato-alveolar	Between portion passing hard palate from alveolar arch and tongue's tip	1	/r/
Palatal	Between hard palate and forward of lingual surface	1	/j/
Velar	Between soft palate and forward of lingual surface	0	/k/
Glottal	Between vocal cords	0	/h/

teeth image). With only this single image, animations can be generated that depict the speaker opening or closing their teeth during speech. The image size is adjusted to fit the animation size (described in Section 4.1). Therefore, the frontal teeth image can be of arbitrary size.

Finally, the syllabic content of the subject's speech must be converted to text. For example, if the subject utters a /te/ syllable in the 36th frame, then "te: 36" is stored in a text file.

3.2 Sets of Tongues-Database

Sets of consecutive tongue images for an arbitrary subject pronouncing *phoneme combinations* were acquired. The use of phoneme combinations preserves original continuous tongue movements as much as possible. In this paper, phoneme combinations are newly defined according to the visibility of the tongue; the phoneme combination must be described as the tongue is not visible at the beginning, but appears in the middle, and then disappears at the end. These combinations must contain all tongue movement variations that appear in spoken English. Therefore, we classified the tongue appearance when each subject uttered vowels and consonants in spoken English [22]. If a tongue is visible, the class is 1; otherwise, the class is 0. The tongue appearance classification for vowels and consonants are shown in **Tables 1** and **2**, respectively. In addition to the tongue appearance classifications shown in Tables 1 and 2, we also combined vowels with consonants to determine phoneme combinations, for example, /i/-/t//e/-/i/, /f/-/e/-/b/, etc. A total of 149 tongue movement variations exist for spoken English. Tongue movements were recorded from a subject pronouncing all 149 phoneme combinations, resulting in 149 tongue image sets labeled with phoneme combinations (/i/-/t//e/-/i/, /f/-/e/-/b/, etc.). The classification of all 149 tongue movements (phoneme combinations) is motivated

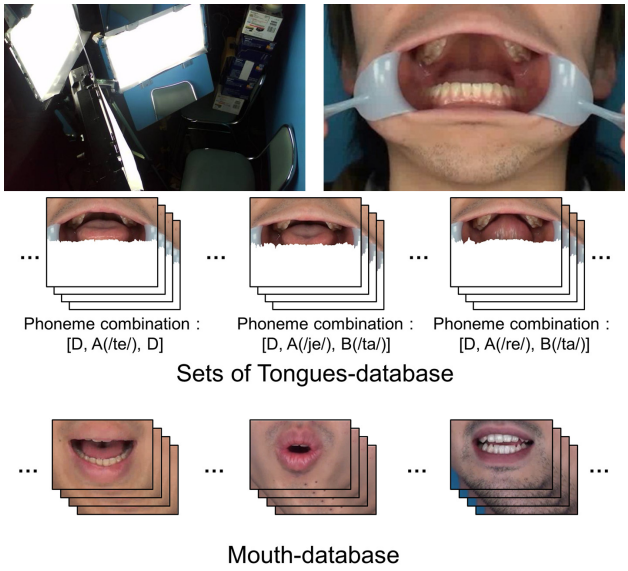

Fig. 2 A recording conditions and an image captured while using an Angle Wider and some representative captured database images.

in Section 4.3. To capture the image sets for the tongue database, we used an *Angle Wider* tool. This tool allows more accurate capture of tongue movements. Recording conditions and an image captured while using an *Angle Wider* and some representative captured database images are shown in **Fig. 2**. Although the Angle Wider does make it difficult for the subject to speak naturally, the inner mouth motions can be typically correct, particularly if the subject uses references (Tongue's appearance), such as Tables 1 and 2.

After capturing all 149 tongue movements, the tongue-database is constructed. The boundary between the lower teeth and tongue image is obtained using the mean shift method [23], and tongue images are then separated from the captured images as shown in Fig. 2.

3.3 Mouth-Database

We captured videos of seven subjects (except for the speaker in the input animation) pronouncing vowels (/aiueo/) and representative symbolic sounds (/te/, /re/, /je/, /θe/, /fa/, /va/) produced by different articulation regions. We used 2,213 images from the captured video to apply the Detai-lization method (later discussion).

The database images were 241 × 201 pixels. Images included movement of the overall mouth from the upper and lower lips. We constructed databases for each of the regions as described later in the text.

