Generating Expressive Facial Mesh Animation : A Survey

HJW
Institution1 address
self@yukinyaa.moe

Abstract

With technology allowing for increasing realism in games and movies, facial animation is still a very challenging task.

"Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."

1. Introduction

Human tend to be very sensitive to facial motion psychologically. Slightest uncanniest in facial animation directly leads to hurt overall experience [3]. Facial animation can be applied to multiple applications such as computer games, e-commerce, immersive VR telepresence, and movies. Yet achieving realistic facial animation is a challenging task. So, delivering natural expressive facial animation is a great interest in graphics field.

Animating high-quality expressive face is very labor-intensive job when done by animator. Another approach to animate face is to capture human face animation in 3D. Face capture is a well-understood field(cite here), yet such approach requires gigabytes of data from expensive capture system, and is hard to manipulate. Therefore, it is necessary to simplify such process.

To simplify such process, one can automatically generate facial animation or can simplify animating produce.

In this survey, I introduce and compare three research that animate expressive facial animation :

- JALI [1] and VisimeNet [7], a linguistic approach to lip-sync.
- MeshTalk [6], a deep learning method.
- D3DExpression [5], LSTM method which replicate facial expression.

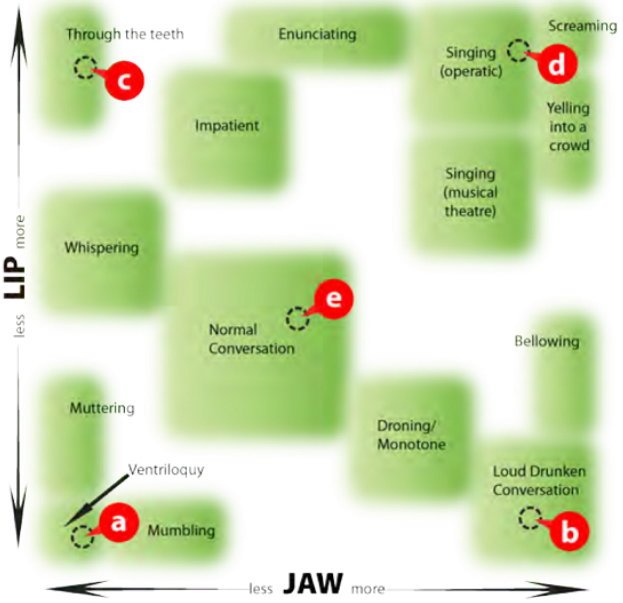

Figure 1. Speaking styles of JALI viseme field

2. Methods

2.1. JALI and Visemenet

2.1.1 JALI

Lip-sync can be done by linguistic approach which is mapping text to phonemes, then phonemes to visemes [2]. Viseme is a specific facial shape when making certain sound. A traditional facial rig can have many-to-one mapping from phonemes to visemes, or many-to-many using dynamic visemes. This approach can achieve realistic acoustic motion on the model.

JALI [1] is a state-of-the-art viseme model that can generate lip-sync animation from English text and voice. JALI takes jaw and lip activation multipliers into consideration, since jaw and lip is the most significant acoustic motion in face. As shown on Fig. 1, different speaking styles shows different jaw lip activation level multipliers, which can be

animated more intuitively.

JALI use text-to-phoneme speech library built in to OSX. This converts English text into a phonemic representation. Next, text is forced aligned, which is a process that aligning phonemes to text. Audio must be annotated with the beginning, middle, and end of each phoneme. This is typically done by training Hidden Markov Model. Several tools exist for this task such as HTK and SPHINX.

JALI animates a facial rig by producing sparse animation keyframes for visemes, by some linguistic rules. When animating, lexical stress and co-articulation rule is considered. Co-articulation is a rule which visemes interact with each other. For example, duplicated visemes such as /p/ and /m/ in "pop man" are co-articulated into one MMM viseme, and tongue-only visemes have no influence on the lips, and lip takes shape of the surrounding visemes.

