Efficient Speech Animation Synthesis with Vocalic Lip Shapes

- Supplementary Supporting Document -

Daisuke Mima Akinobu Maejima Shigeo Morishima

Waseda University

◆ Lip Motion Estimation

Figure 1 shows a situation of lip motion estimation. To synthesize speech animations with our method, we estimate a lip motion which is appropriate for an input speech by the cost function (1).

$$\mathrm{E}(\boldsymbol{a}) = \sum_{i=1}^{N}\sum_{t\in\Omega}(D(t)_i + G(t)_i) \qquad (1)$$

$$= \sum_{i=1}^{N}\sum_{t\in\Omega}(|x(t)_{i-1} - x(t)_i|^2 + |\nabla x(t)_{i-1} - \nabla x(t)_i|^2)$$

$$\boldsymbol{x}_i = \boldsymbol{a}_i\hat{\boldsymbol{x}}_j \quad (0 < j < M)$$

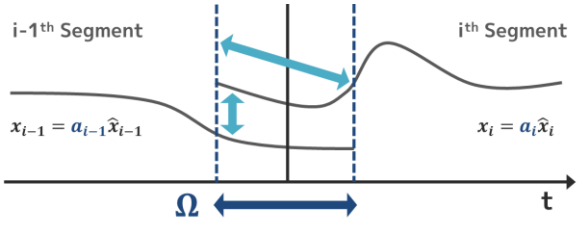
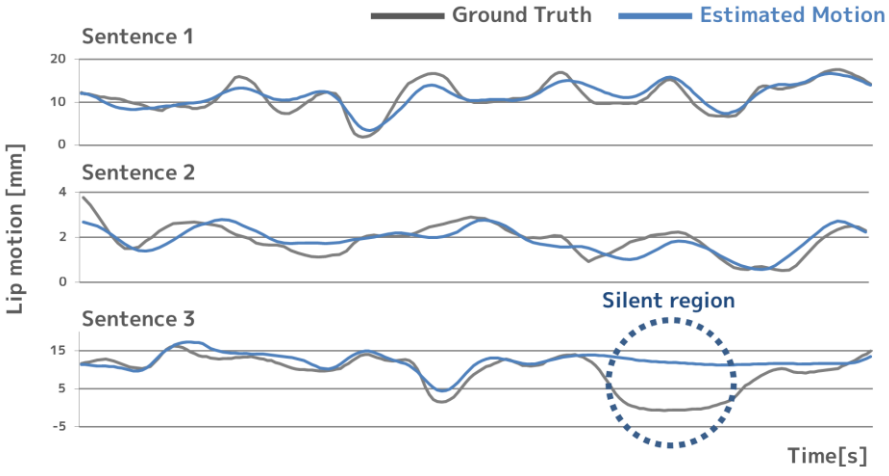


Figure 1: *A situation of estimation.*

where E is an entire cost of a target sentence, N is the number of vowel segments in the sentence and Ω is an overlap region(a consonantal segment) between $i-1^{\text{th}}$ and i^{th} units of vocalic lip motions. Functions D and G represent distance and gradient differences between successive lip motions ($\boldsymbol{x}_{i-1}$ and $\boldsymbol{x}_i$) in the region Ω respectively. The i^{th} lip motion $\boldsymbol{x}_i$ is calculated from $\hat{\boldsymbol{x}}_j$ which is an eigenvector of j^{th} units of vocalic lip motions and $\boldsymbol{a}_i$ which is a principal component score of $\boldsymbol{x}_i$. M is the number of all units of vocalic lip motions learned from training data and the j^{th} unit assigned to i^{th} vowel segment that has a same combination of phonemes. We eventually acquire the optimal $\boldsymbol{a}$ that provides the lowest E by gradient descent method, which means that we obtain an optimum lip motion $\boldsymbol{x}$ for the input speech.

