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

Supplemental Material for SIGGRAPH 2013 Poster

Shin-ichi KAWAMOTO (JAIST)

Introduction

Motivation

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 realtime lip-synching animation is useful for helping speech communication.

Aim

Realizing speech-driven realtime lip-synching based on blendshapes, linear shape interpolation model.

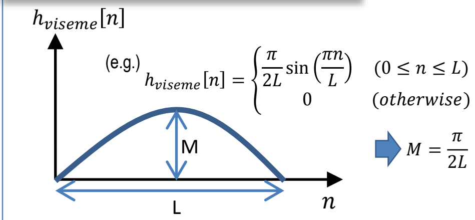
Problem

Simple solution is to construct 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.

Our solution

we consider customization of mouth movement by viseme-dependent filters designed for each mouth shape of given characters.

Viseme-dependent Filter



(e.g.) $$h_{viseme}[n] = \begin{cases} \frac{\pi}{2L}\sin\left(\frac{\pi n}{L}\right) & (0 \le n \le L) \\ 0 & (otherwise) \end{cases}$$

⇒ $M = \frac{\pi}{2L}$

L : filter length (#tap)

$\Longrightarrow$ Transition Time between Mouth Shapes

$$\mathrm{M} = \max_{n}(h_{viseme}[n]) = \frac{\pi}{2L}$$

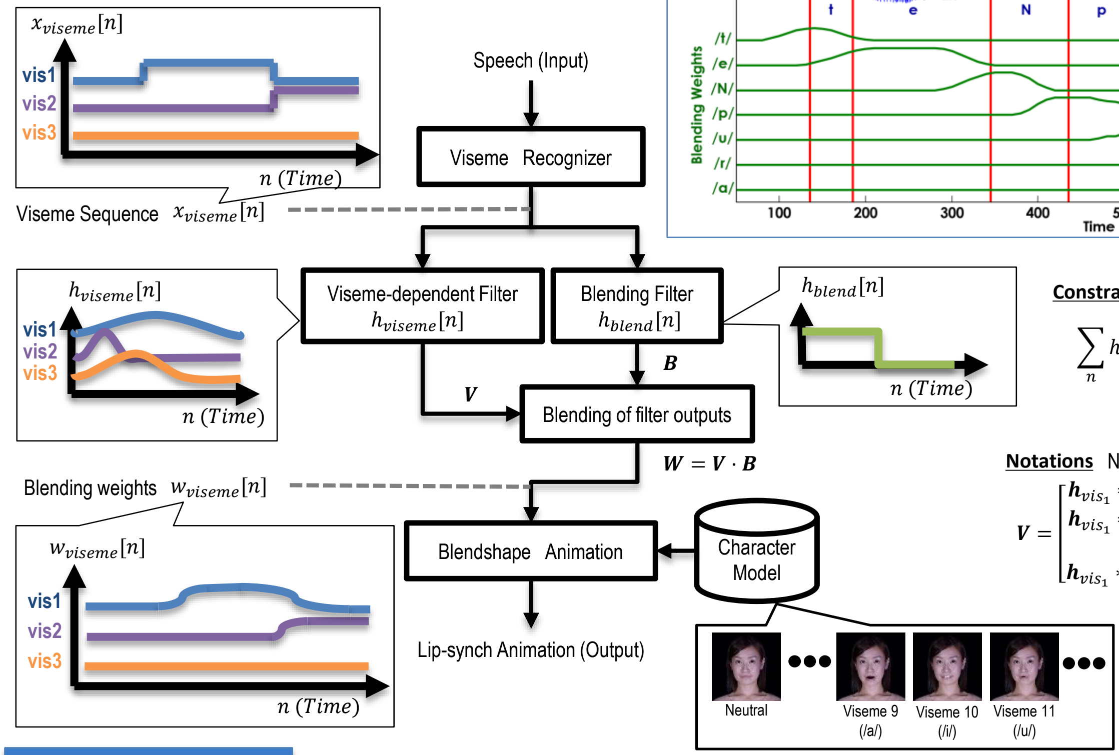
$\propto$ Maximum Mouth Movement Speed

How to decide this parameters?