Unlike other methods, JALI can produce sparse keyframed procedural animation. Generated animation can be edited to achieve more idiosyncratic animation. It is also easy to accompany the method with other animation, since animation only affects jaw and lip parameter of the rig.

Limitation of the work is that it lacks some auditory parameter such as rate of speech, and jitter and shimmer. Also, emotional speech styles are not considered. Emotional styles can affect subtle parameter, but is a critical part of expressiveness of the animation.

2.1.2 Visemenet

Visemenet is a deep-learning based lip-sync method based on JALI model. Unlike JALI, Visemenet takes auditory data only to generate JALI lip-sync animation. Visemenet tries to replicate linguistic model with multiple deep-learning LSTM layers.

As shown on Fig. 2, Visemenet mainly have three stages. Phoneme group stage, landmark stage, and viseme stage. Phoneme and landmark stage is a pretrained model with relatively large amount of dataset, and Viseme stage a main model which predicts three types of JALI animation parameters.

Auditory data is converted to multiple features that has been used in previous studies. 65 feature vectors of 3 type per each frame is used, and 24 frames are used as input of all network, sum 1560 dimension.

Phoneme groups are 20 groups of phonemes which corresponds to certain visemes. Phoneme group stage estimates phoneme group from audio features.

38 artist-defined landmarks in real face with jaw, lip and nose is considered in landmark stage. Landmark stage estimates 2D-displacement of this landmark, resulting in 76 dimension output.

Pretrain networks are trained with about total 15 hour worth of publically available audiovisual facial dataset. Dataset used in training must have English transcript and corresponding voiced video. Most of the video dataset available used for facial recognition has transcript and high quality video recorded with camera at front of the face.

Viseme stage is a main network that predicts JALI model variables. Input of the network is 1560 dimension audio features and 20 dimension predicted phoneme groups, and 76d landmark displacement groups. JALI control activation and visemes, co-articulation parameters are predicted in individual LSTM networks. LSTM layers have 256-dimension memory state, with 2 fully connected decoders.

Viseme stage is trained with 1 hour of rig motion curves. Experienced animator created dataset corresponding audio.

Visemenet successfully mapped audio to speech motion curve. Pretrained network used hand-engineered audio features, replacing such features with learned feature will improve performance, which is happening at image processing.

This research mixes animator-centric technique and deep-learning method. But unlike JALI that shares low-dimension output, Visemenet generates frame-by-frame animation, which is harder to manipulate.

2.2. MeshTalk

MeshTalk is a generic method for generating full facial mesh animation from speech. MeshTalk network can generate lip-sync animation from a single frame of generic human facial mesh and audio signal, and also can in expressive data from mesh animation.

The network resembles Variational Autoencoder with multiple latent space as shown in Figure 3. Target mesh $\hat{h}$ is estimated from template mesh h and latent space c by computing function $\mathcal{D}$.

$$\hat{h}_{1:T} = \mathcal{D}(h, \mathsf{c}_{1:T,1:H}) \tag{1}$$

Sequence of latent space $\mathsf{c}_{1:T}$ is derived from audio sequence $\mathsf{a}_{1:T}$ and expression signal mesh sequence $\mathsf{x}_{1:T}$. c and a are first mapped to $T * H * C$ dimensional latent space, then passed through Gumbel-softmax [4] over every classification head.

