Emotional Voice Puppetry

Ye Pan, Ruisi Zhang, Shengran Cheng, Shuai Tan, Yu Ding, Kenny Mitchell, Xubo Yang

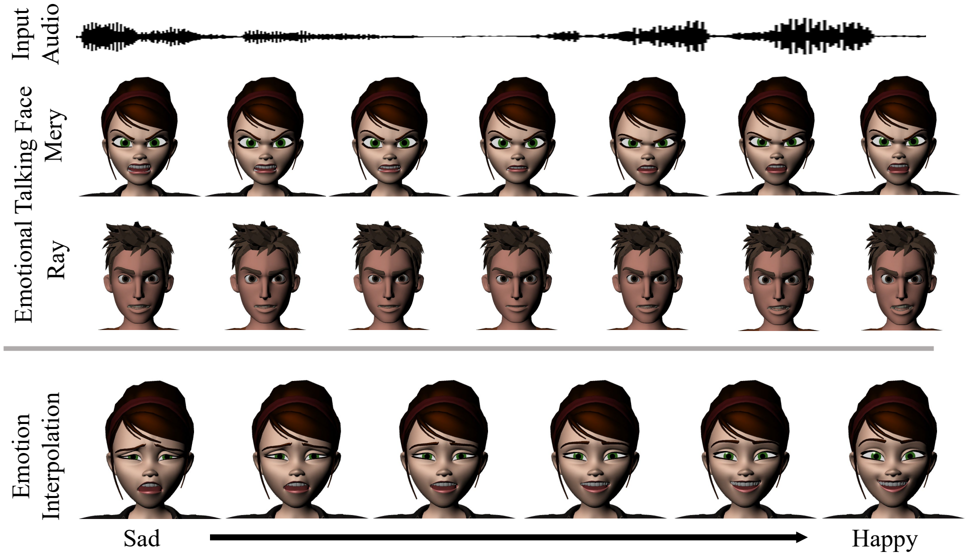


Fig. 1. Given an audio clip, Emotional Voice Puppetry is capable of generating emotion-controllable talking faces for stylized characters in a geometrically consistent and perceptually valid way. In addition, live mood dynamics can also blend smoothly.

Abstract—The paper presents emotional voice puppetry, an audio-based facial animation approach to portray characters with vivid emotional changes. 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. Our approach is exclusive because it takes account of perceptual validity and geometry instead of pure geometric processes. Another highlight of our approach is the generalizability to multiple characters. The findings showed that training new secondary characters when the rig parameters are categorized as eye, eyebrows, nose, mouth, and signature wrinkles is significant in achieving better generalization results compared to joint training. User studies demonstrate the effectiveness of our approach both qualitatively and quantitatively. Our approach can be applicable in AR/VR and 3DUI, namely, virtual reality avatars/self-avatars, teleconferencing and in-game dialogue.

Index Terms—Virtual reality, audio, emotion, character animation

1 Introduction

Many researchers enter the metaverse race with immersive meetings and they are highly interested in animating expressive talking avatars or self-avatars [4, 28]. A common challenge in virtual reality has been the difficulty in supporting the users in controlling their avatars' facial expressions, due to the obstruction of a significant part of the users' faces by the headsets [20]. This obstruction often deters effective facial capture using conventional video-based methods [29]. Consequently, the proposed audio-based facial animation has the capability of offering complementary strengths to the video-based methods, despite not quite matching unobstructed quality vision systems.

Audio-based facial animation methods have the capability of generating lip movements that are seamlessly synchronized with audio speech [42], as well as generating precisely personalized eye blink and head movements [25, 34, 38, 39, 41]. However, the emotion state of these approaches is seldom considered. Emotion is a strong feeling based on a user's circumstances, and mood is often expressed on the face through muscle motion [21]. Recently, Wang et al. proposed an emotional talking face generation baseline (MEAD) that enables the manipulation of emotion and intensity [33]. The MEAD system's primary focus was on animating realistic human faces, but not on stylized characters where the facial geometry might go beyond real human facial geometry. Applying such facial expression generation tools for stylized characters often lacks expressive quality and perceptual validity compared to artist-created animations.

Our paper proposes an emotion-controllable talking face generation framework for stylized characters. Inspired by previous human talking face generation, the study aimed at controlling stylized characters that go beyond the normal human look and act as expressive proxies for the users. The MEAD algorithms were applied at the initial stage to

- *Ye Pan is with Shanghai Jiao Tong University. E-mail: whitneypanye@sjtu.edu.cn*
- *Ruisi Zhang is with UC San Diego. E-mail: ruz032@ucsd.edu*
- *Shengran Cheng is with Shanghai Jiao Tong University. E-mail: SR-Cheng@sjtu.edu.cn*
- *Shuai Tan is with Shanghai Jiao Tong University. E-mail: tanshuai0219@sjtu.edu.cn*
- *Yu Ding is with Virtual Human Group, Netease Fuxi AI Lab. E-mail: dingyu01@corp.netease.com.*
- *Kenny Mitchell is with Roblox & Edinburgh Napier University. E-mail: k.mitchell2@napier.ac.uk.*
- *Xubo Yang is with Shanghai Jiao Tong University. E-mail: yangxubo@sjtu.edu.cn (Corresponding author).*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

map the audio to lip motion. Then the character rig parameters that complimented the mouth shape were retrieved while considering emotional intensity and category. The last stage entailed the development of a new multiple characters generalization network that enables the transfer of expression between the characters.

We also introduce a data-efficient multiple-character generalization network based on the previous ExprGen [2], which automatically learns a function to map the rig parameters of the primary characters to the secondary characters. However, ExprGen requires over 5k samples to train the mapping network. By carefully studying the Facial Action Coding System (FACS) [40] and consulting with our in-house artist, the character rig parameters were collected and categorized into five groups, namely: eye, eyebrows, nose, mouth, and signature wrinkles. Then the rig parameters were trained in parallel and then retargeted concurrently on the secondary characters. Note that our approach only requires a small number of training examples for retargeting.

We demonstrate the effectiveness of our method by comparing it to the state-of-the-art method on recognition, perceived intensity, synchronization, and naturalness & attractiveness, as these are crucial factors for audience engagement [16, 36]. Results show that our method significantly improved scores of the expression recognition & intensity while maintaining the same level of lip sync quality, naturalness & attractiveness compared to MakeItTalk. The proposed technology is highly applicable in impactful fields, including VR, in-game dialogue, and telepresence.

