Speech-Driven Realtime Lip-Synch Animation with Viseme-Dependent Filters

Shin'ichi Kawamoto*
JAIST

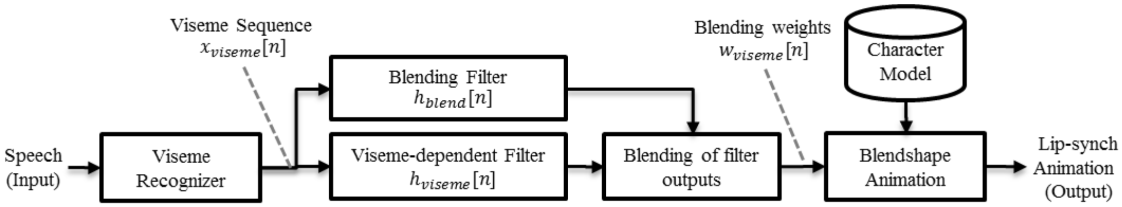

Figure 1: *Processing Flow of Our System*

1 Introduction

This paper presents a speech-driven realtime lip-synching technique by customizable viseme-dependent filters for blending weight sequences based on blendshapes, linear shape interpolation model. Lip-synching is one of fundamental components for creating facial animation. Mouth movement is synchronized along with the speech, when a character utters a word or phrase. Especially, speech-driven real-time lip-synching animation is useful for helping speech communication. Various speech-driven lip-synch animation techniques were proposed by many researchers (e.g. [Morishima 1998]). Most researchers constructed a mapping between speech and mouth-shape directly. These direct mapping approaches can realize lip-synching with small delay. However, it is sometimes unnatural since mouth movement is mismatched between the speaker and the pre-designed characters. In this case, the dynamic property of mouth movement should be designed for pre-designed characters. To solve this, we consider the customization of mouth movement by viseme-dependent filters designed for each mouth shape of given characters. Viseme is a basic unit of mouth shapes that are classified visually, and mapped to some phonemes.

2 Our Approach

If CG characters are animated with the same blending weight sequence for lip-synching, then the mouth movement speed of each character depends on the pre-designed mouth shapes in the blendshapes approach. In such situations, lip-synch animation sometimes seems unnatural when the mouth moves rapidly. To prevent such a situation, the blending weight sequence should be designed in consideration of the mouth movement speed. Since the target application is realtime, the lip-synching, generating process of blending weight sequence can only use the history of the recognized viseme sequence. Thus, we apply a filtering method of observed viseme sequence. The filters depend on the viseme, and are designed in consideration of mouth movement speed based on mouth shapes as target shapes.

Overview of our lip-synch system is shown in Fig. 1. In our approach, a speech signal and a CG character data were given as inputs. This system outputs blending weights of each mouth shape based on blendshapes, which is basic technique of animation and widely used in CG software. The target shapes of the mouth are designed to corresponding viseme. First, we convert speech to a viseme sequence on the fly using a viseme recognizer that is implemented based on the APIs of the speech recognizer 'julius' [Lee and Kawahara 2009]. The viseme recognizer outputs binary sequences of each viseme like step inputs for filters. The number of channels is 13, since 13 Japanese viseme were adopted in this system. If the corresponding viseme were observed in the current analysis frame, then the corresponding element of result vector is 1, and the other elements are 0. Subsequently, each vector sequence is inputed to the viseme-dependent FIR filters $h_{viseme}[n]$ ($h_{viseme}[n] \geq 0$ and $\sum h_{viseme}[n] = 1$). The mouth movement speed can roughly set the upper bound by customizing the parameters of these filters, since the maximum value of filter parameter is proportional to the maximum speed of mouth movement. In addition, the length of the filter means the duration of transition between visemes. The viseme vector sequence is also inputted to the Blending Filter $h_{blend}[n]$ for generating a smoothed viseme vector sequence. The output of the blending filter is used for switching outputs of the viseme-dependent filter smoothly. Afterwards, both of filter outputs were mixed for generating blending weights of the blendshape-based lip-synch animation. In this process, the blending weight of the viseme is calculated by the product of two filter outputs that correspond to the same viseme. Finally, Lip-synch animation is generated using blendshapes with calculated blending weights for each viseme. In this system, we use Galatea-FSM as a facial animation tool [Yotsukura et al. 2003]. The input speech is also outputted with delay in accordance with the processing time of the lip-synch animation.

Currently, our lip-synch system worked well with about 0.3 sec of delay from the input speech (See supplementary video). This delay depends on a few factors. Viseme recognition also needs some processing time. This delay is about 0.2 sec for analyzing the input speech in the current environment. It has a tradeoff between the delay and the accuracy of the viseme recognition results. The length of the filters for generating blending weights is also related to the delay of the animation. It has a tradeoff between the delay and the smoothness of the mouth shape transition. Of course, the minimum delay of speech output depends on the hardware specification, such as the delay of the sound device.

Acknowledgements

This study was supported in part by JSPS KAKENHI Grant Number 24700191.

References

LEE, A., AND KAWAHARA, T. 2009. Recent development of open-source speech recognition engine julius. In *APSIPA ASC*.

MORISHIMA, S. 1998. Real-time talking head driven by voice and its application to communication and entertainment. In *Proc. AVSP*.

YOTSUKURA, T., MORISHIMA, S., AND NAKAMURA, S. 2003. Model-based talking face synthesis for anthropomorphic spoken dialog agent system. In *Proc. 11th ACM MM*, 351–354.

*e-mail:kawamoto@jaist.ac.jp

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

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