Photorealistic Inner Mouth Expression in Speech Animation

Masahide KAWAI* Tomoyori IWAO Daisuke MIMA Akinobu MAEJIMA Shigeo MORISHIMA†
Waseda University

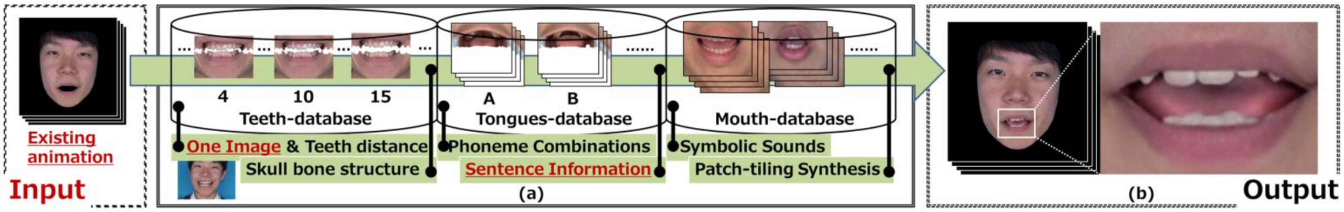

Figure 1: *Overview of our method and result.*

1. Introduction

We often see close-ups of CG characters' faces in movies or video games. In such situations, the quality of a character's face (mainly in dialogue scenes) primarily determines that of the entire movie. Creating highly realistic speech animation is essential because viewers watch these scenes carefully. In general, such speech animations are created manually by skilled artists. However, creating them requires a considerable effort and time.

To solve this problem, we propose a method to automatically synthesize a speech animation by embedding a realistic inner mouth into an existing "low-quality" or "inner mouth-less" speech animation. To the best of our knowledge, there is little comparable work. Chang et al. [2005] proposed the system that is somewhat similar to ours. However, the appearance of inner mouth in the resulting animation is collapsed because the inner mouth morphs along with lip movement. To resolve this morphing artifact, we estimate teeth positions from human anatomy and estimate tongue movements from phonetics. Then, we synthesize more natural looking inner mouth images.

Our contribution is to provide a post effect filter that improves the quality of speech animation created by any previous technique, as shown in the supplementary video.

2. Database Construction

We construct database for each of the regions using the following procedures. First, we capture a video of an arbitrary subject gradually opening his/her mouth. Then, the teeth are cut from the images acquired from the video. Second, sets of consecutive tongue images of an arbitrary subject pronouncing phoneme combinations are acquired to preserve original continuous tongue movements as much as possible. Here the phoneme combinations are defined according to the visibility of the tongue; "the start is invisible, the middle is visible, and the end is invisible." These combinations contain all variations of tongue movements that appear in spoken English. Finally, we capture videos of seven subjects (except for the target subject in the input animation) pronouncing consonants and representative symbolic sounds produced by different articulation regions. If a target subject's image which the teeth can be seen is available, the teeth-database can be automatically updated with the target subject's teeth images. This allows a representation of individuality.

3. Embedding the Inner Mouth

The overview of our method is shown in Figure 1 (a). From a human skull bone structure, we assume that the distance from the position of the *Anterior Nasal Spine* (ANS) to the central position of the upper teeth is always constant. Similarly, the distance from the position of the chin to the central position of the lower teeth is also constant. To embed teeth into an input animation with these assumptions, we must first detect ANS and chin feature points using a feature point detector developed by Irie et al. [2011]. Then the distance between the central position of the upper and lower teeth is calculated using the detected feature points. Next, teeth images that have the closest teeth distance to that of the input animation are selected from the teeth-database. In addition, consecutive tongue image sets are also selected from the database according to sets of phoneme combinations on the basis of sentence information. Image sets are connected at the invisible point of the tongue. Finally, we reconstruct the images around the mouth using patch-based texture synthesis proposed by Mohammed et al. [2009] with the mouth-database.

4. Results and Discussions

The result is shown in Figure 1 (b) and the supplementary video. In the supplementary video, we demonstrate how our method is effective for significantly improving the quality of a "low-quality" input animation. Note that our method can represent inner mouth appearances, such as teeth nipping tongue's tip or tongue's back. Such representations have been difficult to produce with previous methods. One possible limitation of our method is robustness in different lighting conditions between the input animation and images in the databases. However, thanks to a seamless cloning technique, we can synthesize natural inner mouth images even if different lighting conditions exist. Therefore, we conclude that our method is useful as a post effect filter to improve the quality of an input speech animation created by any previous method.

References

CHANG, Y.-J., AND EZZAT, T. 2005. Transferable Videorealistic Speech Animation, *SCA'05*, pp. 143-151.

IRIE, A., TAKAGIWA, M., MORIYAMA, K. AND YAMASHITA T. 2011. Improvements to Facial Contour Detection by Hierarchical Fitting and Regression , *1st ACPR*, pp. 273-277.

MOHAMMED, U., Prince, J. D. -S., AND KAUTZ, J., 2009. Visio-lization: Generating Novel Facial Images, *SIGGRAPH'09*, pp. 57(1)-57(8).

* e-mail: doara-waseda@toki.waseda.jp

† e-mail: shigeo@waseda.jp

SIGGRAPH 2013, July 21 – 25, 2013, Anaheim, California.
2013 Copyright held by the Owner/Author.
ACM 978-1-4503-2261-4/13/07