The main contributions of this work include the following:

- To the best of our knowledge, we presented the first emotional talking heads specially designed for 3D stylized characters in a **geometrically consistent** and **perceptually valid** way.

- We developed our framework on the FERG-3D-DB dataset by adding intensity labels for each character, and extensive user studies validated the effectiveness.

- Our multiple-character generalization network significantly improved the generalization and efficiency of retargeting on new characters.

- We propose new metrics to evaluate the recognition, intensity, synchronization, naturalness, and attractiveness of different talking head animation approaches. Extensive experiments demonstrated their effectiveness.

2 Related work

2.1 Audio driven facial animation

The literature review focuses on audio-based facial animation. Several researchers have studied video-based facial animations and found that video-driven animations have the capability of creating more realistic facial expressions when they capture the facial performance of a human actor. The primary downside of performance capture compared to the animator-generated animation is that it is visually restricted by the performance of the actor and in most cases is void of the expressive quality and perceptual validity [2]. Our objective is to generate expressive and plausible 3D facial animations based solely on audio. The audio-based techniques can be organized in facial enactment, which aims at creating photo-realistic videos of the existing human such as idiosyncrasies, and facial animation, which focuses on expression prediction that can be utilized with a predefined simulator or avatar [30].

There exist many previous studies on audio-based facial animation, however, most of the studies concentrated on the relationship between speech content and the shape of the mouth [11, 23, 44]. The brand pioneered Voice Puppetry to generate full facial animation from an audio track [6]. Karras et al. proposed a neural network, which stacks several convolution layers, to generate the 3D vertex coordinates of a face model from the audio and known emotions [18]. Zhou et al. developed speaker-aware talking head animations from a single image and an audio clip by decomposing the input audio into speaker and content information and then applying a deep learning-based method [43]. Guo et al. adopt the neural radiance field (NeRF) representation for scenes of talking heads [14]. However, the emotion is not addressed.

Additionally, some studies have looked at head motions and eye blinks. One such study is covered in Chen et al. article where the authors synthesize videos of talking face with natural head movements through the explicit generation of head motions and facial expressions [8]. One of the most challenging aspects of synthesizing talking face videos is that the natural poses of a human results in head motions that are either in-plane or out-of-plane. Yi et al.'s article where the authors reconstructed a 3D face animation and then rendered into synthesized frames [37]. Hao et al. present a two-stage approach for the generation of talking-face videos that have realistic controllable eye blinking capabilities [15]. Liu et al. produced talking face that have controllable eye blinking capabilities which are driven by joint features of identity, audio and blinking [22]. Given the extensive level of research on the synthetization of talking face videos, this paper focuses on the animation of 3D stylized characters where the head and eye are animated by rig parameters.

A few studies have looked at human face emotions [27], for instance, Wang et al. developed a large-scale emotional audio-visual dataset (MEAD) that contained talking face videos having varied emotions at varying intensity levels. In addition, the researchers proposed an emotional talking head generation baseline that was essential in manipulating emotions and their strength [33]. Ji et al. attained emotional control by editing video-based talking face generation approaches [17]. Again, the developed systems were used in animating human faces and not 3D stylized characters.

2.2 Facial expression for stylized characters

The successful creation of an animated story depends on the emotional state of a character, which must always be staged unambiguously [19, 32]. Keyframing is a prevalent method for animating characters with clear emotions and artistic expressions [1]. It is a simple method of animating a character but is often time-consuming. Recently, character expressions have been generated by motion capture systems which use modeled features on human faces and geometric markers [35]. Nevertheless, these features do not precisely match the stylized character expressions. Therefore, facial geometric features alone cannot produce the desired and perceptually valid stylized character expression.

On the other hand, photographic precision (for instance, precise drawing of facial wrinkles) is not a certainty of achieving an accurate communication of emotion. Underlying every emotion, only a limited set of elements are the actual basis of our recognition. Primitive artists and cartoonists have invented unexpected and extraordinary graphic substitutes for actions and features. Generally, the stylized or abstracted interpretation of expressions must always rely on the actual nature of the human face [12].

Aneja et al. proposed ExprGen, a multi-stage deep learning system that can generate stylized character expressions from human face images in a way that is geometrically consistent and perceptually valid [2, 3]. This idea inspired this study, and thus the proposed Emotional Voice Puppetry System that develops character expressions using audio as the sole input.

3 Preliminary

3.1 Emotion Categories and Intensities

Categories We utilized four 3D stylized characters, namely Mery, Bonnie, Ray, and Malcolm acquired from the Facial Expression Research Group 3D Database (FERG-3D-DB) possess annotated facial expressions that were categorized into seven groups namely anger, fear, disgust, sadness, joy, surprise, and neutral.

Intensities The original dataset did not label the intensities. They are labeled according to the changes in facial expressions. When the facial expression is categorized under anger, fear, or surprise, the expression is expected to be more pronounced with enhanced eye-opening, while the eye closes and face becomes lackluster if the expression is

