Electronic Preprint for Journal of Information Processing Vol.23 No.5

Regular Paper

Automatic Generation of Photorealistic 3D Inner Mouth Animation only from Frontal Images

Masahide Kawai[1,a)] Tomoyori Iwao[1,b)] Akinobu Maejima[1,c)] Shigeo Morishima[2,d)]

Received: November 6, 2014, Accepted: June 5, 2015

Abstract: In this paper, we propose a novel method to generate highly photorealistic three-dimensional (3D) inner mouth animation that is well-fitted to an original ready-made speech animation using only frontal captured images and small-size databases. The algorithms are composed of quasi-3D model reconstruction and motion control of teeth and the tongue, and final compositing of photorealistic speech animation synthesis tailored to the original. In general, producing a satisfactory photorealistic appearance of the inner mouth that is synchronized with mouth movement is a very complicated and time-consuming task. This is because the tongue and mouth are too flexible and delicate to be modeled with the large number of meshes required. Therefore, in some cases, this process is omitted or replaced with a very simple generic model. Our proposed method, on the other hand, can automatically generate 3D inner mouth appearances by improving photorealism with only three inputs: an original tailor-made lip-sync animation, a single image of the speaker's teeth, and a syllabic decomposition of the desired speech. The key idea of our proposed method is to combine 3D reconstruction and simulation with two-dimensional (2D) image processing using only the above three inputs, as well as a tongue database and mouth database. The satisfactory performance of our proposed method is illustrated by the significant improvement in picture quality of several tailor-made animations to a degree nearly equivalent to that of camera-captured videos.

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

1. Introduction

In movie and video game productions, synthesizing a character's facial expressions, subtle mouth movements, and appearances is essential to creating photorealistic character animations [1]. In speech animation generation, modeling and rendering the inner mouth region and performing lip synchronization are critical to realizing photorealistic quality in terms of the mouth openings, tongue movements, and face rotations synchronized to speech utterance. Therefore, these speech animations must be manually created by skilled artists because of the highly complex appearance changes to be achieved in and around the mouth, particularly for photorealistic human characters. However, creating them requires considerable effort and time.

To address the above issue, researchers have proposed speech animation synthesis techniques. Representative techniques include three-dimensional (3D) model-based methods and two-dimensional (2D) image-based methods. The 3D model-based methods include blendshapes with several shape models [12], [22] and expression retargeting with a motion-capture system [3], [19]. The 2D image-based methods synthesize mouth animations using a prepared video corpus [5], [8]. Both methods can create speech animations with realistic lip movements. However, the detailed appearance of the resulting inner mouth is inadequate because of the complex appearance changes that must be accommodated. In other words, despite the generation of high-quality lip movements, traditional methods still cannot generate realistic inner mouth appearances. If the quality of the inner mouth appearance in ready-made animations is improved through post-processing, the quality of the animations may be significantly improved as an after-effect with no additional labor and cost.

To address the need for realistic inner mouth animation, we propose a method to automatically restore existing speech animations by reconstruction, motion synthesis, and composition of a photorealistic 3D inner mouth in a ready-made animation. The proposed 3D reconstruction method of an inner mouth involves using an approximate approach based on actual measured values and anatomical knowledge. Realistic inner mouth shapes can be represented by implementing the following two reconstruction methods. In the first method, teeth are reconstructed with a quasi-3D model with thickness only by a single frontal image. In the second, the tongue is reconstructed by a simple shape with a Gaussian function and displacement mapping [20], which generates the concavo-convex feature of the tongue in the graphics processing unit. In addition, our 3D simulation method of inner mouth movement is based on physical assumptions and phonetic analysis. For example, the position of the teeth is estimated by considering skull bone structures, and tongue movement is determined by association with symbolic sounds.

Although 3D inner mouth shapes can be reconstructed and sim-

[1] Waseda University, Shinjuku, Tokyo 169–8555, Japan
[2] Waseda Research Institute for Science and Engineering, Shinjuku, Tokyo 169–8555, Japan
[a)] doara-waseda@toki.waseda.jp
[b)] sazabi@akane.waseda.jp
[c)] akinobu@mlab.phys.waseda.ac.jp
[d)] shigeo@waseda.jp

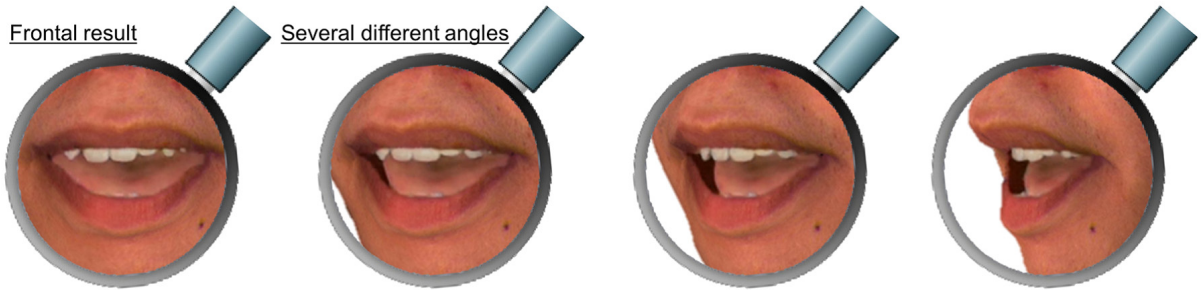

Fig. 1 Detailed view of synthesized speech animation examples. The proposed method can generate 3D inner mouth animation from frontal images. Specifically, it can represent detailed inner mouth shapes, such as the concavo-convex shape, by shape deformation and photorealistic inner mouth appearances using patch-based image synthesis.

ulated, mismatches in image and motion between the internal and external mouth images must be inevitably considered to generate a perfect photorealistic face in movie and video game productions. The mismatches manifested in two primary forms: a sharp boundary between the inner and outer mouth, and a difference in luminance. These mismatches are herein addressed through the combined use of *Multi-view Detai-lization* (advanced *visio-lization* [16]) and seamless transitions [18]. Multi-view Detai-lization in this context is a method that generates novel images that can be applied to detailed areas, such as uneven teeth, and to sequential images, such as animations. This approach is an improvement over visio-lization, which is only applicable to still images, not detailed areas or sequential images. Further, Multi-view Detai-lization can generate 3D inner mouth images using only frontal images. In this paper, we demonstrate that a photorealistic inner mouth animation can be generated by combining the benefits of our Multi-view Detai-lization and the seamless transition method. The main contributions of this paper can be summarized as follows:

(1) **3D reconstruction and simulation of the inner mouth:**
1.1 Reconstruction: Teeth thickness is represented with a multi-layered ellipse cylinder model and the tongue's pixel-wise concavo-convex shape is represented using displacement mapping, depth control by luminance, and frontal images.
1.2 Simulation: Teeth and tongue simulation is based on physical assumptions and phonetic analysis.

