Photorealistic Inner Mouth Expression in Speech Animation

- Supplementary Supporting Document -

Masahide KAWAI Tomoyori IWAO Daisuke MIMA Akinobu MAEJIMA Shigeo MORISHIMA

Waseda University

1. Tongue Classification

We classified tongue appearance when subjects uttered vowels and consonants in spoken English. If a tongue is visible, the classification is 1, the opposite is 0. Classifications for vowels and consonants are shown in Table 1 and Table 2 respectively.

Table 1 Classification of tongue appearance in terms of vowels

Name	Tongue’s appearance	classification	Examples
Front vowel	Forward of inner mouth	1	/e/, /æ/
		0	/i/, /iː/
Back vowel	Backward of inner mouth	0	/ɑː/, /u/

Table 2 Classification of tongue appearance in terms of consonant

Name	Tongue’s appearance	classification	Examples
Bi-labial	Between upper and lower lip	0	/p/, /b/
Labio-dental	Between upper teeth and lower lip	0	/f/, /v/
Dental	Between upper teeth and apex linguae	1	/θ/, /ð/
Alveolar	Between upper alveolar arch and apex linguae	1	/t/, d/
Palato-alveolar	Between portion passing hard palate from alveolar arch and bladal	1	/r/, /ʒ/
Palatal	Between hard palate and forward of lingual surface	1	/j/
Velar	Between soft palate and forward of lingual surface	0	/k/, /g/
Glottal	Between vocal cords	0	/h/

We could classify tongue appearance as above, and we combined vowels with consonants to newly determine phoneme combinations. The phoneme combinations are defined according to the classification (visibility of tongue) ; “the start is invisible, the middle is visible, and the end is invisible”. For example, /i/-/t//e/-/i/, /f/-/e/-/b/, etc.

We took tongue images of a subject pronouncing these phoneme combinations.

2. Gallery of Our Results

Figure 1 shows 2 subjects face images synthesized by our method. The left column represents input images, the middle column represents synthesis results, and the right column represents its close-ups of mouth region.

Note such Figure 1, our method can represent inner mouth appearances such as teeth nipping tongue’s tip and tongue’s back. These appearances are difficult to represent by previous methods.