4. Creating a Post-Effect Filter

An overview of our method is shown in **Fig. 3**. This section describes the inner mouth reanimation technique. The method is composed of four steps. First, images of the upper and lower teeth are created from a frontal teeth image. Second, the teeth images are embedded into sequence images captured from the original animation. Third, tongue movements are estimated and embedded into sequence images using the desired syllabic content and tongue databases. Next, the images around the mouth are recon-

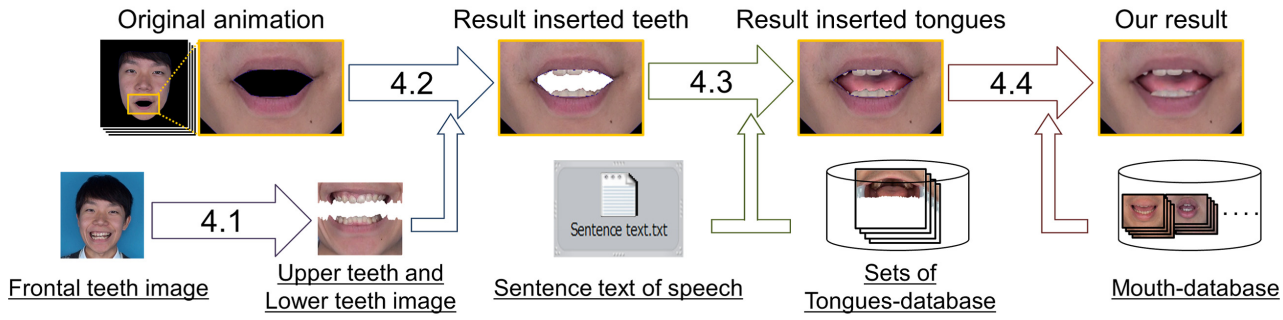

Fig. 3 An overview of our method.

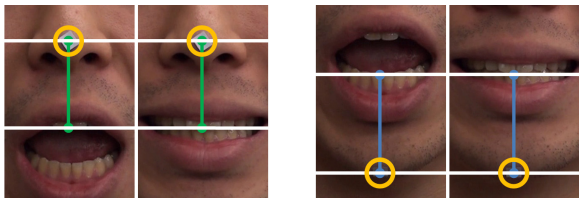

Fig. 4 Verification of the teeth position relative to ANS and chin.

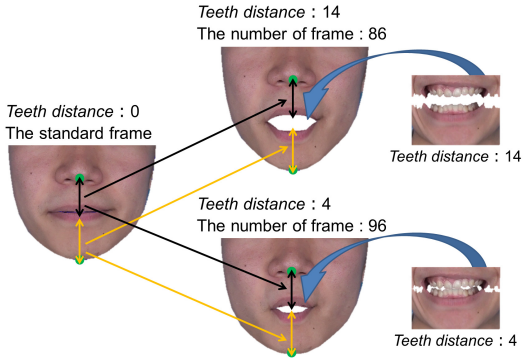

Fig. 5 The selection of teeth image.

structed with the Detai-lization method. Finally, sequence images with a realistic inner mouth appearance are synthesized using the Poisson image editing method [14].

4.1 Upper Teeth and Lower Teeth

The single frontal teeth image is used to create two images, one of the upper teeth and one of the lower teeth. These two images are then used to generate a teeth-database using the following technique. This flexibility allows users to replace teeth images in the teeth-database as needed.

The frontal teeth image is analyzed using a feature point detector [24] to identify 40 feature points. The teeth are isolated using feature points around the mouth and then normalized to fit the animation based on the relative distances between eye-based feature points. The upper and lower teeth images are separated from the whole teeth image using the mean shift method [23]. The combination of feature point detection and mean shift method allows the central positions of the upper and lower teeth images to be obtained. By broadening the distance between central positions of the upper and lower teeth images (*teeth distance*), we can model the opening and closing of the mouth. Teeth distance is defined as the distance between the bottom of the upper teeth and the top of the lower teeth, as measured from the central lateral position. A teeth-database is then generated by synthesizing images using the two images combined at various teeth distances. Based on this concept, a database of 51 teeth images was constructed with teeth distances ranging from 0 to 50.

4.2 Embedding Teeth

In this section, we describe how to embed teeth images into the original animation. The teeth images are obtained from the database constructed in Section 4.1.