(2) **Synthesis of the photorealistic inner mouth:** Photorealistic inner mouth images are created using only frontal images. The proposed Multi-view Detai-lization and seamless transition method are efficient in making the luminance discontinuity naturally smooth and seamless.

As outlined above, our system can create a photorealistic 3D inner mouth animation using only frontal images, while the resulting animation maintains the high-quality lip appearances of ready-made speech animation. As shown in **Fig. 1**, our method can produce photorealistic inner mouth images from not only the frontal viewpoint but also any other viewpoints because our inner mouth model is represented by quasi-3D information. In this paper, our method is focusing on spoken English, although our method can be used for a variety of languages.

2. Related Work

Numerous speech animation techniques have been proposed by many researchers. These techniques are based on different approaches, such as the blendshape method, data-driven method, and blendshape and data-driven combination method. Further, there are two types of speech animations: 2D and 3D. In approaches for generating 2D animations [6], the system generates speech animations that transfer the speaking style of one person to another using a multidimensional morphable model (MMM). When using MMM, however, the inner mouth morphs along with the lip movements; therefore, inner mouth appearances in the resulting animations are expanded and contracted. Accordingly, Anderson et al. could remove inner mouth expansion and contraction by replacing the inner mouth region with a static inner mouth image [2]. Complex tongue movements, however, could not be represented because their method uses a static image. Kawai et al. proposed video-realistic inner mouth animation using inner mouth image databases [13]. Although the method is effective for frontal video animations, 3D inner mouth animation cannot be represented.

In addition, 3D methods have likewise been proposed [15], [21]. Taylor et al. proposed a data-driven method for lip synchronization that achieves realistic lip movements by connecting sequences of active appearance model (AAM) parameters based on phonetic information. Li et al. developed a system to generate facial blendshape rigs with a small number of training poses. The quality of inner mouth shapes of their works entirely depended on the nature of target shapes made by skilled artists. Moreover, the inner mouth animations had to be sequentially created. To create complex inner mouth appearances by hand is an even more difficult task. Some researchers have created facial animations using motion-capturing systems, which are often used to create speech animations [9]. However, the resultant inner mouths were blank in all of the facial animations because the systems were unable to capture inner mouth data. In summary, because inner mouth appearances are blurred or blank, state-of-the-art speech animation techniques have not yet sufficiently achieved the creation of realistic 3D inner mouth animation.

To produce 3D inner mouth animations, some researchers have developed tongue simulation using a tongue model [14], [23]. King et al. controlled a B-spline surface with 60 control points of the tongue model. Unfortunately, the method does not accurately

represent tongue movements. In addition, accurate tongue simulation methods have been proposed [23]. This method achieves accurate simulation of the tongue using the 3D finite element method (FEM). However, this method is unable to produce photorealistic tongue appearances and is limited by the computational cost associated with tongue simulation. As shown by the above related works, the need for creating realistic inner mouth animation remains. Our proposed system addresses this need by creating photorealistic 3D inner mouth animation as a post-production effect with both 2D and 3D approaches which need only a low calculation cost and a small size database. In this paper, we describe this animation restoration method, which improves the quality of facial animations by adding reconstructed and simulated inner mouth animation in the 3D space.

3. Data Acquisition

In this section, we describe data acquisition for the three inputs: original 3D speech animation, teeth image, and syllabic decomposition of speech. We additionally describe the tongue and mouth databases.

3.1 Input Data

As indicated above, three inputs are required for the application of our method. An original 3D speech animation in English is first required. Our method is intended to serve as an upgrade for the original animation by improving the appearance of the inner mouth. Realistic inner mouth animation is thereby provided, while the attributes of the original animation are maintained, such as realistic lip movements synchronized with speech, detailed wrinkles, and so on. In addition, the proposed method can support 3D translation and angular rotation of the face. An image of the speaker's frontal teeth is then required (*teeth image*). With this single image, animations can be generated that depict the opening or closing of the speaker's teeth during speech. The image size is automatically adjusted to fit the animation size. Finally, the syllabic content of the subject's speech must be converted to text. For example, when a subject utters a /te/ syllable in the 79th frame, then "te: 79" is stored in a text file.

3.2 Tongue Set Database

Sets of consecutive tongue images that represent an arbitrary subject pronouncing *phoneme combinations* were acquired for the database [13]. The use of phoneme combinations preserves original continuous tongue movements as much as possible. In this paper, phoneme combinations are defined according to the visibility of the tongue as a group of three phonemes. This group starts and ends with the mouth closing (tongue is invisible) and mouth opening (tongue is visible) in the middle because the tongue animation is smoothly performed and connected. These combinations contain all tongue movement variations that appear in each utterance of vowels and consonants in spoken English [17]. The tongue appearance classifications for vowels and consonants are respectively shown in **Tables 1** and **2**. When the tongue is visible, the class is 1; otherwise, the class is 0. We combined vowels with consonants to determine phoneme combinations, such as /i/-/t//e/-/i/, /f/-/e/-/b/, and so forth. A total of 149

Table 1 Classification of the tongue appearances for vowels.

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

Table 2 Classification of tongue appearances for consonants.

Name	Tongue appearance	Class	Ex.
Bi-labial	Between upper and lower lips	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 tip	1	/r/
Palatal	Between hard palate and front of lingual surface	1	/j/
Velar	Between soft palate and front of lingual surface	0	/k/
Glottal	Between vocal cords	0	/h/

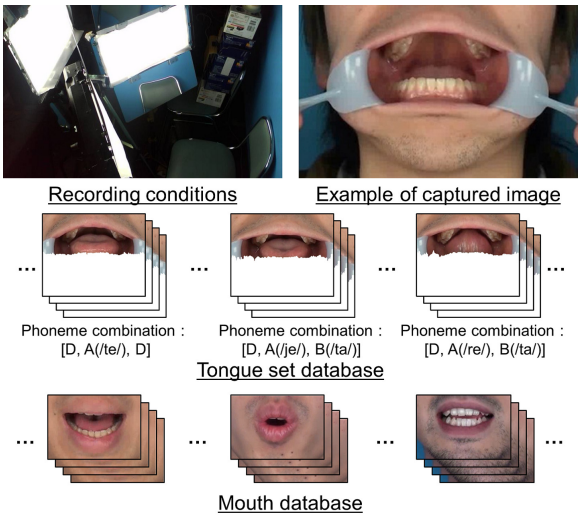

Fig. 2 Recording conditions and an image captured while using an Angle Wider, as well as representative images from a database.

