Variable Rate Speech Animation Synthesis

Akane Yano\textsuperscript{1}  Hiroyuki Kubo\textsuperscript{2}  Yoshihiro Adachi\textsuperscript{3}  Demetri Terzopoulos\textsuperscript{4}  Shigeo Morishima\textsuperscript{5}

\textsuperscript{1,3,5} Waseda University, Tokyo, Japan
\textsuperscript{4} University of California, Los Angeles, CA, USA

1. Introduction

Speech animation has traditionally been achieved by two main approaches: image-based methods [Ezzat et al. 2002] and key-framing methods [Cohen and Massaro 1993]. Both approaches require large databases that include many images or 3D shapes with facial expressions and speech lip shapes.

Recently, mocap-based facial animation techniques have been utilized to create natural speech animation. However, it is difficult to create natural lip postures with variable speech rates by reusing the motions of pre-captured lip markers.

To ameliorate these problems, we present a novel lip-sync interpolation technique that supports flexible speech rates and does not need large databases. Our system can synthesize speech animation with variable speech posture from input speech.

2. Producing Lip-Sync

To synthesize natural lip-sync animation using input speech we must find the significant factors of lip motions in utterance. First, by using a motion capture system, we obtained the motion of each marker on the lips while the actor uttered various sentences. After analyzing the feature-point movement, we found that the lip velocity was almost constant when a specific vowel was uttered, even under variable speech rate (Figure 1). In fact, the faster a phoneme is spoken, the smaller is the range of lip motion.

In our system, we classify the target lip shapes into six groups and determine each target lip shape associated with a vowel. Then we set phoneme units according to the syllable. In the case that a phoneme unit includes a plosive or a stop consonant, such as $b$, $d$, $g$, $p$, $t$, and $k$, we separate this phoneme unit and allocate an appropriate lip shape target to this type of phoneme among the six groups.

Second, we propose an interpolation method based on the analysis of motion capture data in a variety of speech rates. The following equation governs the speech variation process:

$$y = \sum A_i \gamma_i \exp \left[ -\frac{(t - c_i)^2}{2\sigma_i} \right]$$

where $y$ is the motion rate of the $j$th phoneme unit, $c_i$ is the amount of time $t$, $\gamma_i$ is the rate of the scaling function, and $\sigma_i$ is a parameter for the lip shape target corresponding to the phoneme unit. The velocity of lip motion remains constant with eq (1).

3. Speech Animation Synthesis Result

To generate more attractive animation, we applied our lip-sync technique using a parameterized muscle model [Lee et al. 1995]. This model enables users to create a character’s facial expression easily.

Figure 3 shows six kinds of synthesized lip-sync sequences. Group (I) is generated when the duration of phoneme units is short and Group (II) is generated when they are long. Both (I) and (II) are synthesized under the same conditions for the speech posture parameters. The speech shape is reflected by the speech rate in all of the sequences. Figures 3(c) and 3(f) show the results of applying our system to an anime character after adapting the facial muscle structure. As the result, users can reflect various emotions in the lips of a cartoon character.

4. Conclusion

We developed a method for creating natural lip-sync animation which can adapt to changes in the rate of speech. We applied our new method to two types of characters. Both demonstrate speech movements that depend on a variable speech rate. Consequently, we can apply our system to produce entertainment using small facial animation databases.

Acknowledgment

This study was supported by the Japan Science and Technology CREST project.

References