The subject's teeth positions in the animation sequence are estimated using knowledge of the human skull bone structure. **Figures 4** and **5** graphically depict the estimation method. The teeth positions are determined assuming that the distance from the position of the Anterior Nasal Spine (ANS) to the central position of the upper teeth is always constant and similarly, that the distance from the position of the chin to the central position of the lower teeth is also constant [25]. Therefore, the ANS and chin feature points are detected using the feature point detector developed by Irie et al. [24]. The teeth distance between the central position of the upper and lower teeth is then calculated using the detected feature points. The teeth images with the closest teeth distance to that of the original animation are then selected from the teeth-database. Since the teeth-database is already normalized to fit the input animation as described in Section 4.1, the system is effectively computing the absolute distance. The best images labeled with teeth distance are selected by the following equation:

$$\arg\min_{i} |d_{I\_f} - d_{D\_i}| \quad (0 \le i \le N) \tag{1}$$

where

$$N = 50 \tag{2}$$

d_{I_f} is the teeth distance of the original animation, d_{D_i} is the teeth distance of the teeth-database, i is an index between 0 and N, and f is the present frame number. According to above equation, the ith teeth image is selected from the teeth-database and this image is embedded into the fth frame of the original animation.

4.3 Embedding Tongue

After embedding the teeth images, the tongue images are embedded into the original animation. The selection of tongue image sets is depicted in **Fig. 6**. Since tongue movement is closely related to phoneme [26], [27], the phoneme combinations found in the sentence text are used to identify the most appropriate tongue

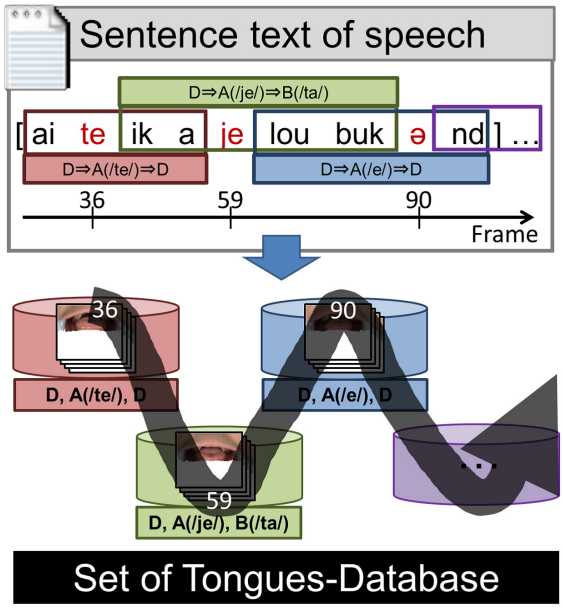

Fig. 6 The selection of tongue image sets.

image sets from the sets of tongues-database. For example, consider the sentence "I take a yellow book and," which can be described phonetically as [ai teik a jelou buk end]. According to the tongue appearance classifications in Table 1 and Table 2, the tongue is visible when /t//e/, /j//e/, and /e/ are pronounced. Using these syllabic sounds (/t//e/, /j//e/, and /e/) as a basis, [ai teik a jelou buk end] can be split into three groups: [ai, te, ik a], [ik a, je, lou buk], and [lou buk, e, nd]. The syllabic sounds are classified as follows:

A(1 or 1+1). "tongue is visible → tongue is visible"
B(1+0). "tongue is visible → tongue is invisible"
C(0+1). "tongue is invisible → tongue is visible"
D(0 or 0+0). "tongue is invisible → tongue is invisible"

Using Table 1 and 2 and the above definitions for A, B, C, and D, the tongue movements for each syllabic sound can be classified as in **Table 3**. A, B, C, and D each represent five, four, four, and five different patterns, respectively. The D classification is typically treated as a single pattern, however, since the tongue is not visible from beginning to end. To discriminate between each of the patterns, a notation such as "A(/te/)" is used to describe the Alveolar + Front vowel (1+1) in refer to Tables 1 and 2. Another example is the use of "A(/θe/)" to describe the Dental + Front vowel pattern (1+1). The same notation is used for the B and C classifications.

Each [• , • , •] described above is a phoneme combination, defined in Section 3.2, that can be described in terms of A, B, C, and D. Phoneme combinations can also be classified as in **Table 4**. We use these phoneme combinations to connect the tongue image sets from the sets of tongues-database to the original animation.

For example, a set of tongue images ([D, A(/te/), D]) are allotted to the original animation images corresponding to the pronunciation of [ai, te, ik a]. Since the number of tongue images from the database is often not identical to the number of animation images corresponding to the pronunciation of a phoneme combination, some of the tongue images are either repeated or removed from the database image sequence. Naturally appearing connections between phoneme combinations are achieved by connecting the sets when the tongue is not visible. Once selected, the tongue images are embedded into the original animation images based on the position of the lower teeth.

