Expressive Talking Avatars

Category: Research

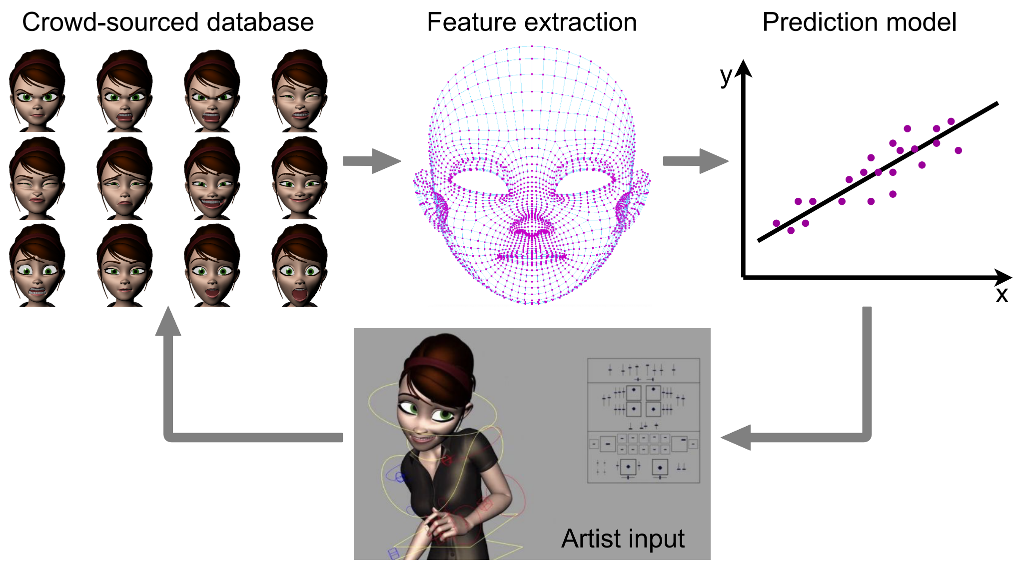


Figure 1: Expressive Talking Avatars: a New Dataset and a New Method

Abstract

Stylized avatars are common virtual representations used in VR to support interaction and communication between remote collaborators. However, explicit expressions are notoriously difficult to create, mainly because most current methods rely on geometric markers and features modeled for human faces, not stylized avatar faces. To cope with the challenge of emotional and expressive generating talking avatars, we build the Emotional Talking Avatar Dataset which is a talking-face video corpus featuring 6 different stylized characters talking with 7 different emotions. Together with the dataset, we also release an emotional talking avatar generation method which enables the manipulation of emotion. We validated the effectiveness of our dataset and our method in generating audio based puppetry examples, including comparisons to state-of-the-art techniques and a user study. Finally, various applications of this method are discussed in the context of animating avatars in VR.

Index Terms: Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies; Human-centered computing—Computer graphics—Graphics systems and interfaces—Virtual reality

1 Introduction

Immersive virtual reality (VR) has been identified as one of the key interaction technologies of the future metaverse, where users use stylized avatars to represent themselves [28]. However, VR head-mounted displays (HMDs) usually occlude a large part of the user's face [18], which rules out most existing video-driven facial animation methods. Although existing audio-driven talking avatar techniques have pushed the boundary of HMD-based avatar mediated communication forward, the following problems remain unresolved. State-of-the-art methods are able to generate lip movements in perfect synchronization with the audio speech [29], but existing methods often neglect facial emotion. However, facial expressions serve as the primary nonverbal means of communication among human beings. A few researchers focus on the emotional talking face generation methods that enables the manipulation of the emotion and its intensity; however, these methods are ~~not~~ developed for human faces [31, 34], not stylized character faces.

A considerable number of recent advancements in the task of talking avatar generation are deep learning based methods [34], where data has a significant influence on performance. We argue that the absence of a high-quality dataset for 3D rigs is the main obstacle to achieve vivid talking avatar generation. As reviewed in the next section, the available datasets are mainly developed for human faces, rather than stylized characters. Aneja et al. developed stylized character datasets with cardinal expression annotations [1, 2]. However, their dataset only includes labeled facial expressions, but not audio-visual animation clips with mapped lip movements with audio. To address this issue, we build the emotional talking avatar dataset, a talking avatar animations featuring six avatars talking with seven different emotions. We carefully selected emotionally consistent speech texts that cover different phonemes, from existing audio-visual human dataset. To ensure the expressiveness and naturalness of the performed emotions, our inhouse professional animator created animation clips in seven categories, including, anger, disgust, fear, joy, neutral, sadness and surprise.

Together with the dataset, we presented the emotional talking avatars specially designed for 3D stylized characters in a geometrically consistent and perceptually valid way. Our method begins by utilizing a pre-trained HuBERT model to extract features, facilitating the precise generation of rig parameters that control the mouth region's movements. Another emotion branch is introduced for generating emotional displacement to the emotionless lip motions. Through a straightforward fusion process, we create emotionally responsive talking heads that react to both audio input and emotional context.

We validate our dataset and our method by measuring emotional recogonition, intensity, synchronization, & naturalness, which are crucial factors for audience engagement. Our dataset were identified with 72% accuracy, which is comparable to the mean accuracy rates for the human condition (videos taken from MEAD [34] with three intensity levels and RAVDESS [20] with 'normal' and 'strong' intensity level) at 73.9%. Results also show that our method maintained the same level of intensity lip sync quality, & naturalness compared to the human condition. Additionally, results also reveal our talking head generation method significantly improved scores of the lip sync quality & naturalness, while maintaining the same level of expression recognition & intensity compared to SOTA. This research makes several contributions, as follows:

- We build a high-quality emotional talking avatar dataset, which is the first emotional audio-visual corpus annotated for 3D rigs of stylized characters.
- We propose an emotional talking avatar generation method that enables the manipulation of the emotion.
- Extensive experiments validated dataset, and our expressive avatar generation methods could be apply to future reference in AR/VR/XR.

2 Related work

2.1 3D Avatars in Virtual Reality

Avatars are digital representations of users that may be abstract, cartoonish, or human-like [38]. These avatars may represent a user in a video call or a character in a video game. Depending on the use case, avatars might be created manually by designers and character artists, e.g. for characters in animated movies [17], or use computer vision or other techniques to generate a digital 2D or 3D avatar in the likeness of a user's visual appearance [22, 23]. Different implementations also allow for different levels of customization of the avatar, such as facial appearance, body type, and clothing and accessories. Avatars can also be animated to allow users to express emotions and help users self-identify with the avatar. This is usually achieved by rigging a skeletal system [8] which is moved by sensing and tracking the user. This can be achieved with cameras and computer vision [5], speech sentiment analysis, or animating visemes through audio input.