tongue movement variations exist for spoken English. We captured videos of tongue movements 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 discussed in Section 5.2. To capture the image sets for the tongue database, we used the *Angle Wider* tool made by Ziecor International, Inc., which is shown in **Fig. 2**. The frontal tongue images were captured under uniform lighting conditions without shade. Although the Angle Wider made it difficult for the subject to naturally speak, the inner mouth motions were typically correct, particularly when the subject used references (tongue appearances), such as in Tables 1 and 2. The boundary between the lower teeth and tongue image was obtained using the

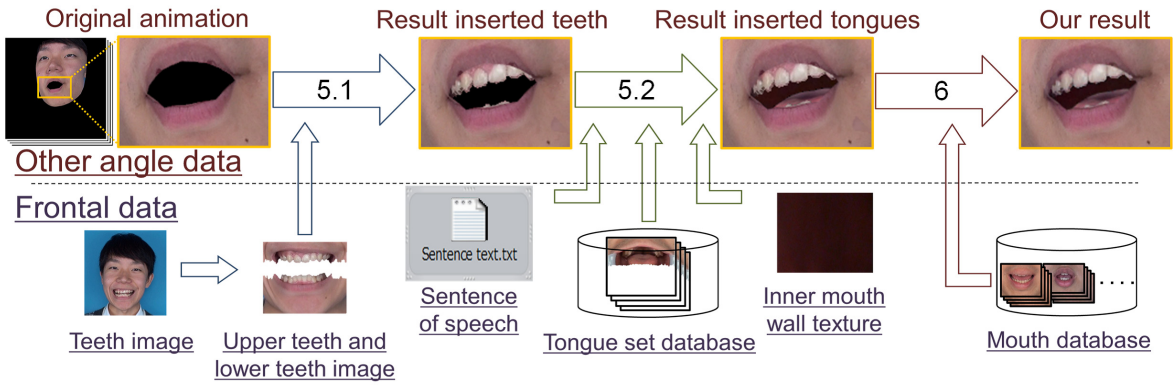

Fig. 3 Workflow of 3D inner mouth restoration method.

mean shift method [7]. Tongue images were then separated from the captured images, as shown in Fig. 2.

3.3 Mouth Database

We used 2,213 images from the captured video to apply the Multi-view Detai-lization method (see Section 6). Therefore, it is desirable to capture database based on the differences of mouth appearances. In general, we required mouth database in regard to (1): mouth openings such as the close vowel (/i/, /u/), half open vowel (/e/, /o/), and open vowel (/a/), (2): tongue positions for vowels as shown in Table 1, and (3): tongue positions for consonants as shown in Table 2. Therefore, we captured videos of seven subjects (except for the speaker in the input animation) pronouncing representative vowels (/aiueo/) for satisfying the above (1) and (2), and representative symbolic sounds (/te/, /re/, /je/, /θe/, /fa/, /va/) produced by different articulation regions for satisfying the above (3). The database images were 241 × 201 pixels and included movement of the overall mouth area from the upper lip to the lower lip. Constructed database images are depicted in Fig. 2.

4. Overview

A workflow of our 3D inner mouth restoration is shown in **Fig. 3**. The method is comprised of three primary steps. First, the 3D shape of the teeth is reconstructed based on elliptic cylinder approximation (Section 5.1.2) and then reconstructed 3D teeth is embedded into the input animation (Section 5.1.3). Second, the 3D tongue shapes are reconstructed (Section 5.2.3) and then embedded into the resulting animation from the first step so that the movement of reconstructed tongue is synchronized with the desired syllabic content (Sections 5.2.1, 5.2.2, and 5.2.4). In the above two steps, there are noticeable artifacts due to color differences on the boundary between the embedded teeth and tongue in the resulting animation. To remove unnatural boundaries and improve their texture, as a third step, inner mouth image synthesis based on a novel Multi-view Detai-lization method is employed (Section 6). Supplemental video shows the outline of the proposed method;

`http://youtu.be/gDLr1LS31JI`

5. Realistic 3D Inner Mouth Restoration

This restoration process is composed of two steps: 3D teeth restoration and 3D tongue restoration. The teeth positions and tongue movements are decided by human anatomy and phonemes.

5.1 Teeth Reconstruction and Simulation

Teeth shape and color are essential for representing an impression of personal characteristics in the inner mouth and for making facial animation photorealistic. A tooth, for example, can be one or more of a variety of shapes, such as a double tooth, misaligned tooth, or canine tooth. In addition to shape, the colors of cavities and dirty teeth can be effective in representing an identity. However, in animation production, a generic model of teeth is used because it is time-consuming to make an individualized 3D teeth model that considers personal characteristics. We therefore propose a simple teeth modeling method based on a single captured image that can represent the appearance of personal characteristics. Accordingly, by replacing only the captured teeth image, a new 3D teeth appearance can be automatically generated that reflects the personal characteristics of teeth shape and color.

5.1.1 Lower and Upper Teeth Extraction

As shown in Section 3.1, a single captured teeth image is analyzed to create the upper and lower teeth models. After 40 feature points are detected using the method by Irie et al. [10], upper and lower teeth images are separated using the mean shift method [7]. From that point, the center tips of the upper and lower teeth are decided (see **Fig. 4** (c)). Moreover, teeth images are size-normalized to fit the animation based on the relative distances between the feature points of the right and left eyes. By controlling the distance between the tips of the upper and lower teeth (*teeth distance*), mouth opening and closing can be generated.

5.1.2 Teeth Reconstruction

3D reconstruction of teeth is performed by approximately aligning an elliptic cylinder to the teeth image. X, Y, and Z axes of the coordinate are defined, as shown in Fig. 4 (a); X, Y, and Z correspond to the left-right, up-down, and front-back (depth) directions, respectively, from the frontal view. We adopt an ellipse whose curve expresses an alignment of the general computer graphics (CG) teeth model in the X-Z plane, and we determine the major and minor axes of the ellipse. Based on the visual mea-

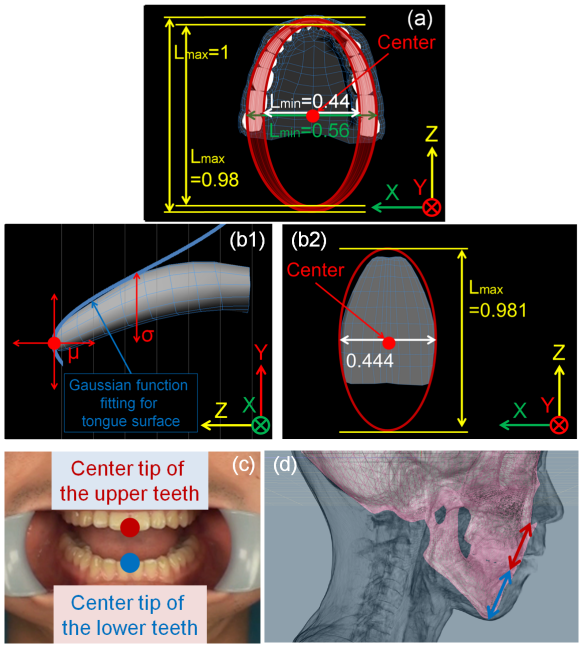

Fig. 4 General CG model: (a) teeth, and (b1) tongue (Z-Y), and (b2) tongue (X-Z). (c) Center position of teeth, (d) Verification of the teeth position relative to ANS and the chin.