Table 3 Classification of the tongue movements for syllabic sounds.

Name	Condition	Examples
Front vowel	A(1)	/e/
Dental + Front vowel	A(1 → 1)	/θ//e/
Alveolar + Front vowel	A(1 → 1)	/t//e/
Palato-alveolar + Front vowel	A(1 → 1)	/r//e/
Palatal + Front vowel	A(1 → 1)	/j//e/
Dental + Back vowel	B(1 → 0)	/θ//a/
Alveolar + Back vowel	B(1 → 0)	/t//a/
Palato-alveolar + Back vowel	B(1 → 0)	/r//a/
Palatal + Back vowel	B(1 → 0)	/j//a/
Bi-labial + Front vowel	C(0 → 1)	/p//e/
Labio-dental + Front vowel	C(0 → 1)	/f//e/
Velar + Front vowel	C(0 → 1)	/k//e/
Glottal + Front vowel	C(0 → 1)	/h//e/
Back vowel	D(0)	/a/
Bi-labial + Back vowel	D(0 → 0)	/p//a/
Labio-dental + Back vowel	D(0 → 0)	/f//a/
Velar + Back vowel	D(0 → 0)	/k//a/
Glottal + Back vowel	D(0 → 0)	/h//a/

Table 4 149 phoneme combinations.

Phoneme combinations	Number
[C, B or D]	$4 \times 5 = 20$
[C, A, B or D]	$4 \times 5 \times 5 = 100$
[D, B]	$1 \times 4 = 4$
[D, A, B or D]	$1 \times 5 \times 5 = 25$

4.4 Making Photorealistic Image

Artificially embedding images into the inner mouth requires some additional steps to produce photorealistic images. In Sections 4.2 and 4.3, we described how to select the teeth and tongue images. Although decoupling the teeth and tongue movement allowed for a substantial reduction in database size, an unnatural boundary between the teeth and tongue image can be created since these images are selected independently. For example, the resulting images will not accurately represent combined effects, such as when teeth affect the tongue's tip. Alternatively, differences between the lighting conditions for the internal and external mouth images can also appear unnatural. The Visio-lization algorithm, proposed by Mohammed et al. [13], and Poisson Image Editing method, proposed by Perez et al. [14], are used to solve these problems. The combination of these two methods is more effective than other smoothing methods because they address issues associated with the fact that the inner mouth is an actual image, while the outer mouth is based on a CG model. **Figure 7** provides an overview of the Visio-lization method.

4.4.1 Visio-lization Method

Images around the mouth are reconstructed using patch-based texture synthesis with the mouth-database. That is, square images around the inner mouth are created by applying Visio-lization with mouth-database as follows.

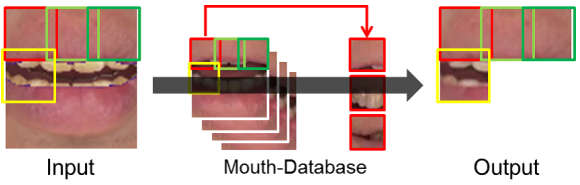

Fig. 7 An overview of the Visio-lization method.

As shown in Fig. 7, the input and database images are separated into multiple small square images called "*patches*." Next, the RGB distance between the patches in the input images and database images are calculated and the best patch image is selected. The best patch image is selected as the image with the smallest RGB distance as defined by the following equation:

$$\arg\min_{i} \sum_{(x,y)\in\Omega_v} \|C_{I\_f\_xy} - C_{D\_i\_xy}\|^2 \quad (0 \le i \le N) \tag{3}$$

where

$$N = 2213 \tag{4}$$

$$C_{I\_f\_xy} = \{R_{I\_f\_xy},\ G_{I\_f\_xy},\ B_{I\_f\_xy}\} \tag{5}$$

$$C_{D\_i\_xy} = \{R_{D\_i\_xy},\ G_{D\_i\_xy},\ B_{D\_i\_xy}\} \tag{6}$$

I indicates an input, D indicates the database, i is an index between 0 and N, f is the present frame number, x is a coordinate value in the horizontal direction of the patch image, y is a coordinate value in the vertical direction of the patch image, Ω_v is the patch image domain, and $R(G, B)_{I\_f\_xy}$ is the R(G,B)-values of the (x, y)position in the fth frame of input sequence. R-values range from 0 to 255, where 0 corresponds to red and 255 indicates white. According to the above equation, the ith patch image is selected from the mouth database and all selected patch images are embedded into each patch position from the top-left to bottom-right. In general, satisfactory results are obtained with the Visio-lization method with a 3 pixel overlap of 20×20 pixel patches.

