Efficient Speech Animation Synthesis with Vocalic Lip Shapes

Daisuke Mima* Akinobu Maejima Shigeo Morishima

Waseda University

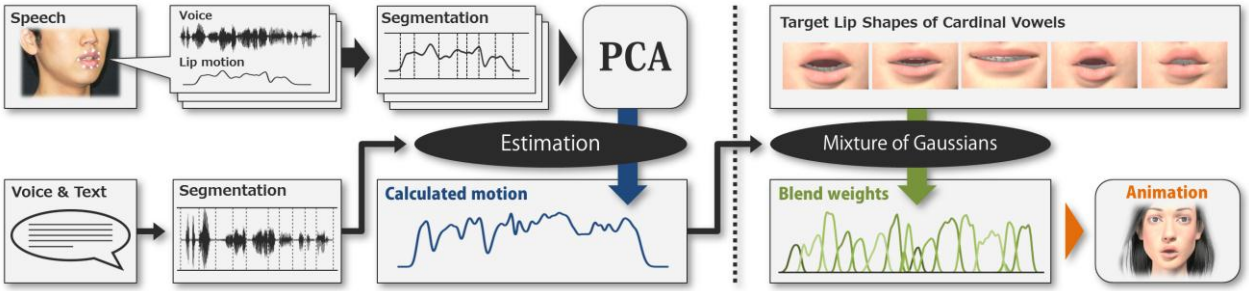

Figure 1: *Animation synthesis using our method.*

1 Introduction

Computer-generated speech animations are commonly seen in video games and movies. Although high-quality facial motions can be created by the hand crafted work of skilled artists, this approach is not always suitable because of time and cost constraints. A data-driven approach [Taylor et al. 2012], such as machine learning to concatenate video portions of speech training data, has been utilized to generate natural speech animation, while a large number of target shapes are often required for synthesis. We can obtain smooth mouth motions from prepared lip shapes for typical vowels by using an interpolation of lip shapes with Gaussian mixture models (GMMs) [Yano et al. 2007]. However, the resulting animation is not directly generated from the measured lip motions of someone's actual speech.

In this study, we propose a novel technique for synthesizing speech animation from an input utterance (voice and text) by interpolating lip shapes of the five cardinal vowels. In addition, we determine the blend weights of the target lip shapes according to calculated lip motions from training data of real speech. Our method has the following two advantages: it can be used for never -heard-before dialogs without complicated computation or advance preparation, and it can generate smooth speech animation with minimum number of target shapes.

2 Lip Motion Estimation

The first step of our work is to classify units of vocalic lip motions from real speech motions. By using a motion capture system, we first obtain the motion of each marker placed on the lips of an actor as the actor uttered various sentences (199 types in this study). Second, the captured lip motions are divided into individual phoneme segments [Lee et al. 2009], and each part of segmented motions is categorized into 11 groups, five types of vowels and six types of consonants. Third, the lip motion that belongs to successive phoneme segments (*/consonant/-/vowel/-/consonant/*) is deemed "a vocalic lip motion", and motions that have same phonemes are grouped as a unit.

The second step is to define a cost function (1) to presume lip motions from an input speech.

$$\mathrm{E}(\boldsymbol{a}) = \sum_{i=1}^{N} \sum_{t \in \Omega} (D(t)_i + G(t)_i), \quad \boldsymbol{x}_i = \boldsymbol{a}_i \hat{\boldsymbol{x}}_j \ (0 < j < M) \tag{1}$$

where functions D and G respectively, represent distance and gradient differences of successive lip motions ($\boldsymbol{x}_{i-1}$ and $\boldsymbol{x}_i$) in the region Ω. We eventually acquire an optimum $\boldsymbol{a}$ that provides the lowest E by the gradient descent method (the details are provided in supplementary supporting document). In this study, because of a small repertoire of target lip shapes for synthesis, we define lip motions as transitions of length between neutral and opening lip shapes in the horizontal and vertical directions; provided, however, that we may add a new category of target shapes.

3 Animation Synthesis Result

We determine the blend weights for the target shapes (five types in this work) by fitting GMMs, whose variables are obtained from equation (2) to an estimated lip motion.

$$\{A, p, \sigma\}_{nm} = \underset{\{A,p,\sigma\}_{nm}}{\operatorname{argmin}} \sum_{t=0}^{T} \left| x_t - \sum_{n=1}^{N} \sum_{m=1}^{M} K_n A_{nm} \exp\left(-\frac{(t-p_{nm})^2}{2{\sigma_{nm}}^2}\right) \right| \tag{2}$$

where $\{A, p, \sigma\}$ are variables of Gaussian functions, x is an estimated lip motion, T is an ending time of an input speech, N is the number of target lip shapes and M is the number of Gaussian functions. The actor's lip shapes equivalent to the targets are K_ns, which vanish in the event that p_{nm} is found in a segment that has a different vowel from that of K_n. With our method, natural speech animations are generated even if there are a few target shapes, because the blend weights for the shapes are calculated according to the well-approximated lip motions of real people.

In this study, we demonstrate speech animation generated by our method in a supplementary video, in which complex lip motions were observed, such as closing the mouth immediately before bilabial consonants. The advantage of our method is to synthesize a speech animation by interpolating a small number of target lip shapes with GMMs. Each parameter of the Gaussian functions is computed on the basis of estimated lip motions, and this means more realistic lip motions can possibly be obtained than those obtained when GMMs are simply applied to an input sentence [Yano et al. 2007]. It remains for future work to automatically control an estimation of lip motions according to utterance speed or facial expressions changes.

References

TAYLOR, S.L., et al. 2012. Dynamic Units of Visual Speech. In Proc. ACM SCA 2012, 275-284.

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

LEE, A., et al. 2009. Recent Development of Open-source Speech Recognition Engine Julius. In Proc. APSIPA ASC 2009, 131-137.

*e-mail : ai-zumi@ruri.waseda.jp

SIGGRAPH 2013, July 21 – 25, 2013, Anaheim, California.

ACM 978-1-4503-2261-4/13/07