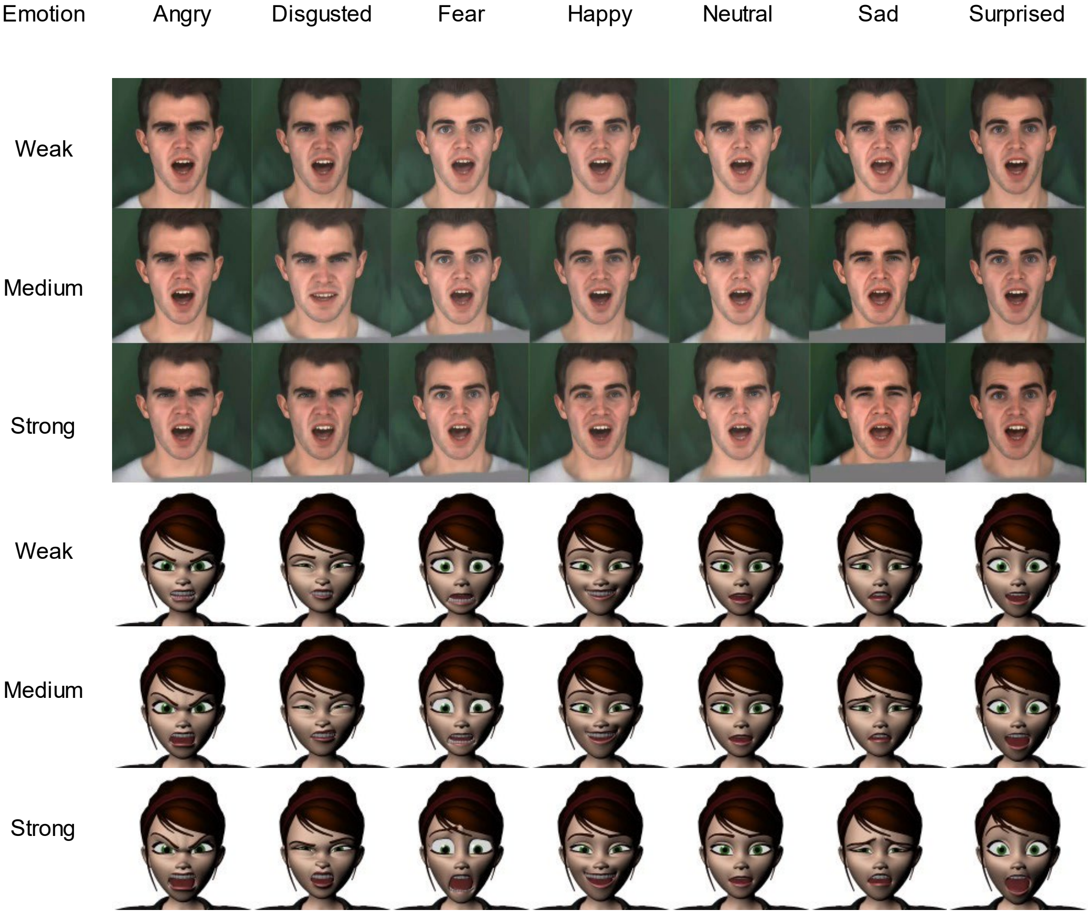


Fig. 2. Resulting images generated from MEAD and our method for seven emotion categories and three intensities.

the opposite. The intensity label for the expression categories was also approximated by first disentangling them into five parts, namely eye, eyebrows, nose, mouth, and signature wrinkle, and then calculating the offset of the expressions relative to the neutral expression for every part identified. Equal weights are given to the five parts, thus offsetting the rig parameters and ranking them into the three intensity levels.

Based on these assumptions, we then define a three levels of emotion intensity.

- WEAK describes the slight or gentle but detectable facial motion.
- MEDIUM describes the normal emotion state or the typical emotion expression.
- STRONG describes the exaggerated facial expressions characterized by intense emotion in the face area.

4 Method

Overview As summarized in Figure 3, we propose an emotional talking-head generation method for stylized characters that is able to automatically manipulate emotion and intensity. We used three-branch architecture to process the audio and emotion distinctly. We first map the audio to lip movements of the base character and retrieve the desired emotion on the upper face from the preprocessed FERG-3D-DB dataset in Section 3.1, and then added the head pose & eye blink. Lastly, the acquired expression parameters are utilized to generate expressions on multiple secondary 3D stylized characters.

4.1 Lip

Audio-to-Landmarks The input audio is converted to lip landmarks of the talking face by first extracting the Mel-Frequency Cepstral Coefficients (MFCC) [24] from the audio. We pair the video frames and audio features using a one-second temporal sliding window with the sample rate set to 30. Based on the audio temporal properties, a long short-term memory (LSTM) network and a fully connected layer are applied to predict the lip's motion. The L2 loss function is established to define the audio-to-landmark task.

$$Loss_{a2l} = \left\| F(x) - l_{gt} \right\|_2 , \tag{1}$$

where x and l_{gt} are the input audio and the corresponding ground truth lip landmark, respectively, and $F(\cdot)$ represents the audio-to-landmark module.

4.2 Emotional rig parameter & Matching

We bridge the generated mouth landmarks with rig parameters of Mery performing the following steps. We first render Mery's 2D images with given rig parameters in FERG-3D-DB. Then, we apply the face detection algorithm [7] to obtain 20 landmarks associated with the mouth. Finally, we use the two-stage matching method to obtain the best-matched rig parameter and mouth landmark pairs: (1) We first filter five rig parameters based on the L_2 distance in Equation2. The l^i_{render} denotes the landmark i of the rendered image. (2) Then, we calculate the open mouth similarity between the generated and detected landmarks based on Equation3. The open mouth distance is given in Equation4, where *upper* and *lower* denote the index of the upper and lower lip, respectively.

$$D_{stage1} = \sum_{i=1}^{20} \left\| F(x)^i - l^i_{render} \right\|_2 , \tag{2}$$