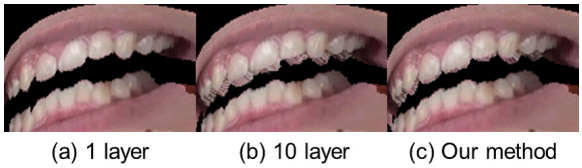

Fig. 5 Comparison layered teeth model. (a) is only single layer, (b) is constant 10 layer model and (c) is our variable depth 10 layered teeth model.

surement, we confirm that the teeth model in the X-Z plane can be approximated by the ellipse whose minor axis is 0.566 when the major axis of the ellipse is 1.

After texture mapping of the upper and lower teeth images to the cylinder surface to represent the thickness of teeth, the shape of the elliptic cylinder surface is slightly deformed. The deformed multiple cylinder surfaces are then accumulated by considering the general CG teeth model (Fig. 4 (a)). In this paper, the number of layers is empirically set to ten to represent teeth thickness. Each length of the major and minor axes in the ten elliptic cylinders is gradually changed to consider the difference of each tooth thickness, while central positions of the cylinders are maintained as the same: thin for anterior teeth, and thick for posterior teeth. Based on actual teeth thickness — from 1.2 [mm] at the anterior to 7.0 [mm] at the posterior — we heuristically adjusted the ratio of the axes from $L_{minor}/L_{major} = 0.556/1$ (1st layer) to $L_{minor}/L_{major} = 0.444/0.981$ (10th layer) in order to represent the difference of teeth thickness when L_{minor} and L_{major} are the length of minor and major axis respectively. **Figure 5** shows the result of a texture-mapped 3D teeth model: (a) reflects only one layer, which appears paper-thin, (b) is a ten-layer accumulation with a constant interval, and (c) is a ten-layer accumulation with our thickness adjustment according to the tooth position.

5.1.3 Teeth Simulation

Our teeth model is controlled in synchronization with the original lip movement. Teeth positions are estimated by the facial model's vertex positions around the mouth relative to human skull bone structure, as shown in Fig. 4 (d). Specifically, we apply the assumption that the distance from the anterior nasal spine (ANS) position to the center tip of the upper teeth is always constant, and the distance from the chin position to the center tip of the lower teeth is likewise constant [4]. Therefore, the ANS and chin positions are obtained from the vertex coordinates of the 3D facial model ANS and chin. The distance between the center tips of the upper and lower teeth is then calculated using the obtained vertex coordinates. We can simulate the opening and closing of the teeth models by controlling the multi-cylindrical teeth models in the 3D space. As a result, the individual characteristics of 3D teeth shape and motion can be reconstructed using only a single captured image of teeth.

5.2 Tongue Reconstruction and Simulation

Although a tongue is an important element to represent an impression of inner mouth as well as teeth, it cannot always be observed because it is located behind teeth and usually shaded. Therefore we propose a simple tongue model that can generate an adequate photorealistic tongue appearance only from actual tongue images without precise 3D modeling of tongue feature.

5.2.1 Tongue Appearance Model

For tongue movement generation, the most appropriate tongue image is selected from a tongue image set database that relates to a phoneme combination [13] in the input sentence text. For example, consider the phrase, "I take a yellow book and $\cdots$," which can be described phonetically as [ái téik a jélou búk énd]. According to the tongue appearance classifications in Tables 1 and 2, the tongue is only 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"

Referencing Tables 1, 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 represent five, four, four, and five different patterns, respectively. However, the D classification is typically treated as a single pattern because the tongue is not visible from beginning to end. To distinctively express each of the patterns, a notation such as "A(/te/)" is used to describe the Alveolar + Front vowel pattern (1+1), which is referenced in 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 classes.

Further, these classifications shown in Tables 1, 2, and 3 are only considered as the "vowel" or "consonant + vowel," not considered as the "consonant + consonant." In case of "consonant + consonant," all combinations of the "consonant + consonant" are exceptionally classified as "tongue is invisible" even if the consonant is classified as 1 (tongue is visible) in Table 2. The rea-

Table 3 Classification of 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/

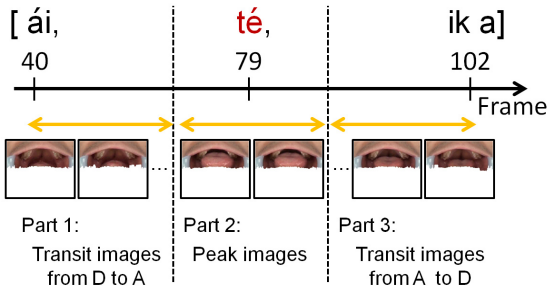

Fig. 6 Selection of tongue image sets.

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]	$1 \times 5 \times 5 = 25$

son is because the "consonant + consonant" cannot have a strong accent and then it is pronounced fast. Therefore, we regard the "consonant + consonant" as a consonant for an invisible tongue. For example, consider speaking a word "strength," which can be described phonetically as [stréntθ]. According to the above classification, the tongue is invisible when /st/ and /nθ/ are pronounced. Therefore, [D, A(/re/), D] is allotted to strength.

5.2.2 Tongue Image Selection

Tongue images are selected from the database according to the sentence text and then aligned with speech, as shown in **Fig. 6**. Each phoneme combination can be classified as in **Table 4**. We use these phoneme combinations to connect the tongue image sets from the tongue set database to the original animation. For example, a set of tongue images ([D, A(/te/), D]) is allotted to the original animation by corresponding to the pronunciation of [ai, te, ik a]. Additionally, a set of tongue images is comprised of three parts: transit images from D to A(/te/), peak images of A(/te/), and transit images from A(/te/) to D. To synchronize the tongue image sequence with the original animation, time scale adjustment is required by extending or contracting the corresponding image sequence in the database. Because all phoneme combinations shown in Table 4 always begin with C or D and end with B or D, any two combinations are inevitably connected through the

Fig. 7 Comparison of results from applying Gaussian approximation and the displacement mapping method. Left: elliptic cylinder approximation; middle: elliptic cylinder approximation + Gaussian approximation; right: elliptic cylinder approximation + Gaussian approximation + displacement mapping method.

term of "tongue is invisible" and therefore output tongue animation is generated without discontinuity.