4.4.2 Detai-lization Method

There are two problems with Visio-lization. First, the method cannot express the full detail of the inner mouth form. Second, since Visio-lization is geared exclusively to still images, time discontinuity can arise between patches. Consequently, the method is not routinely applied to sequential images. To solve these two problems, we propose the use of Detai-lization, an adaptation of Visio-lization that can be used to express details such as each tooth form. With this technique, a very small 6×6 pixel patch size was used with 3 pixel overlap (50% overlap). The use of such a small patch size allows the details of the inner mouth to be represented individually. For this image size, each tooth is 9~12 pixels, therefore, each tooth is synthesized using 2~3 patches. Consequently, we can represent some of the details of each tooth with these smaller patches. The use of small patches also increases the number of patches and patch positions available for analysis. Because such large patches are use in Visio-lization, patch selection is limited to the same position in both input and database images. For example, in Fig. 7, the input patch outlined in red must be replaced by a patch from one of mouth-database images corresponding to the same position (also outlined in red). In the case of Detai-lization, however, any of the small patches from the mouth-database, which has time continuity, may be used. In other words, time continuity between patches can be maintained by selecting patches from the neighborhood, not necessarily the same positions. This flexibility allows us to faithfully reproduce teeth one by one. The approach also accommodates uneven teeth. A comparison of results from Visio-lization and Detai-lization is shown in **Fig. 8**.

Fig. 8 Comparison of Visio-lization's result and Detai-lization's result. Top left row: image before applying Visio-lization or Detail-lization, Top right row: Visio-lization's result, Bottom left row: Detai-lization's result (using 6×6 pixel patch size), Bottom right row: Detai-lization+'s result (using 6×6 pixel patch size and spread reference range).

4.4.3 Seamless Transition

The database images are acquired under lighting conditions other than the original animation. Consequently, the database images must be adjusted to match the lighting conditions of the original image. A seamless transition from the square database patches to the original animation is made with the application of the Poisson Image Editing technique proposed by Perez et al. [14]. This technique interpolates the images captured from the original animation with the square images by solving the Poisson equation, thereby ensuring robustness across different lighting conditions. In summary, a realistic inner mouth animation can be generated with only two databases and three inputs, comprising an original animation, a frontal teeth image, and a syllabic representation of the desired speech.

5. Experiments and Results

To demonstrate this approach, the method was applied to three original animations created by Chang et al., Li et al., and Taylor et al. The sequence images were captured from the same demo movies, we then manually extracted the inner mouth region of each image and applied our method as described in Section 4. **Figure 9** provides a comparison between the new method and prior results. Because our method receive the benefits from real database images, it offers improved resolution over Chang's method and represented the strange teeth from Li's demo more like fangs.

Figure 10 shows a comparison of the new method with Taylor's result. Three different results are shown. The top row corresponds to close-ups of the mouth region from the original sequence images expressed by Taylor et al. The middle and bottom rows demonstrate that the proposed method can be used to embed different teeth images and change luminance values of the tongue images. This capability allows the inner mouth to be replaced with a variety of different features to improve realism and extend flexibility. Therefore, the proposed method can be very useful as a post effect filter to improve speech animation quality.

Table 5 provides the computation time required for each step in the synthesis of an inner mouth animation using an Intel(R) Xeon(R) X4647 processor with 12 GB of RAM. As indicated by the Performance time of Sections 4.2 and 4.3, our system can embed teeth and tongue images in real-time, which allows animators to quickly adjust the inner mouth expressions. The additional processing time associated with Section 4.4 is not problematic since the photorealistic effects can be added as a final step just prior to completion. However, Section 4.4's performance time can be accelerated by effectively selecting patch images using techniques such as FLANN.