We focus our efforts on stylized 3D characters, defined as characters that no human would mistake for another person, but would still be perceived as having human emotions and thought processes. Our goal is to develop an approach for driving stylized character expressions only using sensors from commodity VR hardware as they contain microphones for audio input.

2.2 Audio Driven Facial Animation

Audiovisual speech animation can be categorized into procedural, performance-capture, data-driven, and deep-learning techniques [24, 40]. Procedural speech animation uses visemes, the oral shape at the apex of a given phoneme, to solve overlaps between successive visemes, curve shapes that define attach, apex, sustain, and decay of viseme, and mapping a given phoneme to a viseme [33]. Performance-capture speech animation is visually limited due to the actor's performance. It is visually limited by the actor's performance and is hard for an animator to edit or refine [19]. Data-driven approaches smoothly stitch pieces of facial animation data from a large corpus, to match an input speech track using morphable etc [11]. These data-driven methods tend to be limited in scope to the data available, and the output, similar to performance-capture, is not animator-centric.

Recent research has shown the potential of deep learning to provide a compelling solution to automatic lip-synchronization simply using an audio signal [3, 31, 35]. Current deep learning based methods can be broadly categorized into 2D methods and 3D methods. For 2D methods, Zhou et al. present a new deep-learning based architecture to predict facial landmarks, capturing both facial expressions and overall head poses, from only speech signals [39]. When it comes to 3D methods which is more related to our task, Taylor et al. uses phonemes as an intermediate speech signal to drive coefficients of an Active Appearance Model, representing the lower face and jaw of a head model [32]. Edwards et al. have introduced the JALI model to simulate different speech styles controlling the jaw and lip parameters in a two-dimensional viseme space [10]. Zhou et al. predicts sparsely activated viseme- and co-articulation parameters for a FACS-rig from speech, using both phonemes and raw speech features [40].

In particular, Chen et al. have developed a TTS data augmentation method in talking face tasks by producing augmented audio-animation pairs with a TTS system, and solved the misalignment problem brought by TTS audio with the introduction of soft-DTW loss [6]. The weighted sum of HuBERT features is adopted to fully utilize the underlying information of audio. Their method is proven to boost the few-shot ability of a talking face system in low data resources [6].

Inspired by the previous work developed for human faces [6], we also utilized the pre-trained HuBERT model [13] as a feature extractor, showcasing consistently superior performance across diverse downstream tasks compared to traditional handcrafted acoustic features like MFCC. This leads to improved lip-synchronization. We further implemented the Soft-DTW loss to effectively address potential data misalignment issues.

2.3 Emotion and Expressiveness

Emotion is a factor that plays a strong role in realistic animation. Only a few works consider it in talking face generation due to the difficulty of producing emotion dynamics. Karras et al. animate 3D vertices of a face given a speech signal by utilizing an end-to-end deep network consisting of a formant analysis network, an articulation network, and a learned emotion embedding [16]. Wang et al. collect the MEAD dataset and propose an emotional talking face generation baseline that enables the manipulation of the emotion and its intensity. They designed a two-branch architecture to process the audio and emotional conditions separately. One of the branches is responsible for mapping audio to lip movements and the other branch is responsible for synthesizing the desired emotion on the target face. Finally, the intermediate representations are fused in a refinement network to render the emotional talking-face video [34]. Ji et al. propose to decompose speech in decoupled content and emotion spaces, and then synthesize emotion dynamics from audio [15]. Ji et al. propose the Emotion-Aware Motion Model (EAMM) to generate one-shot emotional talking faces by involving an emotion source video, and their methods can generate satisfactory talking face results on arbitrary subjects with realistic emotion patterns [14]. Gururan et al. present SPACE, which uses speech and a single image to generate high-resolution, and expressive videos with realistic head pose, without requiring a driving video [12]. These methods animate 3D models of faces such as standard FACS-based photorealistic avatars, whereas others directly animate raw images of human faces, but not developed for 3D stylized characters.

Though a reliable parameterization of emotion and expression remains elusive, the six cardinal expressions pervade stories and face to face interactions, making them a suitable focus for educators and facial expression researchers. To guide and automate the process of expression animation, animators and researchers turn to FACS. For example, FACSGen allows researchers to control action units on realistic 3D synthetic faces. However, The strict use of anatomy-based and constrained motion by these systems limits their generalizability to characters with different anatomy and limits their application, because the most believable animation may require the violation of physical laws.

Aneja et al. first looked into learning 3D stylized character expressions from humans in a perceptually-valid and geometrically-consistent manner. They created a database FERG-DB of labeled facial expressions for six stylized characters. This database with expressions is created in collaboration with facial expression artists and initially labeled via Mechanical Turk (MT). Then, they built two systems, DeepExpr, and ExprGen. DeepExpr takes a 2D image of a human, and retrieves the matching stylized character expression image, and ExprGen takes a 2D image of a human, and generates the 3D rig parameters of a stylized character.

Built on the FERG-DB database, Pan et al. contributed a performance-based real-time method to animate believable and accurate facial expressions of 3D stylized characters [26]. Pan et al. also introduced an audio-based facial animation approach to portray characters in a geometrically consistent and perceptually correct way. The lips motion and the surrounding facial areas are controlled by the contents of the audio, and the facial dynamics are established by category of the emotion and the intensity [25]. However, FERG-DB database is labeled facial expressions of stylized characters, and thus difficult to ensure accurate lip synchronization, while generating realistic motions for the entire face with vivid emotional changes.

To tackle issues discussed above, we contribute a novel emotional audio-visual dataset for stylized character rigs dataset with cardinal expression annotations, and then a novel method to accurately retrieve plausible character expressions from audio input only.

3 Dataset

3.1 Design Criteria

Emotion Categories We use seven emotion categories following Aneja et al. [1, 2] (anger, disgust, fear, joy, sadness, surprise & neutral), since there is agreement on their recognition within the facial expression research community, and these seven expressions occur in a wide range of intensities and can blend with each other to create additional expressions.

Design of the Speech Corpus For audio speech content, we follow the MEAD [34] and RAVDESS [20], which are a talking-face video corpus featuring actors and actresses talking with different emotions. We carefully select the sentences covering all phonemes in each emotion category, and the sentences in each emotion category are divided into two parts: 4 common sentences, and 7 emotion-related sentences. We provide more details of the speech corpus in Table 2.3.