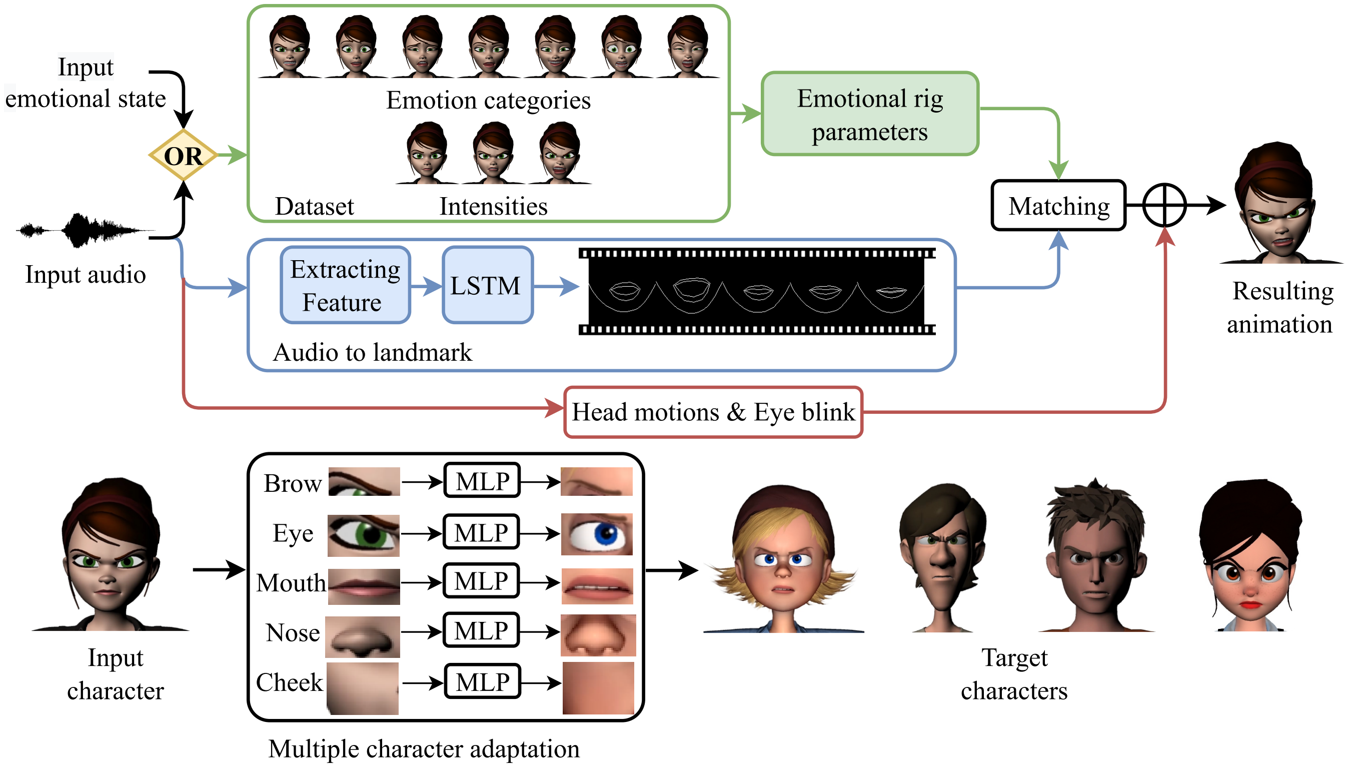


Fig. 3. The overview of emotional voice puppetry. Our method includes three branches to process the mouth shape, emotion and head pose & eye blink respectively.The first branch maps the audio to lip movements of the base character and retrieve the desired emotion on the upper face from the preprocessed FERG-3D-DB dataset in Section 3.1 The second branch adds the additional head pose & eye blink. The third branch utilizes the acquired expression parameters to generate expressions on multiple secondary 3D stylized characters. Finally, a multiple character adaptation network performs input character to target characters expression transfer.

$$D_{stage2} = \|mouth(F(x)) - mouth(l_{render})\|_2 , \tag{3}$$

$$mouth(y) = \left\| y^{upper} - y^{lower} \right\|_2 , \tag{4}$$

4.3 Head motions, Eye gaze & Eye blink

We employ LSTM-based generators to create the relationship between head motions and input audio. We first construct an audio-to-head motion dataset. Our dataset is adapted on the basis of Multi-view Emotional Audio-visual Dataset (MEAD). We then employed OpenFace [5] to create corresponding 3D rig parameters based on data from the MEAD dataset. The LSTM network is then trained to generate head movements based on the audio. The same methodology is also used to generate eye movements.

4.4 Smoothing Optimization

After getting the rig parameter and mouth landmark pairs, we can feed them into the second branch. Due to the clips that may exist in matching, we introduce smoothing to make the frames more consistent and natural. Here, we simply use linear interpolation to add frames between two neighbor keyframes based on their rig parameters.

4.5 Multiple character adaptation

Generalizing facial expressions to multiple characters plays an important role in the animation field. A prior method in ExprGen [2] employs two steps to transfer animation to multiple characters. It first resolves training pairs by measuring the feature distance between two characters. Then, they use the controller values of input character and target character to train MLP models. Though the character pairs' feature distance is close, the two characters' facial geometry details might be different. For example, the input character and target character are both smiling and we retrieved the best-matched target character expression. The retrieved expression might be similar in upper face but different in lower face. Additionally, to control a character's facial expression more precisely, we usually need several hundred parameters for both input and target characters. When training MLP networks, we need a larger dataset to match these pairs (about 10k examples). It requires more manual effort overall and, thus, is harder to generalize to another character.

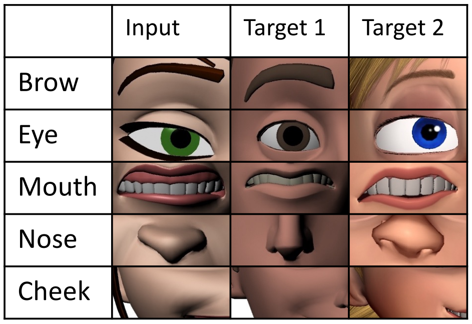


Fig. 4. Matching samples of different parts

The facial expressions are controlled by different independent action units. We divide the face of character into five parts: brow, eye, mouth, cheek and nose 4. For each part, we use the rig parameters as feature vectors and train a separate MLP network to generalize the character's expressions.

We propose a new model to generalize input character expressions to

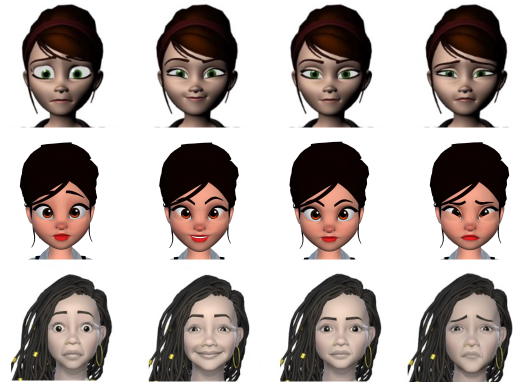


Fig. 5. Multiple character adaptation to new characters. The first row is the input character Mery, the second and third rows are the target character Waitress and Miosha.

multiple target characters. In the matching step, we split the character face into different parts and match their geometry. In the training step, we use different MLP networks to train the facial components separately. Then, the MLP networks' outputs are combined to control target characters.

Matching To match the input character with the target characters, we first split the input character's face into five parts, namely mouth, nose, brow, eye, and cheek. For each part, we perform a two-step filtering similar to the matching algorithm in Section 4.1. In the first step, we directly filter images with the same emotion annotation as the input character. In the second step, we use the following landmark distance to formulate geometry feature vectors: **mouth**: mouth width (left mouth corner to right mouth corner distance), closed mouth height (distance is vertical between the upper and the lower lip); **nose**: nose width (distance is horizontal between leftmost and rightmost nose landmarks); **brow**: left/right eyebrow height (distance is vertical between top of the eyebrow and center of the eye), **eye**: left/right eyelid height (distance is vertical between top of an eye and bottom of the eye), and left/right lip height (distance is vertical between the lip corner from the lower eyelid), **cheek**: the distance from nose to leftist face, and lowest face. Then, for each part in the face at i-th frame, let the input character's landmarks denoted as l^i_{input} and the target character's landmarks denoted as l^i_{target}, we retrieve the closest landmark by minimizing the L_2 distance $\mathbf{D}$ between input and target landmarks as shown in Equation 5.

$$\mathbf{D} = argmin_i ||l^i_{input} - l^i_{target}||_2 \tag{5}$$

Training We create a separate multilayer perceptron (MLP) for each facial part, which consists of N output nodes, M input nodes, and a hidden layer with ReLU activation. The M and N are the dimension of facial parts in input character and target character. The parameters change with different characters and different facial regions used in training. When training the model, gradient descent is used with a mini-batch size of 10 and a learning rate of 0.01 to minimize the square loss between the ground truth and output parameters, where the ground truth is obtained from the matching step.