6. Evaluation and Discussion

The proposed method was applied to an actual movie to demonstrate the realism of the approach. The experiment was conducted using the steps described in Section 4. A comparison is provided in **Fig. 11** between the original images and synthesized images. Comparing column (a) with (b), we found that the synthesized and real person's appearances are very similar. Some viewers even mistook the synthesized movie for the original one. The performance is attributed to our ability to accurately estimate teeth position and tongue movement. We also used a quantitative evaluation of the Peak Signal-Noise Ratio (PSNR) to evaluate our system. PSNR is calculated with the following equation:

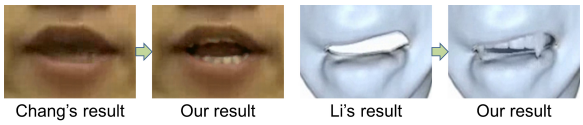

Fig. 9 Comparison with Chang's result and Li's result.

Fig. 10 Comparison with Taylor's result.

Table 5 Performance table of our method.

Section	Performance time [sec/image]
4.1	0.246
4.2	0.0302
4.3	0.0291
4.4	83.9

$$PSNR = 10 \log_{10} \left(\frac{\sum_{i \in \Omega_p} 255^2}{\sum_{i \in \Omega_p} \{y(i) - s(i)\}^2} \right) [\mathrm{d}B] \quad (7)$$

where Ω_p is the set of pixel indices corresponding to the inner mouth region, i is a pixel index, $S(i)$ is the ith luminance value of the synthesized target image, and $y(i)$ is the ith luminance value of real image. The PSNR value was computed for 23 images in which the teeth distance was greater than 10 pixels, ensuring that there was a good view of the inner mouth. For comparison, five different approaches are compared.

(1) Our result
  Teeth ··· created by our method
  Tongue ··· created by our method
(2) Tongue movement opposite to our method
  Teeth ··· created by our method
  Tongue ··· created by letting tongue move reversely
(3) Tongue movement is not accounted for
  Teeth ··· created by our method
  Tongue ··· created by letting tongue stand still
(4) Other result
  Teeth ··· created by using someone else's teeth
  Tongue ··· created by changing the tongue's luminance value
(5) Before applying Detai-lization+
  Teeth ··· created by only embedding teeth
  Tongue ··· created by only embedding tongue

Images from (1)-(5) are shown in **Fig. 12**. Quantitative results are provided in **Table 6**.

Higher PSNR values correspond to better quality images. Since the highest average PSNR values are associated with our proposed method, the inner mouth motion created by our method is more accurate than the alternative methods. The results from approaches (1) through (3) underscore the importance of accurately depicting the tongue movement. In (4), the teeth and tongue are very different from the real image. Since the PSNR is still relatively high, this indicates that it is the movement of the teeth and tongue that are of primary importance. Finally, the results from (5) quantify the effectiveness of the Detai-lization technique. We have demonstrated that the proposed method significantly improves the quality of inner mouth animations for low-quality or empty inner mouth animation. Although previous methods have had difficulty with such tasks, the proposed method is capable of realistically reproducing the appearance of complex inner mouth motions, such as teeth nipping the tip or back of the tongue. The ability to generate realistic inner mouth animations is primarily attributed to the use of Detail-lization. Detai-lization

Fig. 11 Comparison with actual images. The top column (a) represents a synthesis result and close-ups of the mouth region. The bottom column (b) represents the original images and close-ups of the mouth region.

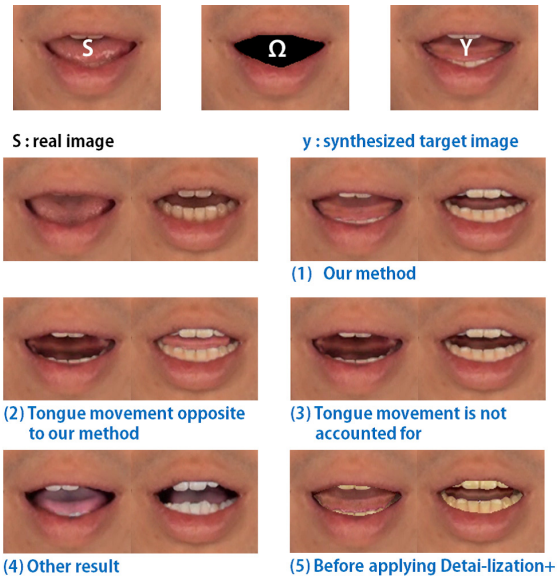

Fig. 12 Examples of the real and synthesized target image (S: real, Y: Synthesized target, Ω: inner mouth region).

Table 6 Results of the quantitative evaluation.

Approach	Average PSNR of each images set [dB]
(1)	17.04
(2)	15.12
(3)	15.30
(4)	15.19
(5)	15.75

is a new method proposed here that is applicable to areas with uneven details and sequential images. We have also demonstrated that our method can be applied to a variety of speech animations.