5.2.3 Tongue Reconstruction

Tongue shapes' differences between people are not perceived normally, because the tongue shape is often occluded and shadowed by the lip or teeth shapes. In this section, therefore, tongue shapes are designed heuristically and coarsely.

For tongue reconstruction, the X, Y, and Z coordinate axes are defined, as shown in Fig. 4 (b1, b2); X, Y, and Z respectively correspond to the left-right, up-down, and front-back (depth) directions from the frontal view. The tongue shape in the X-Z plane is approximately expressed by an ellipse that is completely similar to that of 10th layer's teeth described in Section 5.1.2. Moreover, the Gaussian function is adopted to the general CG tongue model in the Z-Y plane, as shown in Fig. 4 (b2). The tip of tongue model is located in the center position indicated by (μ) in the y-direction (height) of Gaussian function, and the variance of the function (σ) is set to 0.314 when the distance between the right eye and left eye is normalized to 1. The shape of tongue is approximated as a symmetric shape both in x and y direction. Next, we create the elliptic cylinder shape whose bottom face is located in the X-Z plane. We deform the shape in the Z-Y plane along the Gaussian function in the 3D space and perform texture mapping of the tongue image.

We can create an approximate tongue in 3D although the features of its surface cannot be as detailed as a concavo-convex object. Therefore, we apply the displacement mapping method [20] to the approximated tongue shape in which the greater the luminance value of a certain pixel, the smaller the coordinate value of the depth (z). In sum, the dark pixels on the tongue surface are located in the back of the mouth (−z direction), while the light pixels are located in the front (+z direction). In this way, detailed tongue shapes can be created frame by frame from only the selected frontal tongue images as shown in Section 5.2.2 (Tongue Image Selection). A comparison of results from applying Gaussian approximation and the displacement mapping method is shown in **Fig. 7**.

5.2.4 Tongue Simulation

Finally, the generated tongue model is controlled in sync with the original facial animation. In this paragraph, we simulate tongue movement by estimating the tongue position of the z-direction and y-direction frame by frame with the tongue's luminance value. Estimating the tongue positions are depicted in **Fig. 8**. Because a tongue located in the deep position of the inner mouth appears dark, and a tongue located near the teeth appears

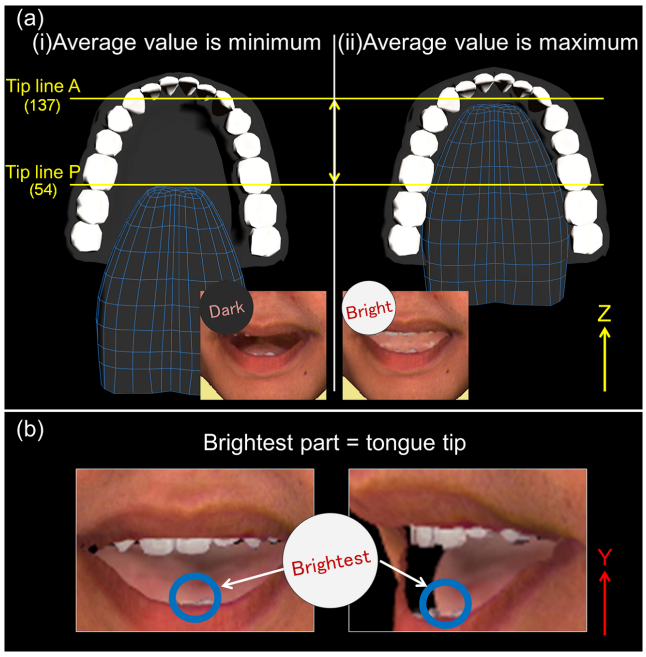

Fig. 8 (a) Estimating the tongue position in z-direction, (b) Estimating the tongue position in y-direction.

bright, the location of the tongue tip in the z-direction is decided by the average luminance values of the total tongue pixels. The tip of the tongue position in the 3D space is now represented by (x', μ, z'). The tongue position in the z-direction (z': tip line P) is located in the z-value of the posterior teeth if the average of all tongue luminance values is the minimum as shown in Fig. 8 (a(i)). On the other hand, the tongue position in the z-direction (z': tip line A) is located in the z-value of the anterior teeth if the average of all tongue luminance values is the maximum as shown in Fig. 8 (a(ii)). In the example case with our tongue database, the minimum value (tip line P) is 54 and the maximum value (tip line A) is 137 in grayscale (black is 0, white is 255).

Next, the tip of the tongue position is controlled in the y-direction (μ). This direction is up and down according to the location of the maximum luminance value (brightest part) of the tongue pixels, which is estimated as a position of the tongue tip as shown in Fig. 8 (b). This rule is applicable to almost all cases; it is not followed only when the sentence includes "re," in which the tip of the tongue appearance is occluded by the upper teeth or lip. In this case, we specifically control the tip of the tongue position toward the alveolar arch position. We have defined a special representation of a curled tongue as the syllabic sound "re." As shown above, only by controlling the tip of the tongue position (x', μ, z'), we can simulate tongue movements in the 3D space. In addition, an inner mouth is composed of teeth, the tongue, and the inner mouth wall. We therefore captured a photo (wall texture) of the inner mouth wall region. Using this wall texture, we synthesize the inner mouth region.

6. Inner Mouth Image Synthesis

Artificially appending images in the inner mouth requires additional steps to produce photorealistic images. In Section 5, we described how to reconstruct the teeth and tongue shapes in the 3D space from frontal images and how to simulate them. Although decoupling the appearance of the teeth, tongue, and inner mouth wall texture enables a substantial reduction in database size, an unnatural boundary between the teeth and tongue texture, or between the tongue and inner mouth wall texture, can result because these textures are independently rendered. Moreover, the resulting texture of the teeth shape will make stripes between layers. Alternatively, differences in lighting conditions for the internal and external mouth textures can also appear unnatural. The Multi-view Detai-lization algorithm (advanced visio-lization [16]) and Poisson image editing method [18] are used to solve these problems. The combination of these two methods is more effective than other smoothing methods because the two methods address issues associated with the fact that the inner mouth is an actual image, while the outer mouth is based on a computer graphics model.