5 Evaluation

5.1 Comparison to the state-of-the-art

We compare our emotion-controllable talking face generation approach with MEAD and MakeItTalk. We include MEAD system, because our system also used Multi-view Emotional Audio-visual Dataset, and their emotional talking-face generation baseline. However, our system is developed for stylized characters, instead of real human. Additionally, we also include MakeItTalk to demonstrate our system achieved comparable performance as state-of-the-art models in terms of lip synchronization and facial geometry, although emotion is not addressed in their approach.

5.1.1 Participants

We recruited 20 participants from Shanghai Jiao Tong university. The average age of participants was 21, with an age range of 20-23 years, of which 10 were female and 10 were male. All of the participants are majoring in engineering. They were naïve to the purposes of the experiment.

5.1.2 Design

The experiment utilized animated characters 6 (Human created via MEAD, MakeItTalk, Mery, Bonnie, Ray & Malcolm) × 7 emotions (Neutral, Anger, Sadness, Fear, Disgust, Happiness, & Surprise) × 3 intensities × 4 audio in a mixed design, with a between-subject design for audio, but a within-subject design regarding characters, emotions, and intensities.

Each participant engaged in 126 (6 characters × 7 emotions × 3 intensities = 126 trials) experimental trials by randomly presenting video clips to them to reduce cases of fatigue.

5.1.3 Procedure

The participants all signed the consent form before engaging in the trial. They were presented with a video clip that they viewed and afterward answered the related questions. The questions were close-ended, and the participants had to choose pre-defined responses.

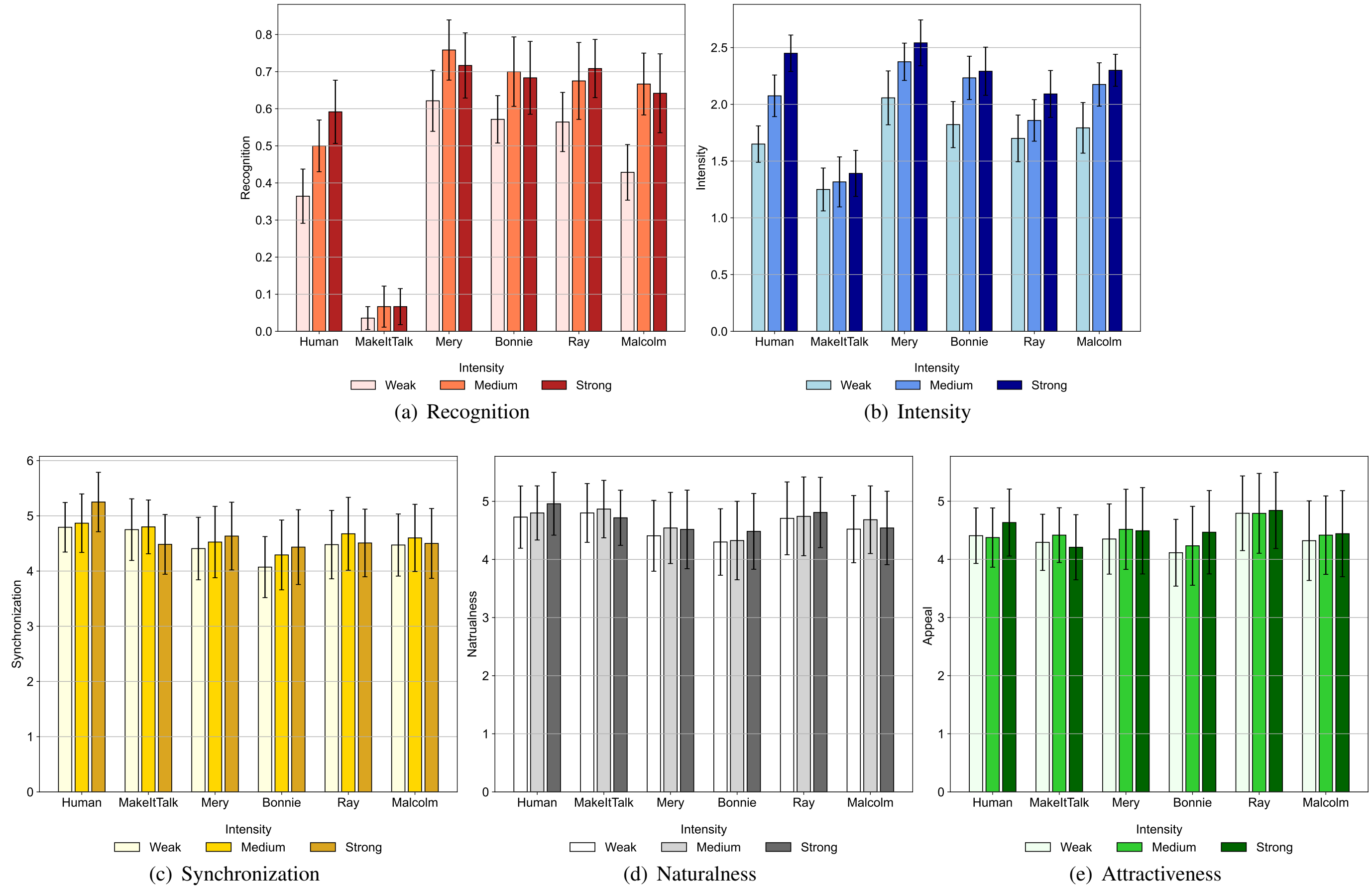
- Which expression did the character depict? Participants were asked to select one of the below responses: neutral, anger, sadness, fear, disgust, happiness, surprise or other.
- How intense was the indicated emotion depicted by the character? Participants rated the intensity on a scale from 1 to 3, where; 1 is weak and 3 is strong.
- 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 talking head overall?" Participants rated naturalness on a scale from 1 to 7, where 1 is not natural at all and 7 is extremely natural.
- How attractive was the character overall?" Participants rated attractiveness on a scale from 1 to 7, where; 1 is not attractive at all and 7 is extremely attractive.

The participants engaged in one practice trial where asking questions was allowed, and then took 126 trials that were measured.