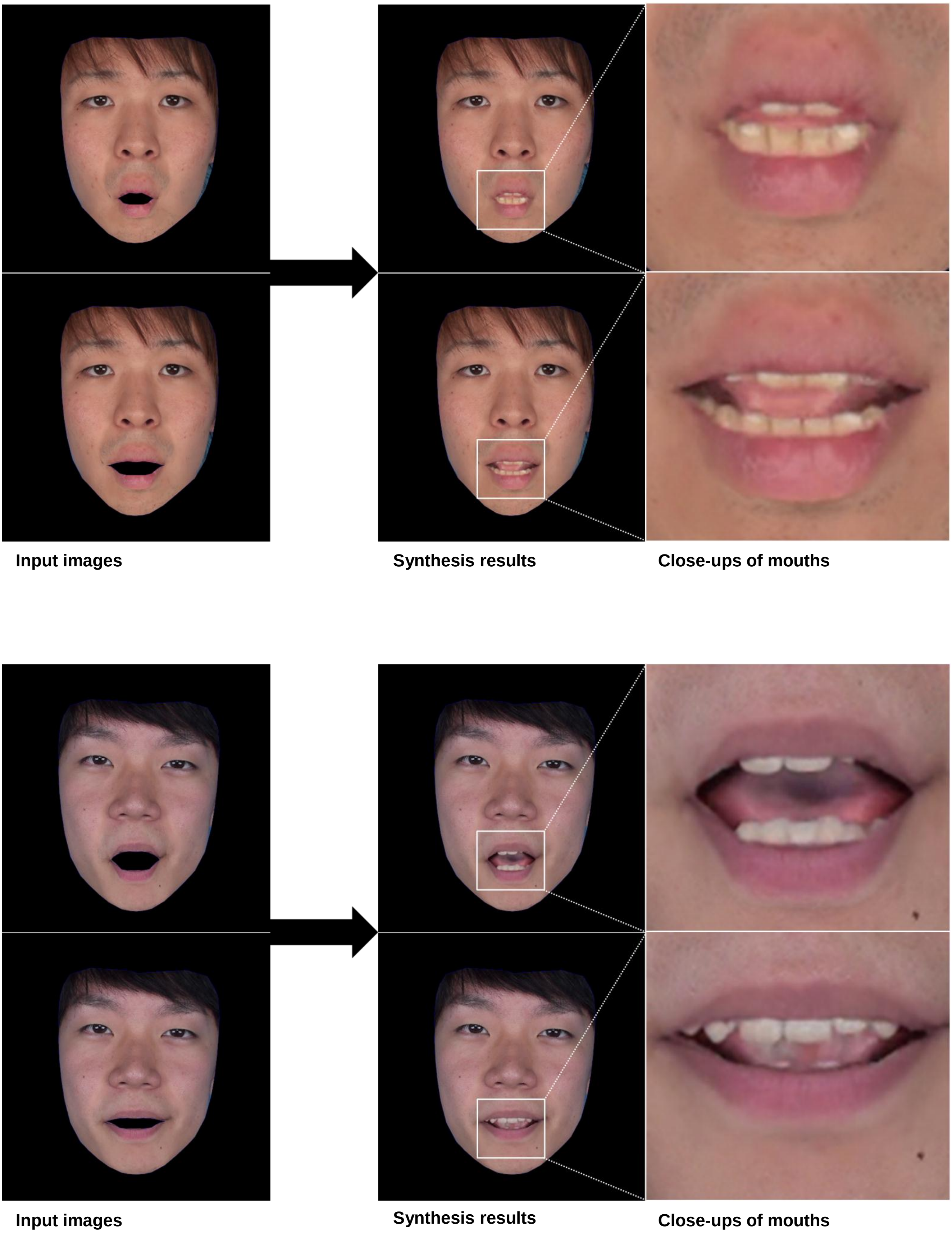


Figure 1: *Gallery of Our Results*

3. Our Result’s Image Compared with Photographed Image

Figure 2 shows the comparison with the original photographed images. We apply our method to “inner mouth-less” photographed images. The first column (a) represents the input images, the second column (b) represents the synthesis results and its close-ups of mouth region, and the third column (c) represents the original photographed images and its close-ups of mouth region. Comparing (b) with (c), we found that the synthesized and photographed appearances are similar. Some people mistake our resulting movie for the photographed movie. Please see the Supplementary Video.

