Pacific Graphics (2013)
B. Levy, X. Tong, and K. Yin (Editors)

Video-Realistic Inner Mouth Reanimation

M. Kawai[1] and T. Iwao[1] and A. Maejima[1] and S. Morishima[1]

[1]Waseda University, Japan

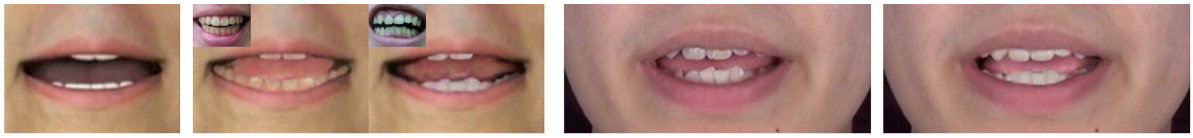

Figure 1: *Example shot of speech animation. Left: Comparison with the result from "Dynamic Units of Visual Speech" [Taylor et al. 2012]. Realistic inner mouth synthesis achieved using small database. Right: Comparison with actual image. Photorealistic inner mouth expression is achieved, closely matching the real person's image and validating the proposed approach.*

Abstract

We propose a novel post-effect method that can make an existing speech animation video-realistic by generating an inner mouth appearance that is tailored to the speaker. The automatic generation of photorealistic inner mouth appearances requires only simple inputs and small databases and is not restricted by video lighting. The approach is also applicable for creature speech animation with human voices. Our system uses two key algorithms; one of the algorithms models the inner mouth appearances based on physical assumptions, and the other algorithm synthesizes the inner mouth images with care taken to maintain time continuity and luminance gradients.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics And Realism—Animation

1. Introduction

Creating realistic and convincing speech animations is still one of the important topics in movie and video game productions[ARL*10]. 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 [TTM11] or retarget real expressions with a motion capturing system [DN08]. 2D method synthesizes mouth motions using a prepared video corpus [BCS97]. 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, realistic inner mouth animations are not yet realized with traditional methods.

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 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 [MPK09]) and seamless transitions [PGB03]. Detai-lization is a method that generates novel images that can be applied to detailed areas like uneven teeth, and sequential images like animations. In this paper, we demonstrate that a photo-

DOI: 10.2312/PE.PG.PG2013short.035-040

realistic 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.

2. Related Work

As previously mentioned, there are several methods for creating speech animations. Chang et al. [CE05] developed a system that can generate speech animations that transfers the speaking style 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. [TMTM12] 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. Motion capture systems are often used to create facial animations [DNL*06]. 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.] used the tongue model, which is composed of a B-spline surface with 60 control points [KP01]. 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) [YGV*13]. 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. Preparation

3.1. Input Data

Our system needs three inputs, an original animation, a frontal image of the teeth, and a syllabic decomposition of the speech, 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 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.

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 1: *Classification of the tongue appearances for vowels*

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; 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. We classified the tongue appearance when each subject uttered vowels and consonants in spoken English [POS94]. 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 Table 1 and Table 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. 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.). A total of 149 tongue movement variations exist for spoken English. The classification of all 149 tongue movements (phoneme combinations) is motivated 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. 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, and tongue images are then separated from the captured images.

3.3. Mouth-Database

We captured videos of seven subjects pronouncing consonants and representative symbolic sounds produced by

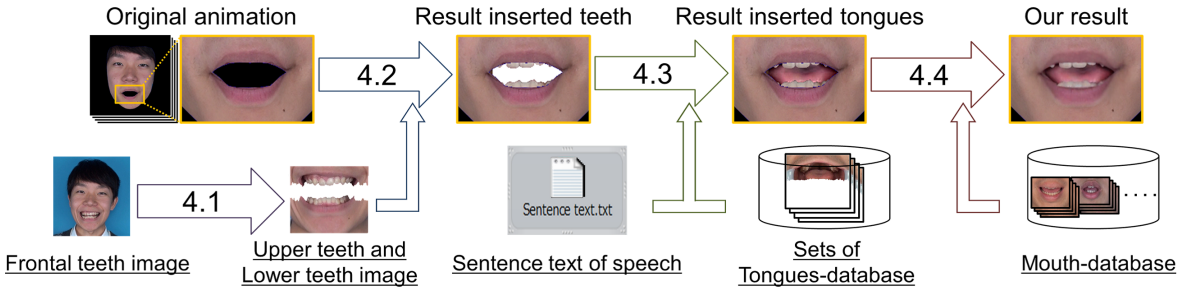

Figure 2: *An overview of our method*

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/

Table 2: *Classification of the tongue appearances for consonants*

different articulation regions. We used 2213 images from the captured video to apply the Detai-lization method. The database images were 241×201 pixels. Images included movement of the overall mouth from the upper and lower lips.