4 Method

4.1 Overview

We propose a facial animation method, which involves generating precise lip movements and expressions based on the input audio and emotion category. Figure 3 shows an overview of the system, which consists of three essential components. Specifically, we first employ a pre-trained HuBERT model [13] to extract HuBERT features and introduce a Mou Decoder to generate accurate rig parameters related to mouth region. Next, we adopt an Audio Encoder and an Emotion Encoder to extract audio features from MFCC and emotion embedding, respectively. Then the features are combined and fed in to the designed Emo decoder to predict emotional parameters, which are further fused with lip parameters to control the 3D stylized characters' emotions as they speak along with the input audio. Lastly, the vivid animations are produced by rendering process in Maya.

4.2 Architecture

Within the network architecture, there are five sub-networks: a pre-trained HuBERT encoder E_h, a mouth decoder D_m, an audio encoder E_a, an emo encoder E_e and a emotion decoder denoted as D_e.

HuBERT Encoder To make full use of the information contained in the audio, we adopt a pre-trained HuBERT model to extract features. Instead of directly taking the final embedding as the subsequent input [36], we predict N hidden layers, which are weighted summed to feed into the Mon Decoder. The obtained HuBERT feature f_h can be represented as:

$$f_h = \sum_{i=1}^{N} (\alpha_i h_i) \quad , \quad \sum_{i=1}^{N} \alpha_i = 1 \tag{1}$$

where h_i and α_i denote the ith hidden layer and corresponding weight.

Mouth Decoder The Mouth Decoder consists of 2-layer 1D convolutional neural network, a 2-layer BiLSTM Network. The former is responsible for downsample the extracted HuBERT feature from 50Hz to 25Hz, while the latter is capable of decoding the feature into meaningful latent representation, which are further to predict the rig parameter sequence of the lip region.

Audio Encoder Considering expression are more correlated to the rhythm and beat instead of phonemes, we extract the Mel-Frequency Cepstral Coefficients (MFCC) [21] aspect from the provided input audio signal, while pairing the video frames and audio signal using a one-second temporal sliding window. Both the audio frame sample rate and video frame rate are set at 25. Subsequently, we apply the audio encoder, which comprises convolutional neural networks (CNN) followed by multi-layer perceptrons (MLP), to process the 28×12-dimensional audio features as input and obtain the desired audio feature.

Emotion Encoder We first encode the emotion label as a one-hot vector e and input it into the emotion encoder. The emotion encoder utilizes a two-layer fully connected (FC) neural network followed by a LeakyReLU activation to map the one-hot vector to an emotion embedding. This embedding is duplicated for each time step.

Emotion Decoder Based on the audio temporal properties, we design the emotion decoder with a long short-term memory (LSTM) network and a fully connected layer to map from the extracted audio feature and emotion embeddings to the rig parameters. The LSTM in our model is composed of three layers with 60 nodes and 100 time steps. In this way, the sequential relationship between audio signals and rig parameters can be better captured.

4.3 Objective Functions

Formulaically, given an audio $a = \left\{a^{(1)}, \ldots, a^{(T)}\right\}$ and the input emotion condition e, we are able to generate the predicted mouth rig parameters $\hat{y_m} = \left\{\hat{y_m}^{(1)}, \ldots, \hat{y_m}^{(T)}\right\}$ and expression rig parameters $\hat{y_e} = \left\{\hat{y_e}^{(1)}, \ldots, \hat{y_e}^{(T)}\right\}$, respectively:

$$\left[\hat{y_m}^{(t)}, h^{(t)}, c^{(t)}\right] = D_m(E_h(a^{(t)}), \hat{y_m}^{(t-1)}, h^{(t-1)}, c^{(t-1)}), \tag{2}$$

$$\left[\hat{y_e}^{(t)}, h^{(t)}, c^{(t)}\right] = D_e(E_a(a^{(t)}), E_e(e), \hat{y_e}^{(t-1)}, h^{(t-1)}, c^{(t-1)}), \tag{3}$$

where $h^{(t)}, c^{(t)}$ represent hidden state and cell state of LSTM unit at time t respectively, and T refers to the frames of the video. Then, we fuse the predicted parameters to obtain the final results $\hat{y}^{(t)}$:

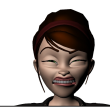
$$\hat{y}^{(t)} = F(\hat{y_m}^{(t)}, \hat{y_e}^{(t)}), \tag{4}$$

Table 1: Description and examples for six different expressions.

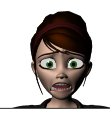
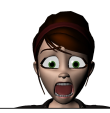
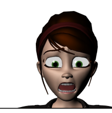
Angry

Item/Part	Eyebrow	Eye	Mouth	Others
Description	Inner corner pulled downward; Lower edge of eyebrow falls;	Eyes widened; The wider, the angrier; Lower lid tight.	Upper lip is lifted in a sneer; Square shape with lots of teeth;	Upper eyelid rises; Lower lip margin straight; Both upper and lower teeth shown;
Example				

Disgusted

Item/Part	Eyebrow	Eye	Mouth	Others
Description	Entire eyebrow lowered; Especially inner corner;	Partly squinted ; Further compressed;	Upper lip raised flattened in an intense sneer; squared-off in shape	Nose wings pulled upward; Nose creases deepened; Upper teeth showed
Example				

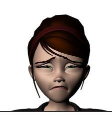
Fear

Item/Part	Eyebrow	Eye	Mouth	Others
Description	Lifted and kinked straight up; Pulled closer together	Opened very wide; Often with taut; Lower lid raised;	Opened and widened; Upper lips tightened; Lower lips stretched	The wider the eyes, the more afraid.; Teeth exposed;
Example				

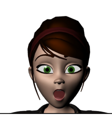
Happy

Item/Part	Eyebrow	Eye	Mouth	Others
Description	Relaxed; May be raised straight up;	upper lids moved downward slightly; Lower lids tightened;	Widened with corner pulled back toward ear; Lips and skin around mouth follow teeth;	Steep edge of cheek; Nose wings pulled upward; Smooth chin;
Example				

Sad

Item/Part	Eyebrow	Eye	Mouth	Others
Description	Entire brow lowered; Lower lids raised; Upper lid pushed;	Reduced to nearly a single line; The thinner, The sadder;	Rectangular Shape; Lips straightened and thinned; Lower lip tucked under upper;	Nose wings raised; Cheeks tight and rounded; Deep nasolabial fold;
Example				

Surprised

Item/Part	Eyebrow	Eye	Mouth	Others
Description	Eyebrows Raised;	Upper eyelid raised; White above iris exposed; Lower lid relaxed	Dropped open; Oval in shape; Protruded lips	Horizontal forehead wrinkles created
Example				