2.3. D3DExpression

2.4. Discussion

References

[1] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. JALI: An animator-centric viseme model for expressive lip synchronization. *ACM Transactions on Graphics*, 35(4):1–11, July 2016. 1

[2] T. Ezzat and T. Poggio. MikeTalk: A talking facial display based on morphing visemes. In *Proceedings Computer Animation '98 (Cat. No.98EX169)*, pages 96–102, June 1998. 1

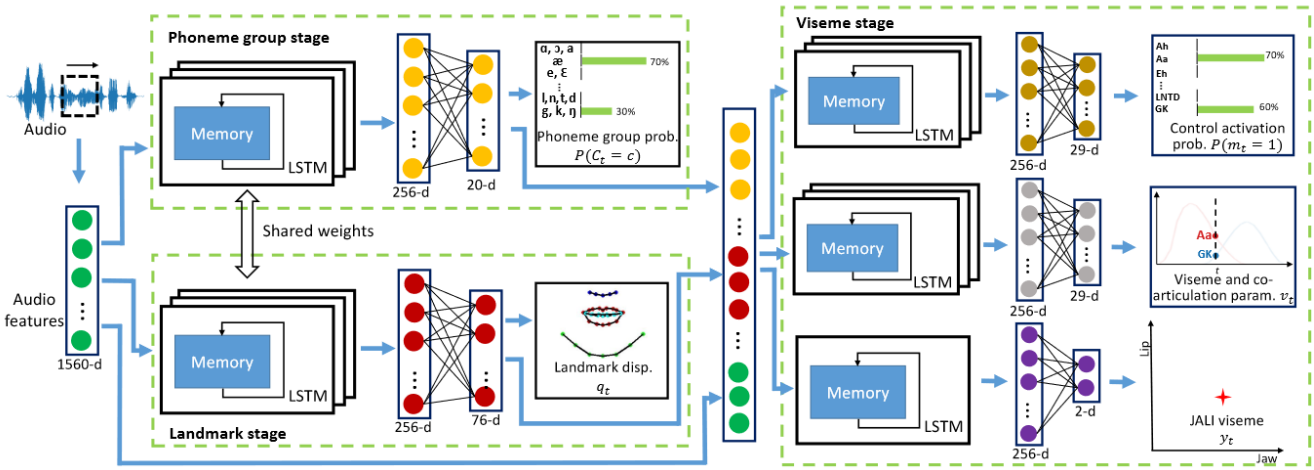

Figure 2. Architecture of Visemenet model.

Method	Input	Output
JALI	Audio, transcript	JALI rigged keyframed lip-sync animation
Visemenet	Audio	JALI rigged lip-sync animation
MeshTalk	Audio, Expression mesh animation	Mesh animation
D3DExpression	??	??

Table 1. Comparison of input and output of each method

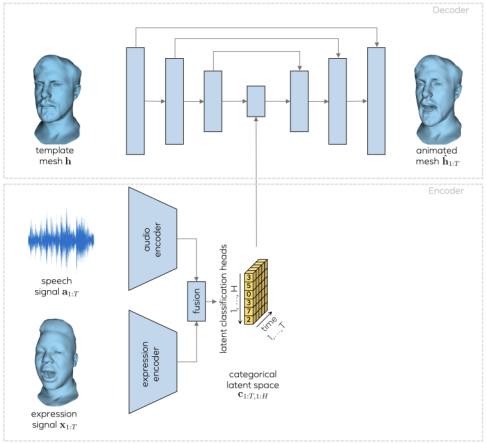

Figure 3. The network diagram [6].

[3] David Hanson. Upending the Uncanny Valley. page 8. 1

[4] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017. 2

[5] Rolandos Alexandros Potamias, Jiali Zheng, Stylianos Ploumpis, Giorgos Bouritsas, Evangelos Ververas, and Stefanos Zafeiriou. Learning to Generate Customized Dynamic 3D Facial Expressions. *arXiv:2007.09805 [cs]*, July 2020. 1

[6] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. MeshTalk: 3D Face Animation From Speech Using Cross-Modality Disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021. 1, 3

[7] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics*, 37(4):161:1–161:10, July 2018. 1

Method	Types	Data(in time)
JALI	No training data	-
Visemenet	Publically available Audio, Transcript, Video	15h pretrain 1h main
MeshTalk	In-house Audio, Mesh animation	13h
D3DExpression	??	??

Table 2. Comparison of traning dataset of each method