The participants were paid 50 RMB. The experiment took about 30 minutes. The experiment was approved by Shanghai Jiao Tong University Research Ethics Committee.

5.1.4 Results & Discussion

We applied Analysis of variance (ANOVA) on separate repeated measures for results on recognition, intensity, synchronization, naturalness and attractiveness. The ANOVA results showed within-participant factors of emotion (7), character (6), and intensities (3). Using the Shapiro-Wilk test, the data were distributed normally for all the assessed conditions, and the boxplot showed no outliers. Furthermore, Mauchly's test was conducted to evaluate data sphericity and possible cases of data sphericity violation. In those cases, we applied with Greenhouse-Geisser correction and marked with an asterisk"*". The post hoc test used was the Bonferroni test for multiple means comparison.



(a) Recognition

(b) Intensity

(c) Synchronization

(d) Naturalness

(e) Attractiveness

Fig. 6. Mean for each controlled intensity level and character on recognition, perceived intensity, synchronization, naturalness & attractiveness. Error bars show standard deviation.

Recognition

For the recognition of expressions, responses were converted to scores, “1” for correct or “0” for incorrect, and then averaged over stimuli repetitions. Figure 6(a) shows the comparison of average scores obtained for three intensities across 6 characters.

First, the MakeItTalk had the lowest mean score $(Mean, M = .193, Standard Error, SE = .016)$. The main effect of characters was significant, $F(5,95) = 80.175, p < .001$. Bonferroni post-hoc comparisons indicated the mean for MakeItTalk is significantly lower than Human, $p < .001$. This is expected, because MakeItTalk animates characters based on geometric markers only.

Second, the average score for Human $(M = .519, SE = .03)$ is significantly lower than the average score for all stylized characters, e.g., primary character Mery $(M = .764, SE = .031).p < .001$. This could be due to the characters’ simpler geometry and stylization, which makes the expressions simpler to discern.

Third, Bonferroni post-hoc comparisons also shows that the mean Mery, Bonnie $(M = .714, SE = .027)$, Ray $(M = .705, SE = .033)$, Malcolm $(M = .624, SE = .031)$ did not significantly differ from each other, $p > .05$. This demonstrated the effectiveness of our multiple character generalization network.

Intensity

Figure 6(b) shows user perceived intensity ratings for three labeled intensities (WEAK, MEDIUM, & STRONG) across 6 characters. Firstly, the main effect of controlled intensities was significant, $F(1.154, 21.918) = 26.266, p < .001^*$. Bonferroni post-hoc comparisons indicated the mean ratings for WEAK $(M = 1.745, SE = .078)$ is significantly lower than MEDIUM, $(M = 1.973, SE = .06).p < .001$ $p = .016$. The mean for MEDIUM also significantly differ from STRONG, $(M = 2.092, SE = .055), p = .022$. This shows our labeled emotion intensities are well distinguished.

Additionally, Figure 6(b) also shows mean ratings for MalkItTalk are the lowest among all characters. We note that the main effect of character was significant, $F(2.892, 54.942) = 47.626, p < .001^*$. Bonferroni post-hoc comparisons indicated the mean ratings for MakeItTalk $(M = 1.371, SE = .084)$ is significantly lower than Human $(M = 1.99, SE = .065)$, Mery $(M = 2.257, SE = .075)$, Bonnie $(M = 2.067, SE = .07)$, Ray $(M = 1.871, SE = .065)$, and Malcolm $(M = 2.062, SE = .066), p < .001$. This is expected, as MakeItTalk’s main contribution is focused on creating better lip synchronization, head motions and personalized facial expressions, instead of generating expressive and emotion for stylized characters.