◆ Estimation Results Compared with Ground Truths



	RMSE (%)
Sentence 1	1.48
Sentence 2	2.64
Sentence 3 (except for silent regions)	2.69

Figure 2: *Comparison between calculated lip motions and Ground Truths. In silent regions some gaps are observed, which causes closing a mouth to take breaths. It is a future task to generate the lip motions without any relation to utterances.*

◆ Subjective Evaluation

To evaluate the performance of our method, we conducted two types of subjective evaluations for twenty subjects:

1) The first evaluation is a comparison between three methods for interpolating vowel target lip shapes, which are linear interpolation, simply using GMMs [Yano et al. 2007] and a combination of lip motion estimation with GMMs (our method). We displayed two types of animations which are generated from a same sentence, and a subject chose one which has more realistic lip motions. The work is conducted for five different sentences (which means that a number of times for one subject is fifteen) to decide which one is used to synthesize the most realistic result (figure 3).

2) The second evaluation is to score reality of speech animations on a zero-to-seven scale to compare with a created animation by a professional artist and the same one whose lip motions are replaced by our results (figure 4). Before the evaluation, we displayed two speech animation, which are synthesized by linear interpolation (which is an example of score 1) and created by an artist (which is an example of score 5), as a guideline for deciding the scores.

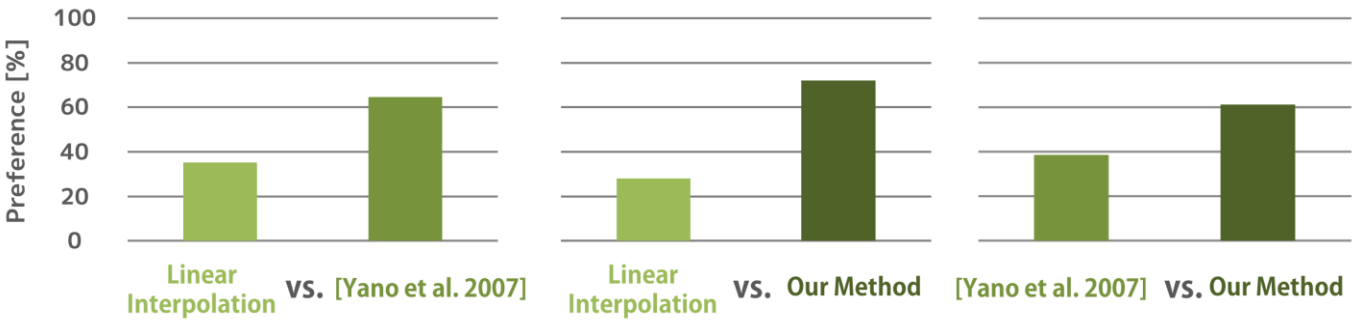


Figure 3: *Comparison between three methods of interpolating target lip shapes.*

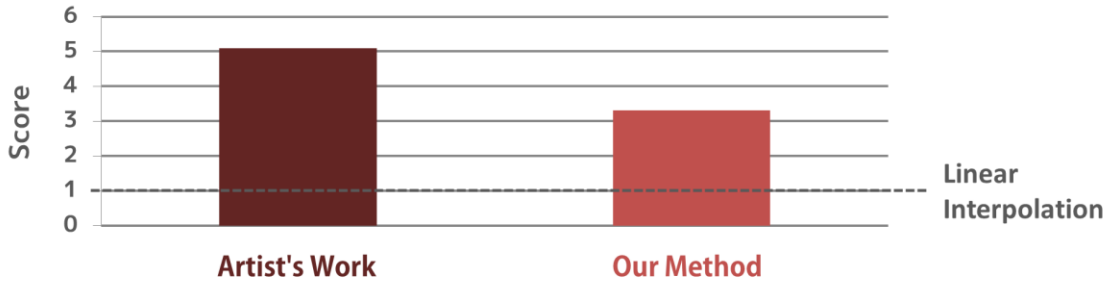


Figure 4: *Comparison between created animation by a professional artist and our result. A speech animation which has average score 3.2 can be synthesized with our method. The score is lower than a handmade animation (which has average score 5.1), however, it is possible with our method to synthesize a reasonable animation easily and quickly, only from an input voice and text.*

References

YANO, A., et al. 2007. Variable Rate Speech Animation Synthesis. In Proc. ACM SIGGRAPH 2007, Poster, no.18.