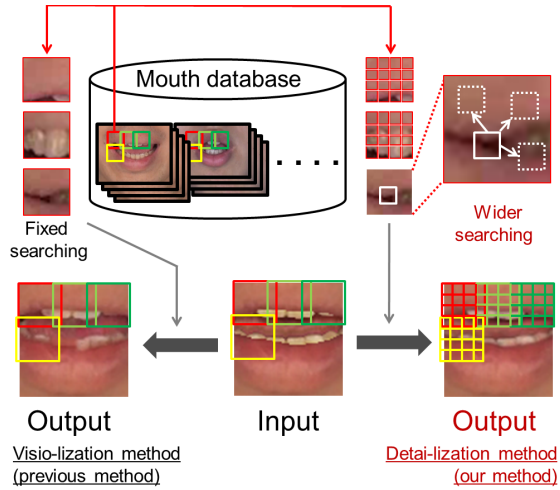

Fig. 9 Overview of visio-lization and Multi-view Detai-lization methods.

We export an image sequence rendered with the teeth and tongue models in Section 5 as an input and then apply image processing to the image sequence using an actual database. In this way, improved photorealistic 3D inner mouth animation can be created. **Figure 9** provides an overview of the visio-lization and Multi-view Detai-lization method.

6.1 Multi-view Detai-lization Method

The novel images can be created by using visio-lization method with the mouth database as follows. As shown in Fig. 9, 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 as the image with the smallest RGB distance. All selected patch images are embedded into each patch position from the top left to bottom right. Satisfactory results are typically obtained using the visio-lization method with a 3-pixel overlap of 20×20 pixel patches. Unfortunately, there is a problem in visio-lization. The method creates image failure in the detail region of the inner mouth form, such as each tooth representation.

To solve the problem, we propose a novel Multi-view Detai-lization method, an adaptation of visio-lization that can be used to express details, such as each tooth form and the delicate bound-

ary between the teeth and lip. Our Multi-view Detai-lization has two special features in which the smaller patch size is used to represent the very delicate feature of teeth. In addition, a wider searching area of patch around the original position is introduced to minimize matching errors and maintain time continuity compared to the conventional Visio-lization method when adapting to face rotation. The very small 6×6 pixel patch size is used with a 3-pixel overlap (50% overlap). For this image size, each tooth is 9~12 pixels; therefore, each tooth is synthesized using 2~3 patches. Alternatively, we can realize the inner mouth appearance of other angles using the mouth database, which is comprised only of frontal images. The use of small patches additionally increases the number of patches and patch positions available for 3D analysis. The best patch is selected as the image with the smallest RGB distance as newly defined by the following equation:

$$\begin{aligned} &\underset{i,m,n}{\operatorname{argmin}} \sum_{(x_m,y_n)\in\Omega_{m,n}} \|\boldsymbol{C}_I(f, x_{m_p}, y_{n_p}) - \boldsymbol{C}_D(i, x_m, y_n)\|^2 \\ &(0 \le i < N = 2{,}213),\ (m_p - 2 \le m \le m_p + 2), \\ &(n_p - 2 \le n \le n_p + 2) \end{aligned} \tag{1}$$

where

$$\begin{aligned} &\boldsymbol{C}_I(f, x_{m_p}, y_{n_p}) \\ &= \{R_I(f, x_{m_p}, y_{n_p}),\ G_I(f, x_{m_p}, y_{n_p}),\ B_I(f, x_{m_p}, y_{n_p})\} \end{aligned} \tag{2}$$

$$\begin{aligned} &\boldsymbol{C}_D(i, x_m, y_n) \\ &= \{R_D(i, x_m, y_n),\ G_D(i, x_m, y_n),\ B_D(i, x_m, y_n)\} \end{aligned} \tag{3}$$

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, m is the column number of the patch, n is the row number of the patch, m_p is the column number of the present referred patch in input, n_p is the row number of the present referred patch in input, $\Omega_{m,n}$ is the patch image domain of the m-th column and n-th row, and $R_I(f, x_{m_p}, y_{n_p})$ ($G_I(f, x_{m_p}, y_{n_p})$, $B_I(f, x_{m_p}, y_{n_p})$) denotes the R (G, B) values of the (x, y) position of $\Omega_{m,n}$ in the f-th frame of the input sequence. According to the above equation, the m-th column and n-th row patch image of the i-th image is selected from the mouth database. All selected patch images are embedded into each patch position from the top-left to bottom-right with a 3-pixel overlap. For the patch syntheses, we take an interpolation of overlap regions to not create artifacts of overlap regions between neighboring patches. Therefore, we use a linear interpolation. Using this interpolation, we can smoothly represent an image.

6.2 Seamless Transition

Because our database images are acquired in a studio under lighting conditions different from those of the original animation, a seamless transition between selected patches and the original animation is introduced by the Poisson image editing technique [18], which can normalize lighting conditions by solving the Poisson equation. A comparison of results from a combination of the seamless transition and visio-lization, and that of the seamless transition and Multi-view Detai-lization, is shown in **Fig. 10**. By using Multi-view Detai-lization, we can create a natural inner mouth appearance, whereas visio-lization results in some image failure parts. Moreover, our method is effective for eliminating the unnatural boundary between the teeth and tongue, and between the teeth and lip, compared to the input images. In summary, a photorealistic 3D inner mouth animation can be generated with only two frontal databases and three inputs: an original animation, a frontal teeth image, and a syllabic representation of the desired speech.

Fig. 10 Comparison of the respective visio-lization and Multi-view Detai-lization results. Top row: image before applying Multi-view Detai-lization; middle row: combination of seamless transition and the visio-lization result; bottom row: combination of seamless transition and the Multi-view Detai-lization result using a 6×6 pixel patch size and spread reference range. Multi-view Detai-lization can create a photorealistic inner mouth appearance without unnatural boundaries and image failure regions.

7. Result

We created 3D inner mouth animations using our method as described in Sections 5 and 6. **Figure 11** provides the image sequence of the resulting animation. The top row presents examples of synthesized speech animation of the human model; the bottom row provides detailed views of examples. Our method can be applied to a variety of models that are mammalian. The human model results confirm that our method can represent realistic tongue movements for speaking the syllabic sound "re." Further, the more resulting animations are included in our movie;

`http://youtu.be/gDLr1LS31JI`

8. Evaluation

To demonstrate the photorealism of the proposed approach, a subjective evaluation and objective evaluation were conducted. The results of these experiments demonstrate that our method is effective.

