Generating Expressive Facial Mesh Animation : A Survey

1. Introduction

Human tend to be very sensitive to facial motion psychologically. Slightest uncanniest in facial animation directly leads to hurt overall experience [4]. 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, We introduce and compare three research that animate expressive facial animation :

- JALI [2] and VisimeNet [8], a linguistic approach to lip-sync.
- MeshTalk [7], a deep learning method.
- D3DExpression [6], LSTM method which replicate facial expression.

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 [3]. 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 [2] 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.

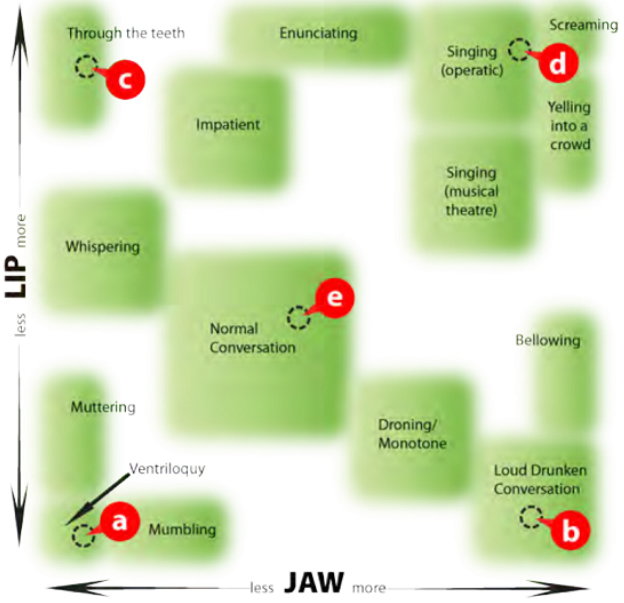

Figure 1. Speaking styles of JALI viseme field

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 and a expression mesh. Since speech does not encode all aspects of the facial expressions. Eye-blinks are a simple example of uncorrelated expressive information. Most existing audio-driven approaches static upper face animation. To overcome this issue, Meshtalk introduces a method that learn a categorical latent space for facial expressions.

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 [5] over every classification head.

Application of Meshtalks are re-targeting animated mesh to other static mesh, and re-synthesizeing sentences with a new language audio snippet.

In traning phase, the method uses In-house audio-visual dataset. At dataset, 1.4 million frames of 13 hours equivalent audio-visual dataset from 250 subjects, reading 50 phonetically balanced sentences.

A novel categorical latent space in combination with a cross-modality loss enables generation of highly realistic animation. Meshtalk demostrates high-accuracy lip mition and plausible motion of eye and eyebrows.

The limitation of Meshtalk is that it cannot compenstate for hidden body parts. For example, hair covering eyebrow. And network size is moderelty large and cannot be run realtime at low-cost computational devices such as mobile phone or VR devices.

2.3. D3DExpression

The goal of paper *Learning to Generate Customized Dynamic 3D Facial Expressions*(D3DExpression) is to generate realistic 3D facial animation given a target expression and a static neutral face.

Dataset is 4DFAB [1], which includes video clips with highly exaggerated six facial expressions, namely happy, sad, surprise, angry, disgust and fear, for 180 subjects. It is assumed that each expression can be characterised by four phases of its evolution as shown on 4. Starting from Netu-