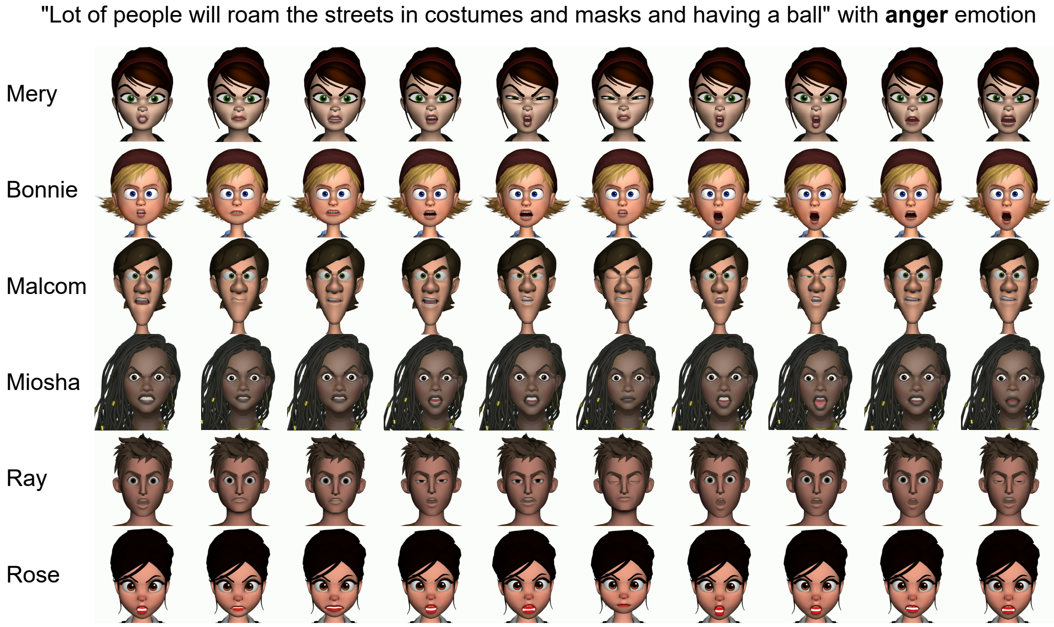


Figure 2: Sampled frames from dataset.

To optimize the network, the L_2 loss function and the Soft-DTW loss [6][1] are established to define the task. Soft-DTW [7] refers to Soft-Dynamic-Time-Warping, which is a variant of the vanilla Dynamic-Time-Warping (DTW) [30]. It addresses the challenge of non-back-propagatable gradients inherent in vanilla DTW. DTW itself aims to optimize the alignment of sequential data and measures their similarity. Recent studies [4, 6, 15] have highlighted DTW's superior alignment capabilities, especially with unequal sequence data. Drawing inspiration from these findings, we integrate Soft-DTW loss into our work to tackle potential synchronization issues between audio and ground-truth rig parameters in our dataset. The prevalence of synchronization challenges in existing datasets, stemming from equipment, transmission, and storage inconsistencies, is amplified in 3D datasets due to animator errors. Therefore, leveraging Soft-DTW loss offers notable advantages in enhancing audio-lip synchronization compared to the conventional MSE loss, which is validated in the following section. Also, we introduce an inter-frame continuity loss [37] to deal with the jitter issue. Given the ground truth rig parameter y, the total loss function $\mathscr{L}$ can be represented as:

$$\mathscr{L} = \mathbb{E}_{a,e,y}\left[(y-\hat{y})^2 + DTW(y,\hat{y})\right] + \lambda_1 \mathbb{E}_{\hat{y}}\left[\sum_{t=0}^{T-1}\left(\hat{y}^{(t+1)} - \hat{y}^{(t)}\right)^2\right] \tag{5}$$

5 Evaluation

5.1 User Study

5.1.1 Participant

We recruited 25 participants from ANONYMOUS University to complete all conditions of this study. 20 participants were assigned to complete task 1 and five additional participants were assigned to complete task 2. The average age of the participants was 22 years, ranging between 19 and 24 years old; 12 were men. They were naïve to the purposes of the experiment.

[1] Implementation refers to https://github.com/Moon0316/T2A

5.1.2 Material

Task 1 To evaluate our dataset, we rendered animation clips from its 3D facial rig in the dataset, which has associated rig control parameters. We divided our dataset into four group (see No. in Table 2.3).

Task 2 To evaluate our method, we randomly selected 75% of dataset to train our model & EVP [25], and used the rest of dataset as ground true. Then, we run our method & EVP method to create 7 animation clips for primary character expression, and apply the Multiple Character Adaptation network [25] transfer the expression on different 5 stylized characters.

5.1.3 Design

Task 1 The experiment utilized 7 characters (Human, Mery, Bonnie, Ray, Malcolm, Rose & Miosha) $\times$ 7 emotions (Neutral, Anger, Sadness, Fear, Disgust, Happiness, & Surprise) $\times$ 2 sentences (Common sentences vs. Emotion-related sentences) $\times$ 4 group of dataset in a mixed design, with a between-subject design for , but a within-subject design regarding characters, emotions, and tracking methods.

Each participant took part in 98 trials to evaluate the human expression and the 6 character expression: 7 $\times$ (7 emotions $\times$ 2 types of sentence) $=$ 98 trials. Thus, there were 1960 trials in total.

Task 2 The experiment utilized 6 characters (Mery, Bonnie, Ray, Malcolm, Rose & Miosha) $\times$ 7 emotions (Neutral, Anger, Sadness, Fear, Disgust, Happiness, & Surprise) $\times$ 3 methods in a within-subject design regarding characters, emotions, and methods.

Each participant took part in 126 trials to evaluate the generated primary character expression, and the expression transfer results on different 5 stylized characters 6 $\times$ (7 emotions $\times$ 3 capturing methods) $=$ 126 trials. Thus, there were 630 trials in total.

To avoid fatigue or carry-over effects, video clips were presented to the participants in random order.

5.1.4 Procedure

Participants were first presented with an information sheet and asked to sign a corresponding consent form. They were instructed to view

Table 2: Designed speech corpus.