8.1 Subjective Evaluation

We conduct a subjective evaluation of the proposed method. Sample sequential images extracted from the evaluation videos are shown in **Fig. 12**. The five videos were captured by a video camera. The proposed method was applied to some of them after manually extracting the inner mouth region of certain videos. Accordingly, we created some synthesized videos (our result). Twenty-six subjects watched the five types (some are synthesized videos and the others are photographed videos) of evaluation videos after we jumbled the synthesized and photographed videos. The subjects were then asked to judge using

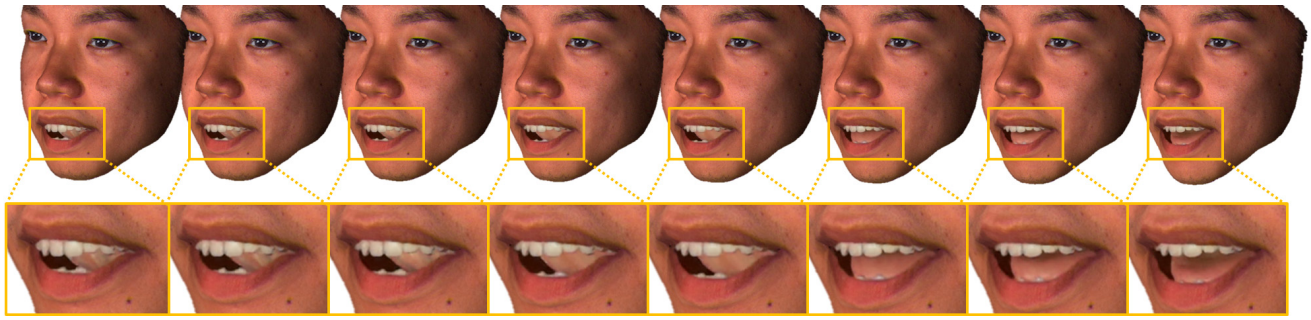

Fig. 11 Sequential images of the resulting animation. The results confirm that our method can represent realistic tongue movements for speaking the syllabic sound "re."

Fig. 12 Sample sequential images from the evaluation videos. Images in the first (top), second, and bottom rows were created by our method.

Table 5 Results of the subjective evaluation. The total number of subjects was 26.

Sample number	Answer: Synthesized	Answer: Photographed	Percentage that answered correctly
1st	12 people	14 people	46.2%
2nd	21 people	5 people	80.8%
3rd	0 people	26 people	100%
4th	0 people	26 people	100%
5th	19 people	7 people	73.1%
Total of 1st, 2nd and 5th (synthesized)	52 people	26 people	66.7%
Total of 3rd and 4th (real)	0 people	52 people	100%

two-alternative questions whether the inner mouth of the character in the given video was a synthesized inner mouth by our method or a photographed inner mouth captured by video camera. To evaluate the quality of inner mouth appearances, we advised all subjects to focus only on the inner mouth region of the character while watching the videos. The subjects' assessments are shown in **Table 5**. The first, second, and fifth rows of the table were created by our method; for those rows, just one-third of the subjects incorrectly answered "photographed." In addition, we determined that this evaluation was properly validated because the answer for watching the videos with photographed mouths (third and fourth sample) was 100%. That is to say, our mouth-synthesized videos compared to videos with photographed mouths; therefore, our synthesized videos can be regarded as having represented high-quality, photorealistic inner mouth appearances. This performance is attributed to our method's ability to accurately estimate teeth position and tongue movement.

Fig. 13 Examples for objective evaluation (S: real, Y: synthesized target, Ω_p: inner mouth region).

8.2 Objective Evaluation

We conducted an objective evaluation to evaluate the validity of three aspects of our method: 1) estimating teeth position, 2) analyzing tongue movements, and 3) the Multi-view Detai-lization technique. We applied our method to a real person's facial image for which the inner mouth region was manually extracted. The face was then laterally inclined at an angle of 33 degrees from the frontal view. We used the peak signal-to-noise ratio (PSNR) for the 78 face images as our evaluation criterion. PSNR was calculated with the following equation:

$$PSNR = 10 \log_{10} \left(\frac{\sum_{i \in \Omega_p} 255^2}{\sum_{i \in \Omega_p} \{y(i) - s(i)\}^2} \right) \text{[dB]} \quad (4)$$

where Ω_p is the set of pixel indices corresponding to the inner mouth region, i is a pixel index, $s(i)$ is the i-th luminance value of the synthesized target image, and $y(i)$ is the i-th luminance value of real image. Example images for the objective evaluation are shown in **Fig. 13**. The PSNR was calculated for the 78 open mouth images in which teeth and tongue were not occluded. For comparison, four different approaches were compared.

(1) **Our result:**
Teeth · · · created by our method,
Tongue · · · created by our method

(2) **Other result with someone else's teeth:**
Teeth · · · created by using someone else's teeth,
Tongue · · · created by our method

(3) **Tongue movement was not accounted for Ref. [2]:**
Teeth · · · created by our method,
Tongue · · · created by letting tongue remain still

Table 6 Results of the quantitative evaluation.

Approach	Average PSNR of each images set [dB]
(1) Our method	15.65
(2)	14.46
(3) Anderson et al.'s result [2]	13.39
(4)	15.19

(4) **Before applying Multi-view Detai-lization:**
Teeth ··· created by only embedding teeth,
Tongue ··· created by only embedding tongue

The evaluation results are provided in **Table 6**. Higher PSNR values correspond to higher quality images. The highest average PSNR values are associated with our proposed method; therefore, the inner mouth motion created by our method is more accurate than the alternative methods. The results from approaches (1) and (2) show that the inner mouth impression differs based on the teeth used. Further, (1) and (3) show the importance of accurately depicting the tongue movement. Finally, the results from (4) quantify the effectiveness of the Multi-view Detai-lization technique. As shown above, our method has demonstrated validity for estimating teeth position, analyzing tongue movements, and performing the Multi-view Detai-lization technique.

9. Limitations and Future Work

Our system can be used for a variety of mammalian computer graphics characters. Because the system must use skull bone structure to estimate teeth position, it is not used for non-mammalian characters, which is a limitation of the system. Additionally, our method requires the actual measurement of the inner mouth; consequently, it cannot be applied to characters for which actual inner mouth measurements cannot be conducted. Our system can generate very photorealistic inner mouth animations, however when the lighting condition is changing, now the specular and delicate shade change cannot be represented.

In future work, shadowing of the inner mouth will be implemented to represent light rays that are occluded by the inner mouth surroundings and those refracted in the inner mouth. We plan to introduce shadowing methods, such as ray tracing and photon mapping, for the inner mouth. Moreover, we must consider light rays transmitted through liquids, such as saliva inside the mouth; rendering of wet materials [11] will solve this problem. Further, we estimated tongue positions by luminance values of the tongues as shown in Section 5.2.4. Although the proposed method can estimate tongue positions easily, it is possible to estimate tongue positions incorrectly. To estimate more accurately, we should simulate tongue movement based on the relationship between tongue appearances and sounds that derive from pronounced sentences. Additionally, teeth thickness should be further considered. In this paper, our method can express differences of each tooth according to the measured teeth thickness. However, it cannot express detail in a piece of tooth; the degree of tooth thickness becomes gradually smaller from the base to the tip. We will consider representing the tooth-piece detail by adjusting not only teeth thickness but also each tooth thickness. Needless to say, the degree of tooth thickness is a topic with room for exploration.