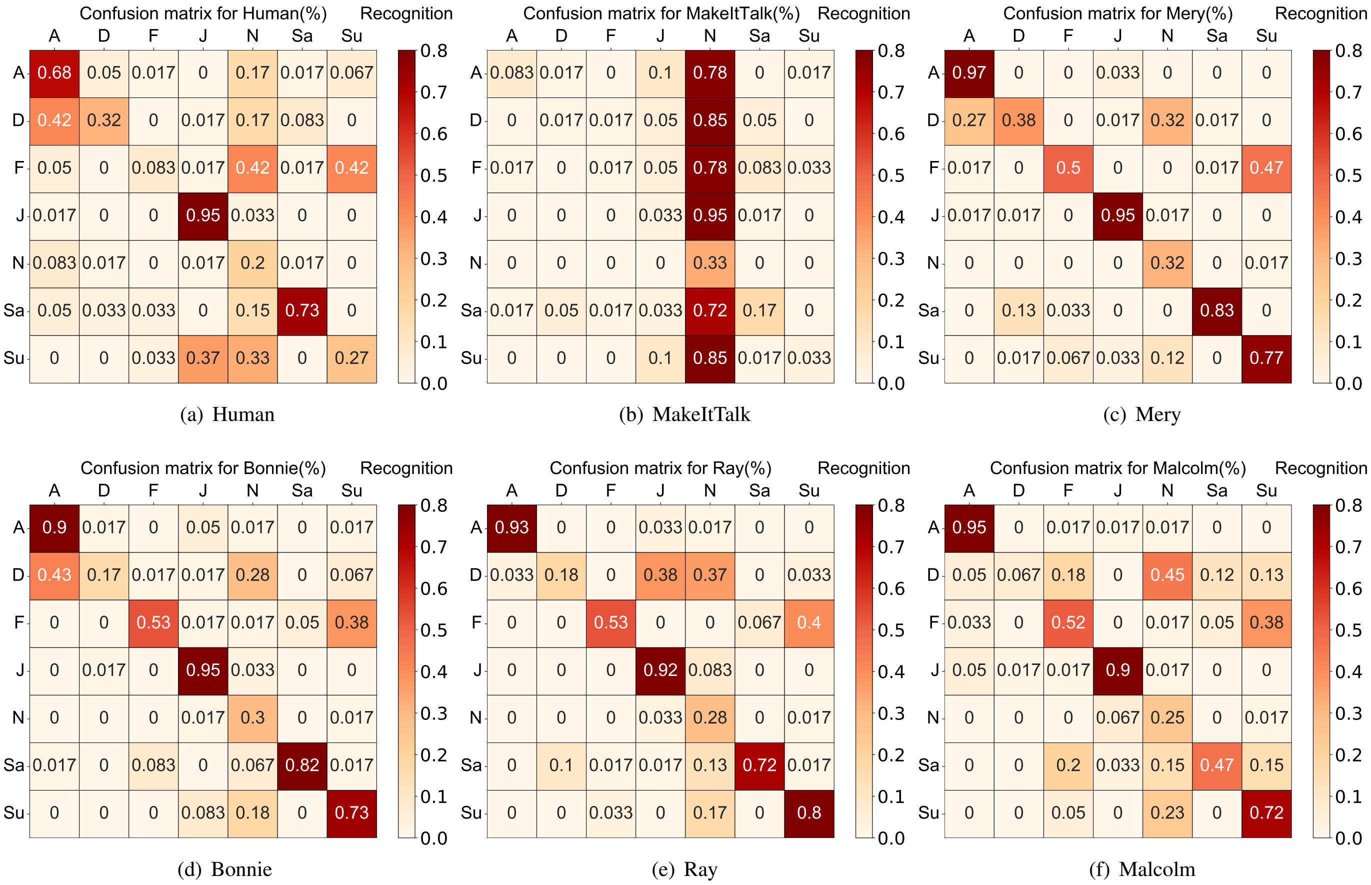
Synchronization

Figure 6(c) shows the lip synchronization scores for 3 intensities across 6 characters. The mean lip sync scores of our four stylized characters are similar to both Human and MakeItTalk. Results reveal that the main effect on intensities and characters were not significant, $F(2,38) = 3.598, p = .037$, and $F(2.602, 49.44) = 1.806, p = .165^*$, respectively. This is expected, because we used the same audio-to-lip method as Human to generate the mouth shape.

Naturalness & Attractiveness

Figure 6(d) shows the rating on naturalness for 3 intensities across 6 characters. Results shows that the main effect on intensities and characters were not significant, $F(1.311, 24.913) = 4.393, p = .057^*$, and $F(2.439, 46.339) = .935, p = .415^*$, respectively.

Figure 6(e) shows the rating on attractiveness for 3 intensities across 6 characters. Results shows that the main effect on intensities and characters were also not significant, $F(2.748, 52.218) = 1.556, p = .214^*$, and $F(1.453, 27.6) = 3.897, p = .063^*$, respectively.



(a) Human

(b) MakeItTalk

(c) Mery

(d) Bonnie

(e) Ray

(f) Malcolm

Fig. 7. Confusion matrix for perceived expression recognition (%) for seven expression classes.

5.2 Multiple Character Adaptation

Separate training results We first use Mery[1] as the input character, Bonnie[2] and Ray[3] as the target characters. The input and target characters can be changed to any other characters. There are 100 rigged parameters to control Mery's facial movements, 59 parameters for Bonnie, and 187 parameters for Ray. We first show how our multiple character adaptation is applied in each facial part in Figure 4.

Generalize to new characters We also show how our adaptation algorithm can be applied to new characters outside FERG dataset in Figure 5. We take Mery as the input character, Waitress and Miosha as the target character. There are 108 rigged parameters in Waitress and 110 rigged parameters in Miosha. We use Mery to match the parameters of Waitress and Miosha by rigging them to speak "Kids are sitting by the door" for seven emotions. 700 frames for each character are used as the training dataset From the Figure 5, we can find our algorithm achieves both geometry and emotion consistent with the input character.

5.3 Ablation Study

In this section, we conduct ablation studies over different components in our emotional talking-head generation method to demonstrate its effectiveness.

5.3.1 Effectiveness of two-step filter

As shown in Figure 8, we display the original mouth landmark, images retrieved using 1-step matching, and 2-step matching. From the images,

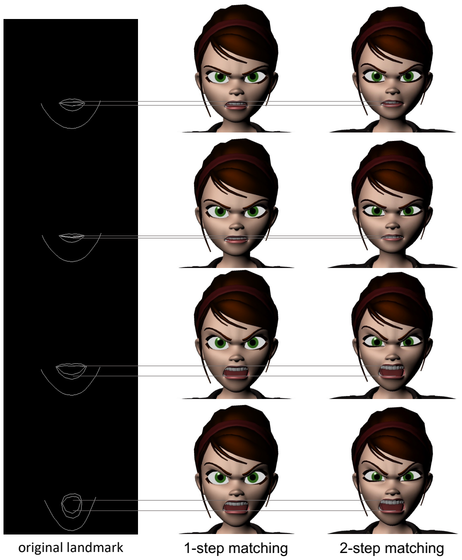


Fig. 8. Retrieving rendered image with and without two-step filter algorithm

[1] https://www.meryproject.com
[2] https://www.joshsobelrigs.com/bonnie
[3] https://www.cgtarian.com

we can find images retrieved using 2-step matching has more accurate representation of the original mouth landmark.

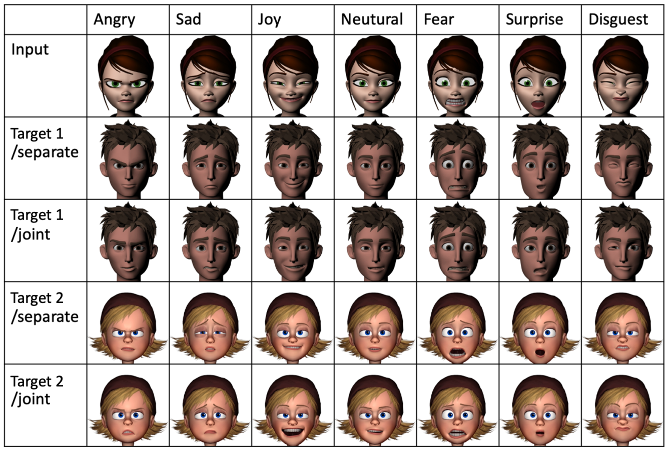


Fig. 9. Separate and joint training results

5.3.2 Effectiveness of separate training

Figure 9 shows the training results with different samples for both separate training and joint training. We trained both separate MLP and joint MLP with about 700 examples(for each expression, we used 100 examples). When fewer examples are available for training, a separate training strategy can maintain better geometry representation of the input character.

6 Discussion

Characters We used 'Mery' as the base character, 'Bonnie', 'Ray' & 'Malcolm' as the secondary character. Results show there was no significant effect among these characters in terms of expression recognition & intensity. This indicates our solutions were effective irrespective of whether that character is primary or secondary.

Emotion We look into recognition for seven expression classes. The main effect of emotion was significant, $F(6,114) = 40.622, p < .001$, according to our preliminary data on expression recognition. Figure 7 depicts the confusion matrix for each expression class's perceived expression recognition. For a specific row (e.g., anger) in each subfigure, the columns show the percentage (e.g., averaging nearly over all observed individual anger expressions) of respondents agreeing on the associated expression classes.