4. Creating a Post-Effect Filter

An overview of our method is shown in Figure 2. This section describes the inner mouth reanimation technique. The method is composed of four steps.

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 [ITMY11] 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. 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

The subject's teeth positions in the animation sequence are estimated using knowledge of the human skull bone structure. The teeth positions are determined using an assumption 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 [BBPV03]. Therefore, the ANS and chin feature points are detected using the feature point detector developed by Irie et al. [ITMY11]. 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:

$$\arg\min_i |d_{I\_f} - d_{D\_i}| \ (0 \leq i \leq N) \quad (1)$$

where

$$N = 50 \quad (2)$$

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

Since tongue movement is closely related to phoneme [GLY10], the phoneme combinations found in the sentence text are used to identify the most appropriate tongue 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. A, B, C, and D each represents 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 [C, B or D], [C, A, B or D], [D, B], and [D, A, B or D]. The number is 4×5+4×5×5+1×+1×5×5=149. 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]. Once selected, the tongue images are embedded into the original animation images based on the position of the lower teeth.

4.4. Making Photorealistic Image

Artificially embedding images into the inner mouth requires some additional steps to produce photorealistic 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. 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. [MPK09], and Poisson Image Editing method, proposed by Perez et al. [PGB03] 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 3 provides an overview of the Visio-lization method.

Images around the mouth are reconstructed using patch-based texture synthesis with the database. That is, square images around the inner mouth are created by applying Visio-lization with mouth-database as follows. As shown in Figure 3 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:

$$\arg\min_i \sum_{(x,y)\in\Omega} \|\boldsymbol{C}_{I\_f\_xy} - \boldsymbol{C}_{D\_i\_xy}\|^2 \ (0 \leq i \leq N) \quad (3)$$

where

$$N = 2213 \quad (4)$$

$$\boldsymbol{C}_{I\_f\_xy} = \{R_{I\_f\_xy},\ G_{I\_f\_xy},\ B_{I\_f\_xy}\} \quad (5)$$

$$\boldsymbol{C}_{D\_i\_xy} = \{R_{D\_i\_xy},\ G_{D\_i\_xy},\ B_{D\_i\_xy}\} \quad (6)$$

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 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 3 pixel overlap of 20×20 pixel patches.

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. To solve these two problems, we propose the use of Detai-lization, an adaptation of Visio-lization that can be used to

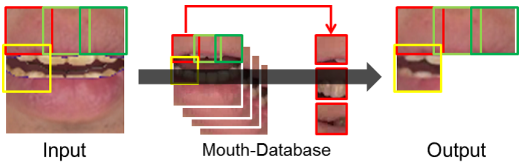

Figure 3: *An overview of the Visio-lization method*

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. 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.

The database images are acquired in a lighting condition 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. [PGB03]. 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.

5. Experiment and Result

To demonstrate this approach, the method was applied to original animation created by Taylor et al. [TMTM12]. The sequence images were captured from the demo movie, we then manually extracted the inner mouth region of each image and applied our method as described in Section 4. Figure 4 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 3 provides the

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

Table 3: *performance table of our method*

Figure 4: *Comparison with Taylor's result*

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 in 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. As for section 4.4, performance time will be able to be accelerated by effectively selecting patch images using techniques such as FLANN (Fast Library for Approximate Nearest Neighbors).

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 Figure 5 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 as 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:

$$PSNR = 10\log_{10}\left(\frac{\sum_{i\in\Omega} 255^2}{\sum_{i\in\Omega}\{y(i)-s(i)\}^2}\right) \text{ [dB]} \qquad (7)$$

where Ω 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. Quantitative results are provided in below.

Figure 5: *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.*

(1) Our result → 17.40[dB]
Teeth ··· created by our method
Tongue ··· created by our method
(2) Tongue movement opposite to our method → 15.12[dB]
Teeth ··· created by our method
Tongue ··· created by letting tongue move reversely
(3) Tongue movement is not accounted for → 15.30[dB]
Teeth ··· created by our method
Tongue ··· created by letting tongue stand still
(4) Other result → 15.19[dB]
Teeth ··· created by using someone else's teeth
Tongue ··· created by changing the tongue's luminance value
(5) Before applying Detai-lization+ → 15.75[dB]
Teeth ··· created by only embedding teeth
Tongue ··· created by only embedding tongue

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.