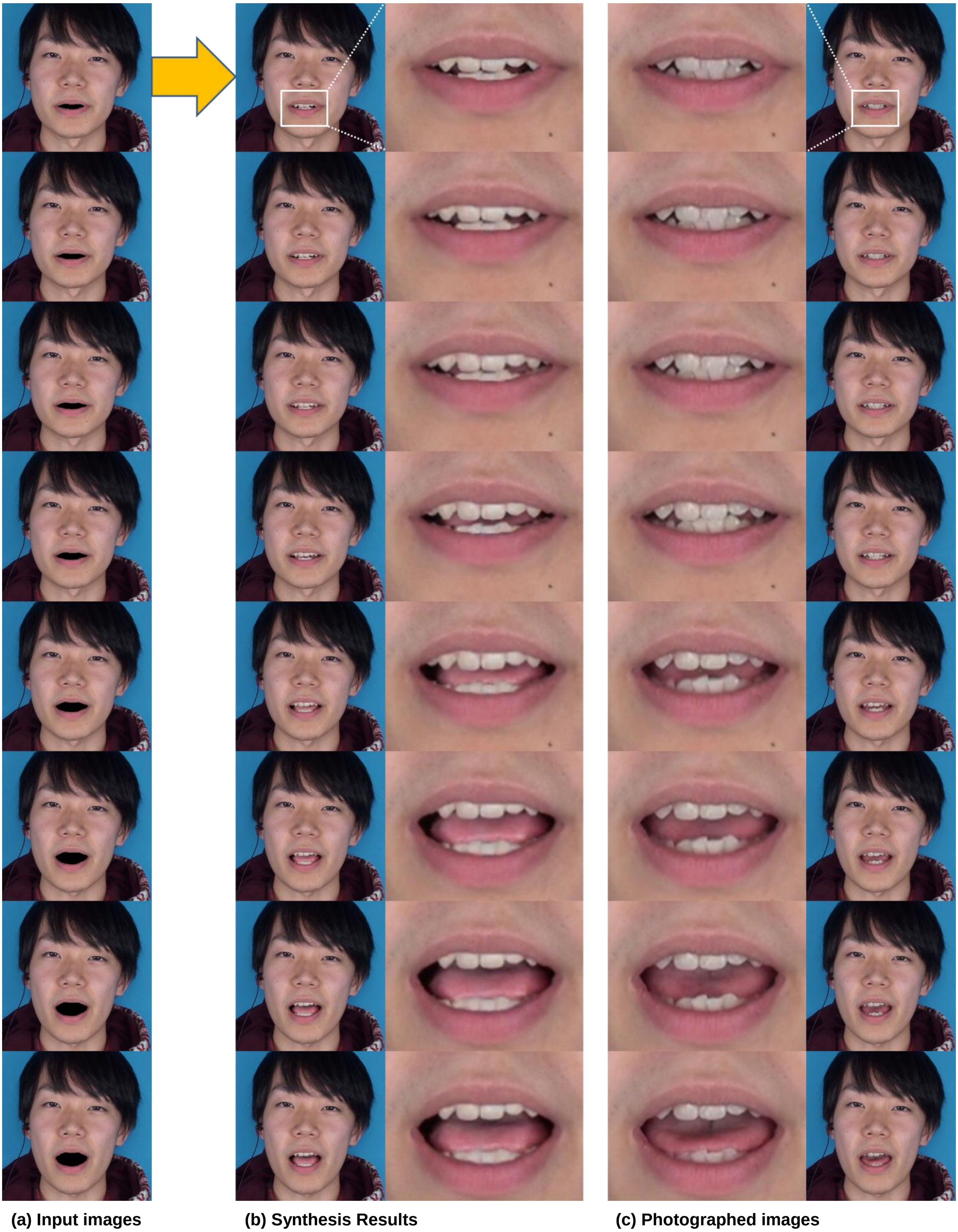

(a) Input images
(b) Synthesis Results
(c) Photographed images

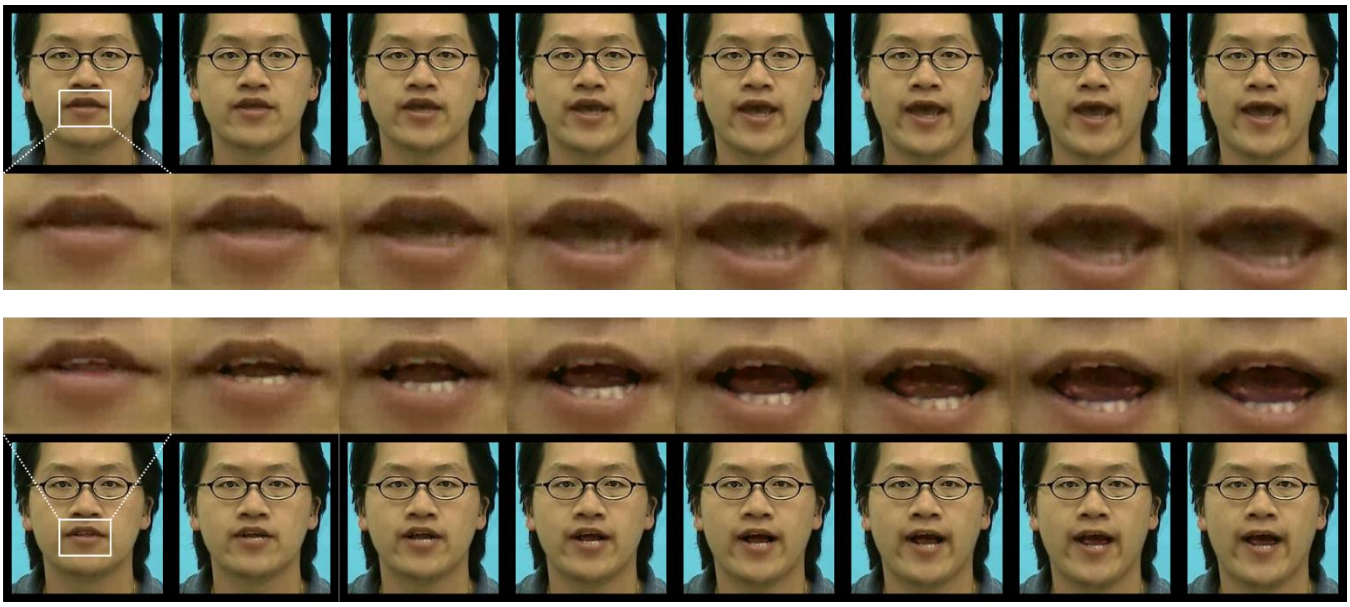
Figure 2: *Comparison Results (vs. Original photographed Images)*