		Common sentences read in 7 emotions
Emotion	NO.	Speech Corpus
All	1	Kids are talking by the door
	2	Dogs are sitting by the door
	3	She had your dark suit in greasy wash water all year
	4	Don‘t ask me to carry an oily rag like that
		Emotion-related sentences
Emotion	NO.	Speech Corpus
Angry	1	Right now may not be the best time for business mergers
	2	You’re so preoccupied that you’ve let your faith grow dim
	3	Lot of people will roam the streets in costumes and masks and having a ball
	4	Then he would realize they were really things that only he himself could think
Disgust	1	Please take this dirty table cloth to the cleaners for me
	2	Young children should avoid exposure to contagious diseases
	3	You’re not living up to your own principles she told my discouraged people
	4	Pretty soon a woman came along carrying a folded umbrella as a walking stick
Fear	1	Call an ambulance for medical assistance
	2	The fish began to leap frantically on the surface of the small lake
	3	We will achieve a more vivid sense of what it is by realizing what it is
	4	This is a problem that goes considerably beyond questions of salary and tenure
Happy	1	The eastern coast is a place for pure pleasure and excitement
	2	By that time perhaps something better can be done
	3	Obviously the bridal pair has many adjustments to make to their new situation
	4	His artistic accomplishments guaranteed him entry into any social gathering
Sad	1	There’s no chance now of all of us getting away
	2	The diagnosis was discouraging however he was not overly worried
	3	The prospect of cutting back spending is an unpleasant one for any governor
	4	But the ships are very slow now and we don’t get so many sailors any more
Surprise	1	He ate four extra eggs for breakfast
	2	I just saw Jim near the new archeological museum
	3	He further proposed grants of an unspecified sum for experimental Hospitals
	4	Properly used the present book is an excellent instrument of enlightenment
Neutral	1	The best way to learn is to solve extra problems
	2	As such it was beyond politics and had no need of justification by a message
	3	Keep your seats boys I just want to put some finishing touches on this thing
	4	Bridges tunnels and ferries are the most common methods of river crossings

an animation clip and then asked to answer three questions:

- “Which expression did the character depict?” Participants were asked to select one of the words: Neutral, Anger, Sadness, Fear, Disgust, Happiness, Surprise or Other.
- “How intense was the indicated emotion depicted by the character?” Participants rated the intensity on a scale from 1 to 7, where 1 represents a rating of “Not at all”, and 7 represents “Extremely”.
- “Whether the lip motion sync with the speech?” Participants rated on lip sync qualities on a scale from 1 to 7, where 1 is not synchronized at all & 7 is synchronized extremely well.
- “How natural was the character overall?” Participants rated attractiveness on a scale from 1 to 7, where 1 represents a rating of “Not at all”, and 7 represents “Extremely”.

Each participant undertook one practice trial where they could ask questions, and then undertook measured trials.

The participants were paid ANONYMOUS amount. The experiment took about 30 minutes. The experiment was approved by ANONYMOUS University Research Ethics Committee.

5.1.5 Results on datasets

For the statistical analysis, we conducted separate repeated measures Analysis of Variances (ANOVAs) for videos, looking at the results on recognition, intensity, synchronization, and naturalness. There were no outliers, and the data was normally distributed for each condition as assessed by boxplot and Shapiro–Wilk test $(p > 0.05)$, respectively. We ran Mauchly’s test for validating sphericity of the data, and whenever it is significant, we report results with Greenhouse-Geisser correction applied and marked with an asterisk “*”. Post hoc tests were conducted using the Bonferroni test for the comparison of means.

Recognition For the recognition of emotions, responses were converted to scores “1” (correct) or “0” (incorrect) and averaged over stimuli repetitions.

Figure 5(a) shows that anger ($M = .921, SE = .014$), joy ($M = .925, SE = .019$), neutral ($M = .886, SE = .027$), and sadness

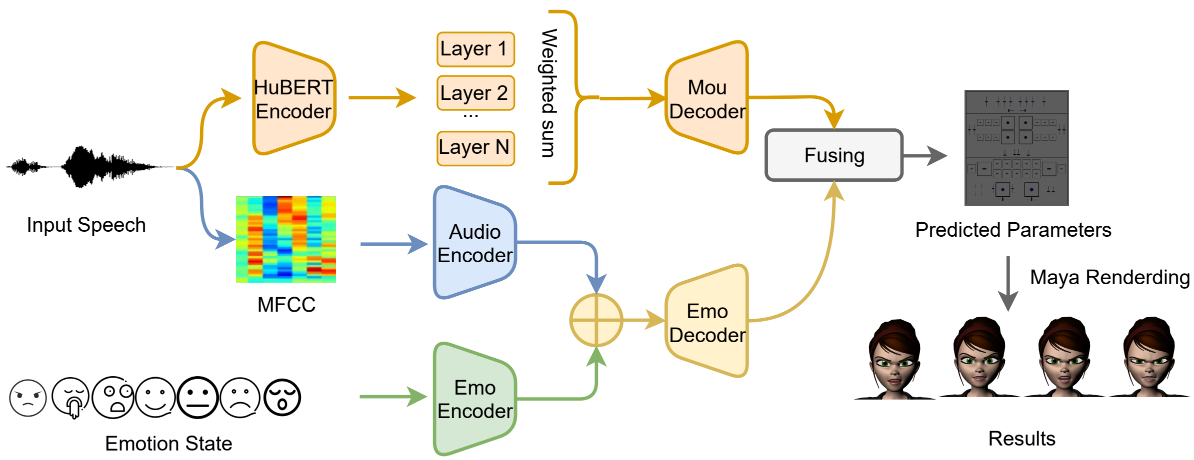
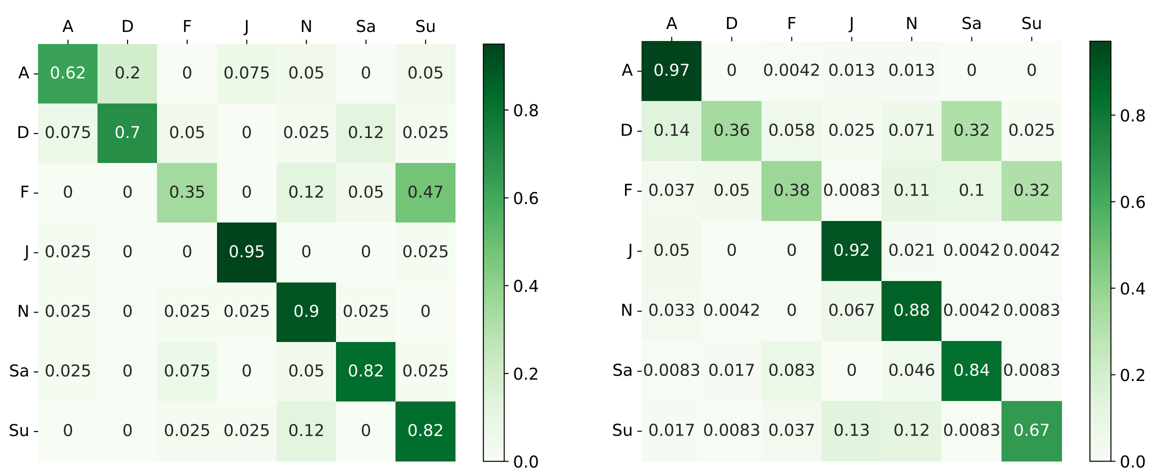


Figure 3: The overview of our system.