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 motion, 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 is a new method proposed here that is applicable to areas with uneven details and sequential images.

In future work, we have to consider quasi 3D animation from any arbitrary camera angles and lighting condition. Our method can be easily extended by preparing multi-angle database theoretically. The method of high speed retrieval and compression of database is problem to be overcome. Moreover, we have to consider lighting vector. Image-based animation does not have normal vector. Therefore, our system cannot represent shadow change of inner mouth when light source position changes.

References

[ARL*10] ALEXANDER, O., ROGER, M., LAMBETH, W., CHIANG, J-Y., MA, W-C., WANG, C-C., DEBEVEC, P.: The Digital Emily Project: Achieving a Photorealistic Digital Actor. In *IEEE CGA* (2010), pp. 20–31. doi:10.1109/MCG.2010.65.

[BBPV03] BLANZ V., BASSO C., PODDIO T., VETTER T.: Reanimating faces in images and video. In *Proc. Eurographics '03* (2003), vol. 22, pp. 641–650. doi:10.1111/1467-8659.t01-1-00712.

[BCS97] BREGLER C., COVELL M., SLANEY M.: Video rewrite: Driving visual speech with audio. In *Proc. SIGGRAPH '97* (1997), pp. 353–360. doi:10.1145/258734.258880.

[CE05] CHANG Y., EZZAT T.: Transferable videorealistic speech animation. In *Proc. the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation* (2005), pp. 143–151. doi:10.1145/1073368.1073388.

[DN08] DENG, Z., NEUMANN: Expressive Speech Animation Synthesis with Phoneme-Level Controls. *Computer Graphics Forum 27* 8 (Sep 2008), pp. 2096–2113. doi:10.1111/j.1467-8659.2008.01192.x.

[DNL*06] DENG, Z., NEUMANN, U., LEWIS, J. P., KIM, T-Y., BULUT, M., NARAYANAN, S: Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces. *IEEE Trans. on Visualization and Computer Graphics 12* 6 (Nov/Dec 2006), pp. 1523–1534. doi:10.1109/TVCG.2006.90.

[GLY10] GIBBON F E., LEE A., YUEN I.: Tongue-palate contact during selected vowels in normal speech. *The Cleft Palate-Craniofacial Journal 47* 4 (July 2010), pp. 405–412. doi:10.1597/09-067.1..

[ITMY11] IRIE A., TAKAGIWA M., MORIYAMA K., YAMASHITA T.: Improvements to facial contour detection by hierarchical fitting and regression. In *The First Asian Conference on Pattern Recognition* (2011), pp. 273–277. doi:10.1109/ACPR.2011.6166689.

[KP01] KING, A. S., PARENT, E. R.: A 3D Parametric Tongue Model for Animated Speech. In *The Journal of Visualization and Computer Animation 12* 3 (Sep 2001), pp. 107–115. doi:10.1002/vis.249.

[MPK09] MOHAMMED U., PRINCE S J. D., KAUTZ J.: Visiolization: generating novel facial images. In *Proc. SIGGRAPH '09* (2009), No. 57. doi:10.1145/1576246.1531363.

[PGB03] PEREZ P., GANGNET M., BLAKE A.: Poisson image editing. In *Proc. SIGGRAPH '03* (2003), pp. 313–318. doi:10.1145/1201775.882269.

[POS94] PELACHAUD C., OVERVELD C V., SEAH C.: Modeling and animating the human tongue during speech production. In *Proc. Computer animation '94* (1994), pp. 40–49. doi:10.1109/CA.1994.324008.

[TMTM12] TAYLOR S L., MAHLER M., THEOBALD B-J., MATTHEWS, I.: Dynamic units of visual speech. In *Proc. the 2012 ACM SIGGRAPH/Eurographics symposium on Computer animation* (2012), pp. 275–284. doi:10.1145/1073368.1073388.

[TTM11] TENA J R., TORRE F D., MATTHEWS I.: Interactive region-based linear 3D face models. In *Proc. SIGGRAPH '11* (2011), No. 76. doi:10.1145/2010324.1964971.

[YGV*13] YANG, Y., GUO, X., VICK, J., TORRES, G. L., CHAMPBELL, T.: Physics-Based Deformable Tongue Visualization. In *IEEE Trans. on Visualization and Computer Graphics 19* 5 (May 2013), pp. 811–823. doi:10.1109/TVCG.2012.174.