In future work, we will consider continuing imaged-based 2D animation with alternative camera angles, or possibly expanding the method to 3D animation. The proposed method does not apply directly to natural speech animation. As for realizing 2D animation with alternative camera angles, we require construction of the database with alternative camera angles. However, this will require substantial additional time and effort to construct all possible angles for the teeth and tongue-database. Therefore, methods will be explored to obtain images from multiple angles using as few images as possible. Moreover, we extracted the inner mouth region manually. Thereby, our result sometimes suffers from flicking artifacts because of low extraction. We would like to extract the inner mouth region automatically and accurately. In addition, we will consider lighting vectors. Image-based animation does not have normal vectors. Therefore, our system cannot represent shadow changes of the inner mouth when the light source position changes.

References

[1] Alexander, O., Roger, M., Lambeth, W., Chiang, J.-Y., Ma, W.-C., Wang, C.-C. and Debevec, P.: The Digital Emily Project: Achieving a Photorealistic Digital Actor, *IEEE Trans. CGA*, pp.20–31 (2010).
[2] Joshi, P., Tien, W.C., Desbrun, M. and Pighin, F.: Learning Contrals for Blend Shape Based Realistic Facial Animation, *Proc. SCA '03*, pp.187–192 (2003).
[3] Tena, J.R., Torre, F.D. and Matthews, I.: Interactive Region-Based Linear 3D Face Models, *Proc. SIGGRAPH '11*, No.76 (2011).
[4] Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R.W. and Gross, M.: High-Quality Passive Facial Performance Capture using Anchor Frames, *Proc. SIGGRAPH '11*, No.75 (2011).
[5] Seol, Y., Lewis, J.P., Seo, J., Choi, B., Anjyo, K. and Noh, J.: Spacetime Expression Cloning for Blendshapes, *TOG '12 Journals*, Vol.32 (Issue 2), No.14 (2012).
[6] Deng, Z. and Neumann, U.: Expressive Speech Animation Synthesis with Phoneme-Level Controls, *Computer Graphics Forum*, Vol.27, No.8, pp.2096–2113 (2008).
[7] Perez, P., Gangnet, M. and Blake, A.: Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces, *IEEE Trans. Visualization and Computer Graphics*, Vol.12, No.6, pp.1523–1534 (2006).
[8] Bregler, C., Covell, M. and Slaney, M.: Video Rewrite: Driving visual speech with audio, *Proc. SIGGRAPH '97*, pp.353–360 (1997).
[9] Buck, I., Finkelstein, A., Jacobs, C., Klein, A., Salesin, H.A., Seims, J., Szeliski, R. and Toyama, K.: Performance-Driven Hand-Drawn Animation, *Proc. 1st International Symposium on Non Photorealistic Animation and Rendering*, pp.101–108 (2000).
[10] Cosatto, E. and Graf, H.: Photo-realistic Talking-heads from Image Samples, *IEEE Trans. Multimedia*, Vol.2, No.3, pp.152–163 (2000).
[11] Ezzat, T., Geiger, G. and POGGIO, T.: Trainable Videorealistic Speech Animation, *Proc. SIGGRAPH '02*, pp.388–398 (2002).
[12] Chang, Y. and Ezzat, T.: Transferable Videorealistic Speech Animation, *Proc. SCA '05*, pp.143–151 (2005).
[13] Mohammed, U., Prince, S.J.D. and Kautz, J.: Visio-lization: generating novel facial images, *Proc. SIGGRAPH '09*, No.57 (2009).
[14] Perez, P., Gangnet, M. and Blake, A.: Poisson Image Editing, *Proc. SIGGRAPH '03*, pp.313–318 (2003).
[15] Taylor, S.L., Mahler, M., Theobald, B.-J. and Matthews, I.: Dynamic Units of Visual Speech, *Proc. SCA '12*, pp.275–284 (2012).
[16] Li, H., Weise, T. and Pauly, M.: Example-Based Facial Rigging, *Proc. SIGGRAPH '10*, No.32 (2010).
[17] Huang, H., Chai, J., Tong, X. and Wu, H.-T.: Leveraging motion capture and 3D scanning for high-fidelity facial performance acquisition, *Proc. SIGGRAPH '11*, No.74 (2011).
[18] Bickel, B., Botsch, M., Angst, R., Matusik, W., Otaduy, M., Pfiste, H. and Gross, M.: Multi-scale capture of facial geometry and motion, *Proc. SIGGRAPH '07*, No.33 (2007).
[19] King, A.S. and Parent, E.R.: A 3D Parametric Tongue Model for Animated Speech, *The Journal of Visualization and Computer Animation*, Vol.12, No.3, pp.107–115 (2001).
[20] Vogt, F., Lloyd, E.J., Buchaillard, S., Perrier, P., Chabanas, M., Payan, Y. and Fels, S.S.: Efficient 3D Finite Element Modeling of a Muscle-Activated Tongue, *The 3rd International Conference on Biomedical Simulation*, pp.19–28 (2006).
[21] Yang, Y., Guo, X., Vick, J., Torres, G.L. and Champbell, T.: Physics-Based Deformable Tongue Visualization, *IEEE Trans. Visualization and Computer Graphics*, Vol.19, No.5, pp.811–823 (2013).
[22] Pelachaud, C., Overveld, C.W.A.M.V. and Seah, C.: Modeling and Animating the Human Tongue during Speech Production, *Computer Animation '94*, pp.40–49 (1994).
[23] Comaniciu, D. and Meer, P.: Mean shift: A robust approach toward feature space analysis, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.24, No.5, pp.603–619 (2002).
[24] Irie, A., Takagiwa, M., Moriyama, K. and Yamashita, T.: Improvements to Facial Contour Detection by Hierarchical Fitting and Regression, *The 1st Asian Conference on Pattern Recognition*, pp.273–277 (2011).
[25] Blanz, V., Basso, C., Poddio, T. and Vetter, T.: Reanimating faces in images and video, *Eurographics '03*, Vol.22 (2003).
[26] Gibbon, F.E., Lee, A. and Yuen, I.: Tongue-Palate Contact During Selected Vowels in Normal Speech, *The Cleft Palate-Craniofacial Journal*, Vol.47, No.4, pp.405–412 (2010).
[27] Stone, M., Morrish, K.A., Sonies, B.C. and Shawker, T.H.: Tongue curvature: A model of shape during vowel production, *Folia Phoniat*, Vol.39, pp.302–315 (1987).