10. Conclusion

In this paper, we proposed a novel restoration method for existing speech animations. Our restoration technique consists of a few key approaches. The 3D reconstruction provides an approximated 3D inner mouth shape that is not accurate but is adequate in achieving a naturalistic representation. The reconstructed 3D teeth are simulated under anatomical constraints. In addition, the reconstructed 3D tongue is simulated based on the result of phonetic analysis. We can represent satisfactory 3D inner mouth simulations from frontal images by representation of teeth thickness and tongue's concavo-convex feature while considering human anatomy and phonetic analysis. Our synthesis offers a detailed image at any viewpoint from only frontal images using Multi-view Detai-lization which can also correct an unnatural boundary between the lip and inner mouth. Our proposed Multi-view Detai-lization method is much more effective for smoothing discontinuities in both texture and time-sequential domains compared to conventional visualization. The ability to generate realistic inner mouth animations is primarily attributed to the use of Multi-view Detai-lization. Our proposed method significantly improves the quality of existing inner mouth animations while maintaining original lip movement. Further, it can contribute towards considerably improving efficiency in movie and video game productions.

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] Anderson, R., Stenger, B., Wan, V. and Cipolla, R.: Expressive Visual Text-To-Speech Using Active Appearance Models, *CVPR*, pp.3382–3389 (2013).

[3] 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*, Vol.30, No.4, pp.75:1–75:10 (2011).

[4] Blanz, V., Basso, C., Poddio, T. and Vetter, T.: Reanimating faces in images and video, *Proc. Eurographics '03*, Vol.22, pp.641–650 (2003).

[5] Bregler, C., Covell, M. and Slaney, M.: Video rewrite: Driving visual speech with audio, *Proc. SIGGRAPH '97*, pp.353–360 (1997).

[6] Chang, Y. and Ezzat, T.: Transferable videorealistic speech animation, *SCA*, pp.143–151 (2005).

[7] Comaniciu, D. and Meer, P.: Mean shift: A robust approach toward feature space analysis, *IEEE Trans. PAMI*, Vol.24, No.5, pp.603–619 (May 2002).

[8] Cosatto, E. and Graf, H.: Photo-realistic talking-heads from image samples, *IEEE Trans. Multimedia*, Vol.2, No.3, pp.152–163 (Sep. 2000).

[9] 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*, Vol.30, No.4, pp.74:1–74:10 (2011).

[10] Irie, A., Takagiwa, M., Moriyama, K. and Yamashita T.: Improvements to facial contour detection by hierarchical fitting and regression, *ACPR*, pp.273–277 (2011).

[11] Jensen, H.W., Legakis, J. and Dorsey, J.: Rendering of Wet Materials, *Proc. Eurographics Workshop*, pp.273–281 (2011).

[12] Joshi, P., Tien, W.C., Desbrun, M. and Pighin, F.: Learning controls for blend shape based realistic facial animation, *SCA*, pp.187–192 (2003).

[13] Kawai, M., Iwao, T., Maejima, A. and Morishima, S.: Video-Realistic Inner Mouth Reanimation, *Pacific Graphics*, pp.35–40 (2013).

[14] King, A.S. and Parent, E.R.: A 3D Parametric Tongue Model for Animated Speech, *The Journal of VCA*, No.12, pp.107–115 (2001).

[15] Li, H., Weise, T. and Pauly, M.: Example-based facial rigging, *Proc. SIGGRAPH '10*, Vol.29, No.4, pp.32:1–32:6 (2010).

[16] Mohammed, U., Prince, S.J.D. and Kautz, J.: Visio-lization: Gen-

erating novel facial images, *Proc. SIGGRAPH '09*, Vol.28, No.3, pp.57:1–57:8 (2009).
[17] Pelachaud, C., Overveld, C.V. and Seah, C.: Modeling and animating the human tongue during speech production, *Proc. Computer Animation '94*, pp.40–49 (1994)
[18] Perez, P., Gangnet, M. and Blake, A.: Poisson image editing, *Proc. SIGGRAPH '03*, pp.313–318 (2003).
[19] Seol, Y., Lewis, J.P., Seo, J., Choi, B., Anjyo, K. and Noh, J.: Spacetime expression cloning for blendshapes, *TOG*, Vol.31, No.2, No.14 (Apr. 2012).
[20] Szirmay-Kalos, L. and Umenhoffer, T.: Displacement Mapping on the GPU – State of the Art, *Comput. Graph. Forum.*, pp.1567–1592 (2008).
[21] Taylor, S.L., Mahler, M., Theobald, B.-J. and Matthews, I.: Dynamic units of visual speech, *SCA*, pp.275–284 (2012).
[22] Tena, J.R., Torre, F.D. and Matthews, I.: Interactive region-based linear 3D face models, *Proc. SIGGRAPH '11*, Vol.30, No.4, pp.76:1–76:9 (2011).
[23] Yang, Y., Guo, X., Vick, J., Torres, G.L. and Champbell, T.: Physics-Based Deformable Tongue Visualization, *IEEE Trans. VCG*, Vol.19, pp.811–823 (2013).

Masahide Kawai was born in 1990. He received his B.S. and M.S. degrees in Engineering from Waseda University, Tokyo, Japan, in 2013 and 2015. He is currently working at Sony Corporation from 2015. His research interests are Computer Graphics and Computer Vision. He received the IPSJ GCAD Award in June, 2014 and the Master Thesis Award of Waseda University in March, 2015. He is a member of IPSJ and ACM SIGGRAPH.

Tomoyori Iwao was born in 1989. He received his B.S. and M.S. degrees in Engineering from Waseda University, Tokyo, Japan, in 2012, 2014. He is currently working at Canon Corporation from 2014. His research interest is Computer Graphics.

Akinobu Maejima was born in 1978. He received his B.S. and M.S. degrees, all in Electrical Engineering and Electronics from Seikei University in 2002 and 2004. The Ph.D. degree in Science and Engineering from Waseda University in 2010. He was a research associate of Waseda University from 2007 to 2010. He was a junior researcher of Information Technology Research Organization, Waseda University from 2010 to 2014. Currently, he is working at OLM Digital, Inc. from 2014. His research interests are Computer Graphics and Computer Vision. He is a member of 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, in 1982, 1984, and 1987, respectively. Currently, he is a professor of School of Advanced 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 Security and Safety Laboratories, Information Technology Research Organization. He received the IEICE-J Achievement Award in May, 1992.