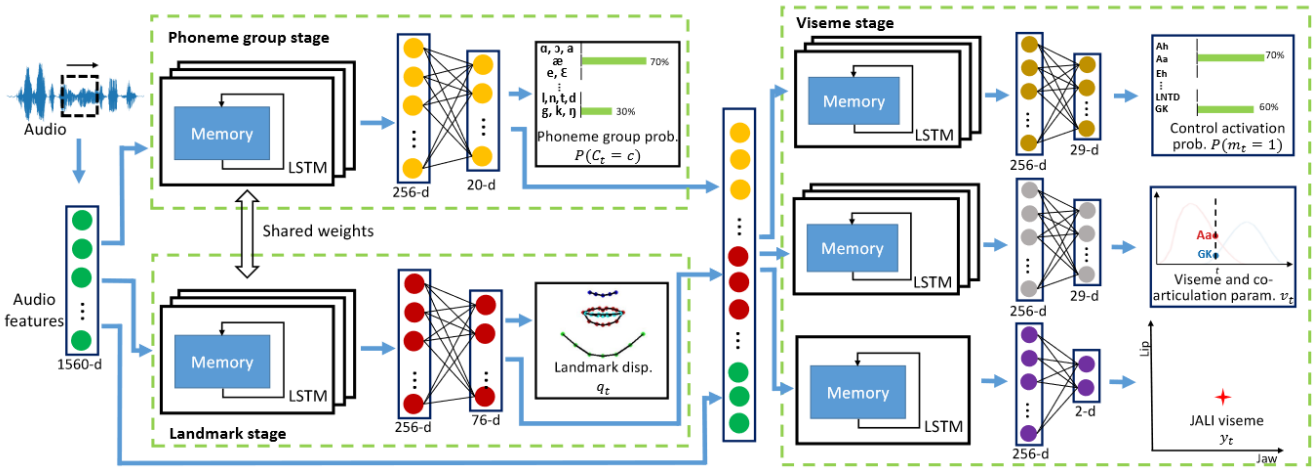

Figure 2. Architecture of Visemenet model.

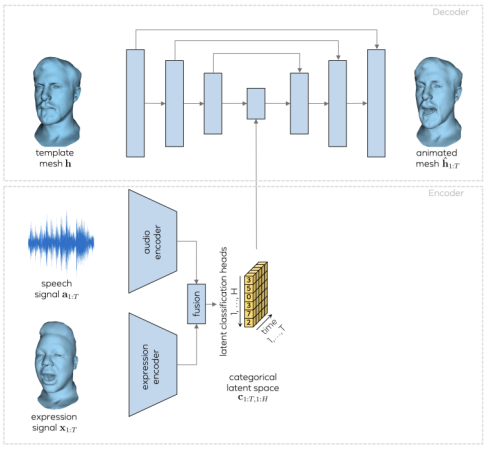

Figure 3. The network diagram [7].

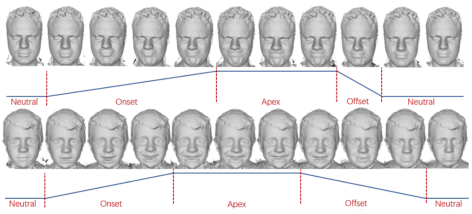

Figure 4. Four phases of emotion, [6]

ral value of 0, onset that linerlly increment value, then apex value of 1, offset that decreses then back to netural face.

The overall architecture of the model is structured by two major components. The first one contains a temporal encoder, using an LSTM layer that encodes the expected facial motion of the target expression. Input of the LSTM network is 6 dimension, which is a one-hot encoding of one of the six exprssion, with amplitude. The second component is a frame decoder, with four layers of mesh convolutions, where each one is followed by an upsampling layer. Each upsampling layer increases the number of vertices by five times, and every mesh convolution is followed by a ReLU activation. Finally, the output of the decode is added to the neutral face.

The model can predict exaggerated facial mesh, and interpolate between diffrent expressions. Results show that the proposed method outperforms expression blendshapes. But the dataset is limited to extreme variation of the expression and limited to six emotions.

2.4. Discussion

We went through methods that generates rich facial animation. JALI takes phonemic approach to generate realistic animator centric lip-sync animation. Visemenet took this approach, an replaced phonemic component with deep learning layers, further automating the process. MeshTalk suggests novel loss function and categorical latent space to encourage the model to have an accurate upper and lower face reconstruction. D3DExpression proposed a method that can recreate facial expression animation.

Three of the paper targets lip-sync animation as a goal and shows high quality results (see Tab. 1). Lip-sync is well understood field, but expressive animation of upper face like eyes and eyebrows is relatively unexplored, yet shows potential.

Meshtalk and D3DExpression generates mesh animation. Yet rendering mesh requires various texture and maps, like bump-map for wrinkles. Further research can be conducted to generate such textures.

Method	Input	Output
JALI	Audio, transcript	JALI rigged keyframed lip-sync animation
Visemenet	Audio	JALI rigged lip-sync animation
MeshTalk	Audio, Expression mesh animation, template mesh	Mesh animation
D3DExpression	Expression encoding, template mesh	Mesh animation with given expression

Table 1. Comparison of input and output of each method

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	Expression labeled mesh animation	1000+ clips

Table 2. Comparison of traning dataset of each method

References

[1] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4DFAB: A Large Scale 4D Facial Expression Database for Biometric Applications. *arXiv:1712.01443 [cs]*, June 2018. 2

[2] 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

[3] 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

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

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

[6] 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, 3

[7] 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

[8] 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