4. Our Result Compared with Related Work

Figure 3 shows the comparison with "Transferable Videorealistic Speech Animation" proposed by Chang et al [2005]. The top row is image sequences synthesized by Chang et al, and the second row represents its close-ups of the mouth region, and the third and the fourth rows are our results.

We found that teeth and a tongue in the images generated by Chang's method are expanded and contracted. In contrast, our method can solve this problem and improve their resolution (the edges are more clearly).

[Chang *et al.* 2005]



Our method

Chang et al, "Transferable Videorealistic Speech Animation",
Proceedings of the 2005 ACM SIGGRAPH/Eurographics Symposium on Computer Animation.

Figure 3: *Comparison Results (vs. Chang et al.)*

5. 16 Close-ups from the Supplementary Video

Figure 4 shows 16 close-ups from the Supplementary Video.

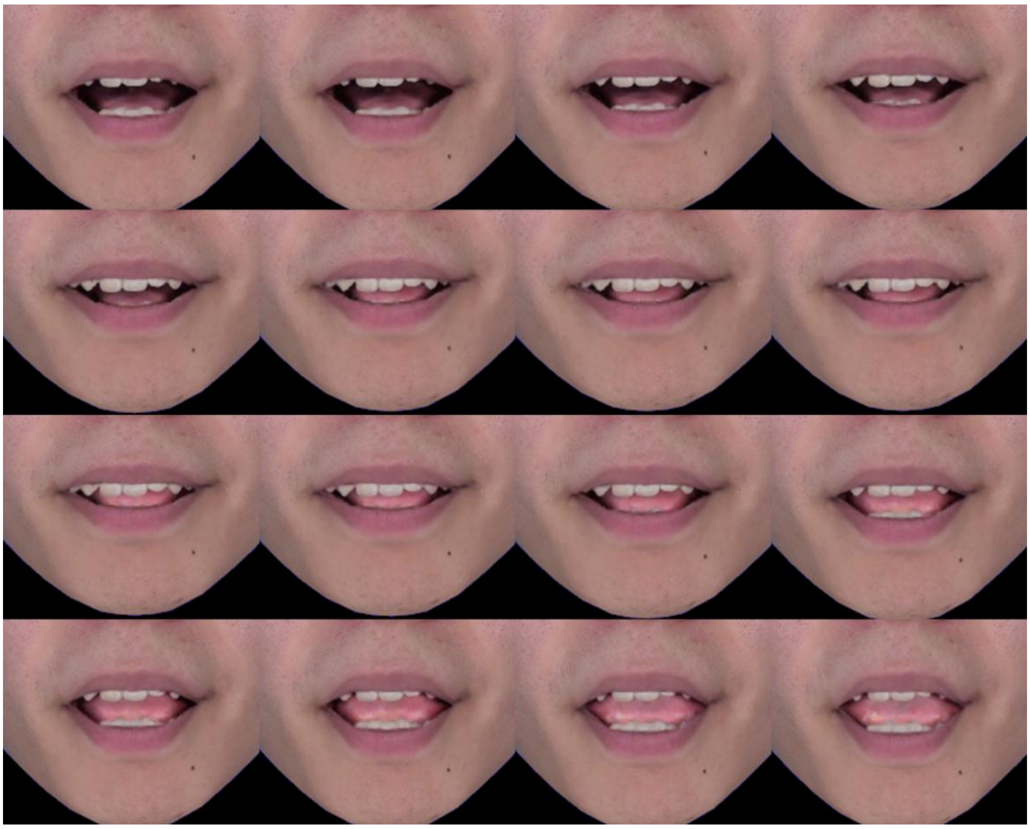


Figure 4: *16 Close-ups from the Supplementary Video*