(a) Human (b) Characters

Figure 4: Confusion matrix for perceived expression recognition (%) for seven expression classes. A = anger, D = disgust, F = fear, J = joy, N = neutral, Sa = sadness, Su = surprise.

($M = .836, SE = .034$) show high accuracy, whereas disgust ($M = .407, SE = .041$) and fear ($M = .375, SE = .05$) are very difficult for users to perceive. We also observed the character expression recognition accuracy could occasionally be higher than humans, this might be because the characters have simpler geometry and stylization can make the expressions relatively easier to perceive.

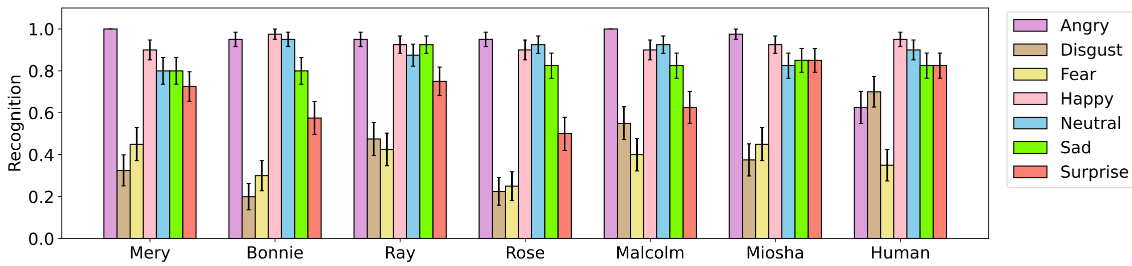
We compared average scores obtained for 7 emotions across 7 characters. We found the main effect of the character was significant, $F(6,114) = 3.215, p = .006$, and the characters $\times$ emotions interaction was also significant, $F(36,684) = 3.957, p < .001$. This indicates the recognition score for different characters present differently for different emotions.

Our initial results on expression recognition shows that the main effect of emotions was significant, $F(3.79,72.003) = 53.139, p < .001^*$. Thus, we look into participants' rating for seven expression classes. Figure 4 shows the confusion matrix for perceived expression recognition for each expression class. In each sub-figure, for a given row (e.g. anger), the columns represent the percentage (averaged over all the perceived human anger expressions) of participants agreeing on the corresponding expression classes.

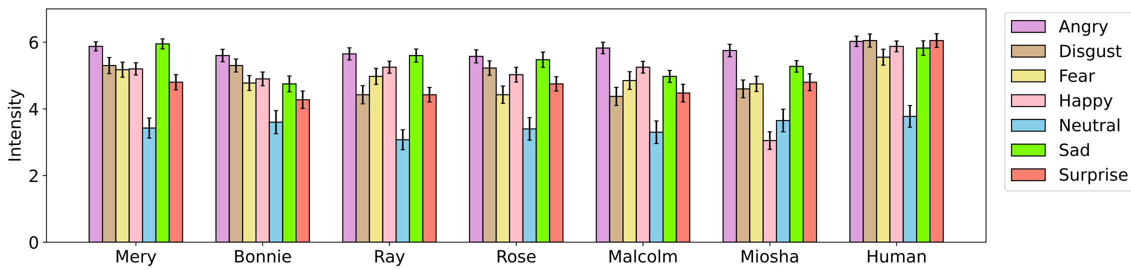
Intensity Intensity ratings for our characters were high in general, which is expected for exaggerated cartoon animation. Figure 5(b) shows the mean intensity ratings for 7 emotions across 7 characters. The average score over all characters for neutral ($M = 3.461, SE = .398$) is significantly lower than the average score for the rest of emotions.

We found the main effect of character, the main effect of emotion, and characters $\times$ emotions interaction were all significant, $F(1.963,37.288) = 8.515, p < .001^*$, $F(1.874,35.606) = 27.03, p < .001^*$, and $F(36,684) = 3.386, p < .001$, respectively. This indicates the intensity score for different characters present differently for different emotions.

Synchronization & Naturalness We look at the rating on synchronization & naturalness for 7 emotions across 7 characters conditions. However, no statistically significant effects were found, in terms of characters, emotions and the characters $\times$ emotions interaction, thus we did not include the figure results for synchronization & naturalness.



(a) Recognition



(b) Intensity

Figure 5: Results on datasets

5.1.6 Results on our methods

Recognition Figure 7(a) shows the comparison of average recognition scores obtained for three generation methods across six stylized characters. The average score over all characters for GT ($M = .786, SE = .046$) and our methods ($M = .743, SE = .033$) are slightly higher than the average score for the EVP ($M = .619, SE = .03$). We found the main effect of the generation method was significant, $F(2,8) = 5.056, p = .038$. However, Bonferroni post-hoc comparisons did not find any significant effect among these conditions.

Intensity Figure 7(b) shows the comparison of average intensity scores obtained for three generation methods across six stylized characters. The average score over all characters for GT ($M = 5.171, SE = .603$), our methods ($M = 5.09, SE = .62$), and EVP ($M = 5.133, SE = .668$), respectively. Results revealed the main effect of methods on intensity scores was not statistically significant, $F(1.021, 4.085) = .056, p = .83^*$.

Synchronization Figure 7(c) shows the mean synchronization ratings for three generation methods across six characters. The average score over all characters for GT ($M = 5.024, SE = .703$), and our methods ($M = 4.133, SE = .633$) are significantly higher than the average score for EVP ($M = 1.929, SE = .347$).

The main effect of the generation method was significant, $F(2,8) = 16.531, p = .001$. Bonferroni post-hoc comparisons indicated the mean synchronization ratings for GT and our methods are significantly higher than EVP, $p = .038$ and $p = .023$, respectively.

Naturalness Figure 7(d) shows the mean naturalness ratings for three generation methods across six characters. The average score over all characters for GT ($M = 4.983, SE = .681$), and our methods ($M = 3.805, SE = .672$) are significantly higher than the average score for EVP ($M = 2.186, SE = .411$).

The main effect of the generation method was significant, $F(2,22) = 15.077, p = .002$. Bonferroni post-hoc comparisons indicated the mean naturalness ratings for GT and our methods are significantly higher than EVP, $p = .035$ and $p = .028$, respectively.

Table 3: Quantitative results for comparison & ablation study.