Editor's Recommendation

This paper proposes a novel method for lip-sync animation. The method generates intraoral motion pictures by using a minimum necessary set of separate motion image databases for teeth and tongue, and synthesizing them on the fly. Their results are closer to real mouth motion images than those from previous research. In addition, their method can easily be embedded into existing lip-sync systems without inner mouth images.

(Chairman of SIGCG Masanori Kakimoto)

Masahide Kawai was born in 1990. He received his B.S. degree in Engineering from Waseda University, Tokyo, Japan, in 2013. He is currently attending master's course at the Graduate School of Physics and Applied Physics, Waseda University. His research interest is Computer Graphics. He is a student member of the IPSJ and ACM SIGGRAPH.

Tomoyori Iwao was born in 1989. He received his B.S. degree in Engineering from Waseda University, Tokyo, Japan, in 2012. He is currently attending master's course at the Graduate School of Physics and Applied Physics, Waseda University. His research interest is Computer Graphics.

Daisuke Mima was born in 1988. He received his B.S. and M.S. degrees in engineering from Waseda University, Tokyo, Japan, in 2011, 2013. His research interest is Computer Graphics.

Akinobu Maeiima was born in 1978. He received his B.S. and M.S. degrees, all in Electrical Engineering and Electronics from Seikei University in 2002, 2004, and Ph.D. degree in Science and Engineering from Waseda University in 2010. He was a research associate of Waseda University from 2007 to 2010. Currently, He is a junior researcher of Information Technology Research Organization, Waseda University. His research interest includes Computer Graphics and Computer Vision. He is a member of IEIECE, IIEEJ and ACM SIGGRAPH.

Shigeo Morishima was born in 1959. He received his B.S., M.S. and Ph.D. degrees, all in Electrical Engineering from the University of Tokyo, Tokyo, Japan, in1982, 1984, and 1987, respectively. Currently, he is a professor of School of Advenced Science and Engineering, Waseda University, Tokyo, Japan. His research interests include 3D Reconstruction and Modeling of Face, Motion Analysis and Synthesis of Human Body, Analysis and Synthesis of Facial Expression and Retargetting, and all concerning about Future Interactive Entertainment using speech and image processing. He was a visiting professor at University of Toronto from 1994 to 1995. He is also a project leader of Seculity and Safety Laboratories, Information Technology Research Origanization. He received the IEJCE-J Achievement Award in May, 1992.