⇒ **[default]** depends on Euclid distance between a target-shape and a neutral-shape (*normalize maximum mouth movement speed*)

[customize] depends on the desirable transition time (*eg. Consonants /p/, /b/, and /m/*)

Processing Flow



Viseme Recognizer

Viseme recognizer iterates the following processes at a constant frequency:

[Step.1] Estimates phonemes and its duration by using an HMM-based speech recognizer from observing input speech at this time.

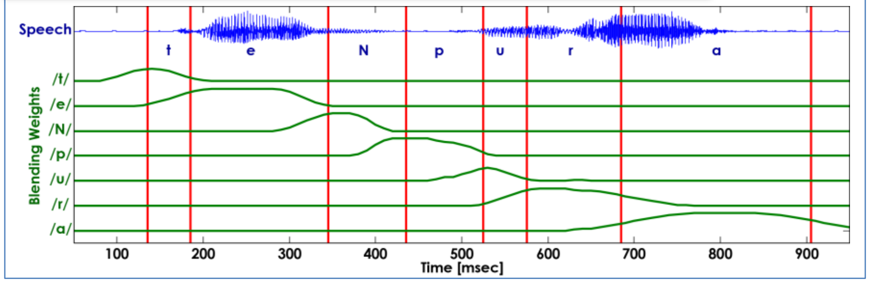
[Step.2] Convert phoneme to viseme based on table lookup.

Phoneme-Viseme Mapping Table (Japanese)

Viseme ID	Phonemes
1	r, ry
2	b, by, m. my, p, py
3	t
4	d, n, ny
5	g, gy, hy, k, ky, N
6	f
7	ch, dy, j, s, sh, ts, z
8	w
9	a
10	I
11	u
12	e
13	o

Note: Viseme is a basic unit of mouth shapes that are classified visually

Example of Our Result (Speech & Blending Weights)



Constraints of Blending Filter

$$\sum_{n} h_{blend}[n] = 1, \quad h_{blend}[n] > 0$$

Notations Note: '*' means a convolution operator

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{h}_{vis_1} * \boldsymbol{x}_{vis_1} & \boldsymbol{h}_{vis_2} * \boldsymbol{x}_{vis_1} & \cdots & \boldsymbol{h}_{vis_N} * \boldsymbol{x}_{vis_1} \\ \boldsymbol{h}_{vis_1} * \boldsymbol{x}_{vis_2} & \boldsymbol{h}_{vis_2} * \boldsymbol{x}_{vis_2} & \cdots & \boldsymbol{h}_{vis_N} * \boldsymbol{x}_{vis_2} \\ \vdots & \vdots & \ddots & \cdots \\ \boldsymbol{h}_{vis_1} * \boldsymbol{x}_{vis_N} & \boldsymbol{h}_{vis_2} * \boldsymbol{x}_{vis_N} & \cdots & \boldsymbol{h}_{vis_N} * \boldsymbol{x}_{vis_N} \end{bmatrix}$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{h}_{blend} * \boldsymbol{x}_{vis_1} \\ \boldsymbol{h}_{blend} * \boldsymbol{x}_{vis_2} \\ \vdots \\ \boldsymbol{h}_{blend} * \boldsymbol{x}_{vis_N} \end{bmatrix} \qquad \boldsymbol{W} = \begin{bmatrix} w_{vis_1} \\ w_{vis_2} \\ \vdots \\ w_{vis_N} \end{bmatrix}$$

Result & Discussion

Our lip-synch system worked well with about 0.3 sec of delay from the input speech.

⇒ Factors of its delay:

(1) Viseme recognition also needs some processing time.

Delay ⇔ tradeoff ⇔ Accuracy of viseme recognition results

(2) Length of the filters for generating blending weights is also related to the delay of the animation

Delay ⇔ tradeoff ⇔ Smoothness of mouth shape transition

(3) Minimum delay of speech output depends on the hardware specification (*eg. delay of sound device*)

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