Metric/Method	EVP [25]	w/o Soft-DTW	w/o Emo Encoder	Ours
F-RPD	0.179	0.057	0.089	**0.014**
M-RPD	0.094	0.029	0.037	**0.005**

5.2 Comparison & Ablation Study

We introduce two new metric, facial rig parameter distance (F-RPD) and mouth rig parameter distance (M-RPD), which calculates the distance between facial and mouth rig parameters generated by different methods (EVP, w/o Soft-DTW, w/o Emo Encoder and Ours) and ground truth (GT). EVP [25] tackles the same task with our method. For ablation study, we replace Soft-DTW loss with MSE loss, denoted as w/o Soft-DTW. w/o Emo Encoder refers to our framework without Emo Encoder, which generates neutral talking faces without any expressions. The results in Table 3 illustrate the Soft-DTW loss and the emotion branch in our framework plays a great role in performing lip-sync and expressions, respectively.

We select some frames from the videos (left: sad emotion; right: surprised emotion) generated by ground-truth (GT), EVP [25], w/o emo encoder(ablation study) and our method. Specifically, EVP predicts wrong lip motions and w/o emo encoder fails to generate vivid expressions. Our method not only produces accurate mouth shapes, but also presents the similar expressions as GT.

6 Discussion

6.1 Comparison to the State-of-the-art

Although MEAD [34] and EVP [25] tackle the similar task to our method, we would like to emphasize our major differences and superiority. Firstly, MEAD [34] consists of three sub-networks: audio-to-landmarks module, neutral-to-emotion transformer and refinement network. Such a three-stage structure is complex and consumes more computational resources, in addition to the discontinuity between the generated video frames. In contrast, we generate the rig parameters directly from the audio and emotion label, which is much more concise. MEAD is used in the realm of real human

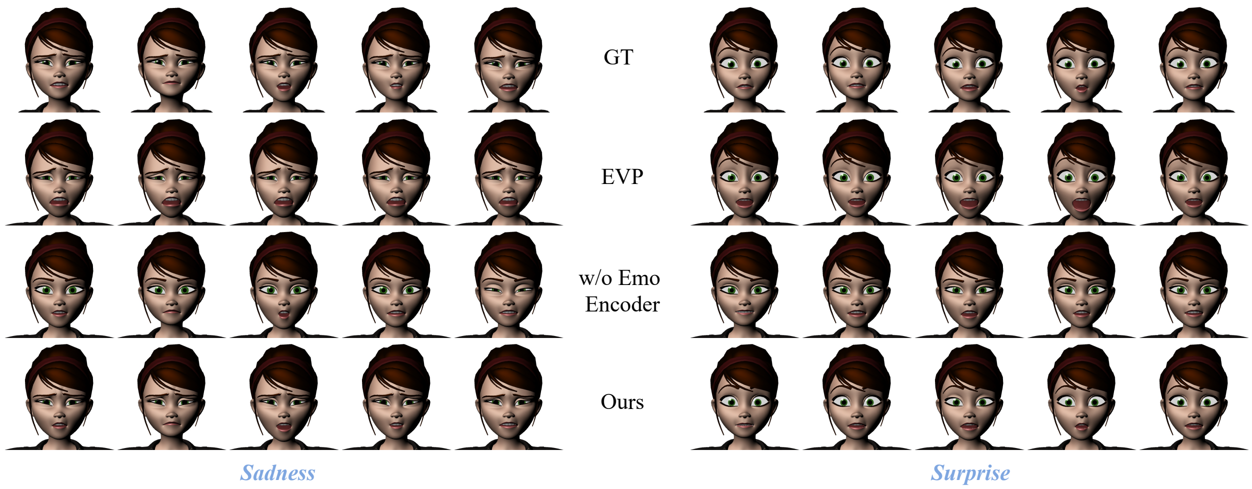


Figure 6: Sampled frames from ground-truth (GT), EVP [25], w/o emo encoder and our method.

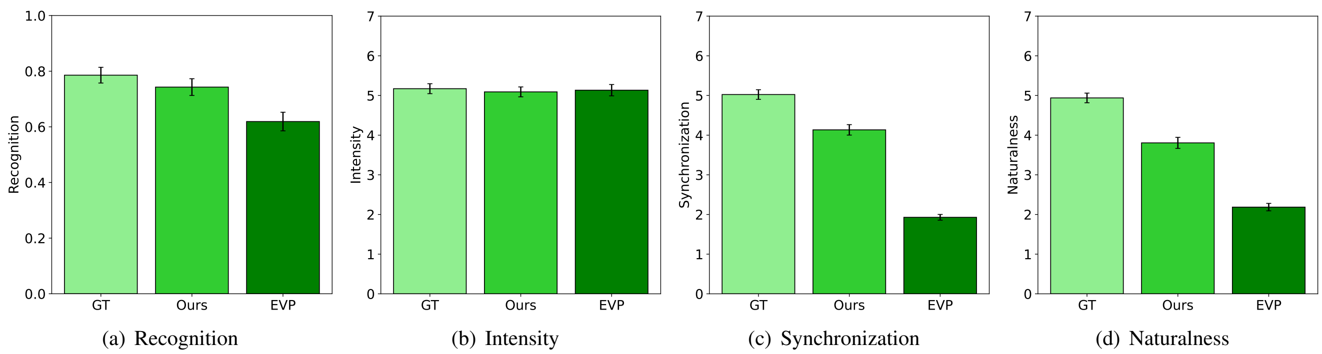


Figure 7: Results on our methods

animation, not stylized characters. EVP [25] is used for stylized characters, which utilizes the audio-to-landmarks module in MAED to predict landmarks which are then matched with the landmarks in the FERG-3D-DB dataset [1], but there are two main problems: a) there are huge domain gaps between the landmarks in such two different datasets, which leads to inaccuracies and error accumulation. b) each frame needs to be be found one by one in more than 10,000 images contained in FERG-3D-DB, which is very time-consuming. By comparison, we utilize our published dataset to perform end-to-end prediction on rig parameters via a simple network, which greatly speeds up the prediction generation efficiency.

6.2 Audio, Video & Text

Talking avatar generation is a typical multi-modal task involving the creation of videos featuring characters speaking, driven by an audio clip, a text script or a video sequence. In this paper, we introduce an audio-visual dataset and present a system tailored for generating talking avatars driven by audio input. Importantly, our system can be easily adapted for text-driven or video-driven scenarios. On the one hand, the prosperous text-to-speech (TTS) [9, 27] systems are capable of synthesizing high-quality audio from text, which equips our method with the necessary audio inputs for processing. On the other hand, we aim to build an avatar-human dataset that establishes a direct one-to-one relationship between avatars and real humans, enabling us excel in the generation of video-driven avatars in the future.

7 Conclusion

Depict characters with clear, unambiguous expressions in talking head generation task is often neglected in previous works due to the absence of suitable emotional audio-visual dataset for 3D rigs. We contribute a novel high-quality Emotional Talking Avatar Dataset providing rich and accurate affective visual and audio information with great detail. We then developed a novel approach to facial animation combines input audio and emotional cues to achieve precise lip movements and expressive expressions for stylized characters. Incorporating the built Emotional Talking Avatar Dataset, our method outperforms the SOTA emotional talking face method by applying the Soft-DTW loss, the pre-trained HuBERT feature extractor and the emotion branch. We believe our new dataset and the new method would benefit the community of expressive avatar animation and can be applied in various fields such as social VR experience, teleconferencing, visual games and storytelling.

References

[1] D. Aneja, B. Chaudhuri, A. Colburn, G. Faigin, L. Shapiro, and B. Mones. Learning to generate 3d stylized character expressions from humans. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 160–169. IEEE, 2018.

[2] D. Aneja, A. Colburn, G. Faigin, L. Shapiro, and B. Mones. Modeling stylized character expressions via deep learning. In Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13, pp. 136–153. Springer, 2017.

[3] L. Bao, H. Zhang, Y. Qian, T. Xue, C. Chen, X. Zhe, and D. Kang. Learning audio-driven viseme dynamics for 3d face animation. arXiv preprint arXiv:2301.06059, 2023.

[4] D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In Proceedings of the 3rd international conference on knowledge discovery and data mining, pp. 359–370, 1994.

[5] C. Brito and K. Mitchell. Repurposing labeled photographs for facial tracking with alternative camera intrinsics. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 864–865. IEEE, 2019.

[6] Q. Chen, Z. Ma, T. Liu, X. Tan, Q. Lu, X. Chen, and K. Yu. Improving few-shot learning for talking face system with tts data augmentation. arXiv preprint arXiv:2303.05322, 2023.

[7] M. Cuturi and M. Blondel. Soft-dtw: a differentiable loss function for time-series. In International conference on machine learning, pp. 894–903. PMLR, 2017.

[8] F. Danieau, I. Gubins, N. Olivier, O. Dumas, B. Denis, T. Lopez, N. Mollet, B. Frager, and Q. Avril. Automatic generation and stylization of 3d facial rigs. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 784–792. IEEE, 2019.

[9] C. Du and K. Yu. Rich prosody diversity modelling with phone-level mixture density network. In Interspeech 2021, Aug 2021. doi: 10.21437/interspeech.2021-802

[10] P. Edwards, C. Landreth, E. Fiume, and K. Singh. Jali: an animator-centric viseme model for expressive lip synchronization. ACM Transactions on graphics (TOG), 35(4):1–11, 2016.

[11] T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. ACM Transactions on Graphics (TOG), 21(3):388–398, 2002.

[12] S. Gururani, A. Mallya, T.-C. Wang, R. Valle, and M.-Y. Liu. Space: Speech-driven portrait animation with controllable expression. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20914–20923, 2023.

[13] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460, 2021.

[14] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In ACM SIGGRAPH 2022 Conference Proceedings, pp. 1–10, 2022.

[15] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu. Audio-driven emotional video portraits. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14080–14089, 2021.

[16] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (TOG), 36(4):1–12, 2017.

[17] J. P. Lewis, K. ichi Anjyo, T. Rhee, M. Zhang, F. H. Pighin, and Z. Deng. Practice and theory of blendshape facial models. In Eurographics, 2014.

[18] H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls, and C. Ma. Facial performance sensing head-mounted display. ACM Transactions on Graphics (ToG), 34(4):1–9, 2015.

[19] H. Li, J. Yu, Y. Ye, and C. Bregler. Realtime facial animation with on-the-fly correctives. ACM Trans. Graph., 32(4):42–1, 2013.

[20] S. R. Livingstone and F. A. Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one, 13(5):e0196391, 2018.

[21] B. Logan. Mel frequency cepstral coefficients for music modeling. In In International Symposium on Music Information Retrieval. Citeseer, 2000.

[22] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. De La Torre, and Y. Sheikh. Pixel codec avatars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 64–73, 2021.

[23] C. Malleson, M. Kosek, M. Klaudiny, I. Huerta, J.-C. Bazin, A. Sorkine-Hornung, M. Mine, and K. Mitchell. Demonstration: Rapid one-shot acquisition of dynamic vr avatars. In 2017 IEEE Virtual Reality (VR), pp. 447–448, 2017. doi: 10.1109/VR.2017.7892372

[24] Y. Pan, C. Landreth, E. Fiume, and K. Singh. Vocal: Vowel and consonant layering for expressive animator-centric singing animation. In SIGGRAPH Asia 2022 Conference Papers, pp. 1–9, 2022.

[25] Y. Pan, R. Zhang, S. Cheng, S. Tan, Y. Ding, K. Mitchell, and X. Yang. Emotional voice puppetry. IEEE Transactions on Visualization and Computer Graphics, 29(5):2527–2535, 2023.

[26] Y. Pan, R. Zhang, J. Wang, Y. Ding, and K. Mitchell. Real-time facial animation for 3d stylized character with emotion dynamics. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23). ACM, 2023.

[27] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. Learning,Learning, Jun 2020.

[28] A. Richard, C. Lea, S. Ma, J. Gall, F. De la Torre, and Y. Sheikh. Audio- and gaze-driven facial animation of codec avatars. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 41–50, 2021.

[29] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1173–1182, 2021.

[30] H. Sakoe. Dynamic-programming approach to continuous speech recognition. In 1971 Proc. the International Congress of Acoustics, Budapest, 1971.

[31] S. Tan, B. Ji, and Y. Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22146–22156, 2023.

[32] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. ACM Transactions On Graphics (TOG), 36(4):1–11, 2017.

[33] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews. Dynamic units of visual speech. In Proceedings of the 11th ACM SIGGRAPH/Eurographics conference on Computer Animation, pp. 275–284, 2012.

[34] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In European Conference on Computer Vision, pp. 700–717. Springer, 2020.

[35] H. Wu, S. Zhou, J. Jia, J. Xing, Q. Wen, and X. Wen. Speech-driven 3d face animation with composite and regional facial movements. arXiv preprint arXiv:2308.05428, 2023.

[36] Z. Ye, Z. Jiang, Y. Ren, J. Liu, J. He, and Z. Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430, 2023.

[37] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu. Audio-driven talking face video generation with learning-based personalized head pose. arXiv preprint arXiv:2002.10137, 2020.

[38] B. Yoon, H.-i. Kim, G. A. Lee, M. Billinghurst, and W. Woo. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 547–556, 2019. doi: 10.1109/VR.2019.8797719

[39] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li. Makelttalk: speaker-aware talking-head animation. ACM Transactions On Graphics (TOG), 39(6):1–15, 2020.

[40] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh.

Visemenet: Audio-driven animator-centric speech animation. ACM Transactions on Graphics (TOG), 37(4):1–10, 2018.