For MakeItTalk, We note that the majority of expressions are incorrectly perceived as neutral. For both human and our charaters, surprise and joy are highly accurate, while fear and disgust are extremely difficult for people to recognize and express.

Application, limitations and future work Our approach supports an immersive teleconferencing experience where users holding face-to-face meetings in virtual meeting rooms. A key challenge that is holding back VR communication is the fact that the majority of the face is occluded by the HMD. Employing our audio driven approach in the development of 3D facial animation makes it possible to generate believable, recognizable emotional avatars. These types of avatars have the capacity to facilitate interaction between VR users using head-mounted displays. Most high-level approaches [13, 20, 26, 31] towards the development of talking heads in the VR space largely focus on the generation of geometrically correct 3D avatars, but tend to lack the perceptual validity and expressive quality of animator-created animations. Given the importance of emotions in communication, this paper presents a deployable solution for presenting a fully immersive and compelling teleconferencing experience by enhancing the recognition and believability of the avatars' expression.

Our method achieves real-time. It processes approximately 639 frames per second. The method employed is also easily applicable to networks with low bandwidth. This is made possible by the fact that the methodology only requires transferring the character's 3D rig parameters from the server which is then directly animated at the client side, rather than transferring videos.

Our approach can also facilitate character animation for in-game dialogue and immersive storytelling. Traditionally, believable animated characters are created and refined iteratively by artists, but they require significant expertise and a tremendous amount of time. Our approach can provide a more efficient and timely way to generate animation in a geometrically consistent and perceptually correct way.

There are many potential avenues for next steps. It would be interesting to validate the results using quantitative metrics. To evaluating lip sync for real human talking face, SyncNet [10] or LMD [9] scores are commonly used for comparison with other baselines. However, these metrics are difficult to apply to stylized characters because of their exaggerated mouth shape and artistic expressions. Thus, we developed user studies demonstrate the effectiveness of our approach. An obvious route for next step is to develop quantitative evaluation metrics for stylized characters.

7 Conclusions

We propose a new approach for animating live 3D talking heads that is capable of manipulating emotion and its intensity by solely relying on audio. Our system is dependent on the existing architecture but has a unique contribution of creating 3D stylized characters that can amplify over expression clarity instead of geometric markers. Another main contribution is that our method can more easily generalize to multiple character targets compared to state-of-the-art methods (e.g., ExprGen). Using the FACS, the facial characters are disentangled into eye, eyebrows, nose, mouth, and signature wrinkles, which allows the expression transfer of a primary character to the secondary characters only through annotation of a few poses for every new character. An evaluation model was developed, and the effectiveness of the developed voice puppetry system was demonstrated. The new system can be applied in various fields such as social VR experience, teleconferencing, visual games and storytelling.

Acknowledgments

This work was supported by Shanghai Sailing Program (No. 20YF1421200), National Natural Science Foundation of China (NSFC, NO. 62102255), CAROUSEL+ (EU FET PROACT, NO.101017779) and Hangzhou Key Science and Technology Innovation Program (No. 2022AIZD0054).

References

[1] A. Agarwala, A. Hertzmann, D. H. Salesin, and S. M. Seitz. Keyframe-based tracking for rotoscoping and animation. *ACM Transactions on Graphics (ToG)*, 23(3):584–591, 2004.

[2] 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.

[3] D. Aneja, R. Hoegen, D. McDuff, and M. Czerwinski. Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2021.

[4] M. Baloup, T. Pietrzak, M. Hachet, and G. Casiez. Non-isomorphic interaction techniques for controlling avatar facial expressions in vr. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10, 2021.

[5] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 59–66. IEEE, 2018.

[6] M. Brand. Voice puppetry. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 21–28, 1999.

[7] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[8] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pp. 35–51. Springer, 2020.

[9] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu. Lip movements generation at a glance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 520–535, 2018.

[10] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2017.

[11] 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.

[12] G. Faigin. *The artist's complete guide to facial expression*. Watson-Guptill, 2012.

[13] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8649–8658, 2021.

[14] Y. Guo, K. Chen, S. Liang, Y.-J. Liu, H. Bao, and J. Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5784–5794, 2021.

[15] J. Hao, S. Liu, and Q. Xu. Controlling eye blink for talking face generation via eye conversion. In *SIGGRAPH Asia 2021 Technical Communications*, pp. 1–4. 2021.

[16] J. Hyde, E. J. Carter, S. Kiesler, and J. K. Hodgins. Using an interactive avatar's facial expressiveness to increase persuasiveness and socialness. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 1719–1728, 2015.

[17] 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.

[18] 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.

[19] J. Lasseter. Principles of traditional animation applied to 3d computer animation. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pp. 35–44, 1987.

[20] 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.

[21] L. Li, S. Wang, Z. Zhang, Y. Ding, Y. Zheng, X. Yu, and C. Fan. Write-a-speaker: Text-based emotional and rhythmic talking-head generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 1911–1920, 2021.

[22] S. Liu and J. Hao. Generating talking face with controllable eye movements by disentangled blinking feature. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[23] Y. Liu, F. Xu, J. Chai, X. Tong, L. Wang, and Q. Huo. Video-audio driven real-time facial animation. *ACM Transactions on Graphics (TOG)*, 34(6):1–10, 2015.

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

[25] Y. Lu, J. Chai, and X. Cao. Live speech portraits: real-time photorealistic talking-head animation. *ACM Transactions on Graphics (TOG)*, 40(6):1–17, 2021.

[26] K. Olszewski, J. J. Lim, S. Saito, and H. Li. High-fidelity facial and speech animation for vr hmds. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016.

[27] H. X. Pham, Y. Wang, and V. Pavlovic. End-to-end learning for 3d facial animation from speech. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 361–365, 2018.

[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] 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.

[30] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European Conference on Computer Vision*, pp. 716–731. Springer, 2020.

[31] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *arXiv preprint arXiv:1610.03151*, 2016.

[32] F. Thomas, O. Johnston, and F. Thomas. *The illusion of life: Disney animation*. Hyperion New York, 1995.

[33] 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.

[34] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021.

[35] T. Weise, S. Bouaziz, H. Li, and M. Pauly. Realtime performance-based facial animation. *ACM transactions on graphics (TOG)*, 30(4):1–10, 2011.

[36] P. Wisessing, K. Zibrek, D. W. Cunningham, J. Dingliana, and R. McDonnell. Enlighten me: Importance of brightness and shadow for character emotion and appeal. *ACM Transactions on Graphics (TOG)*, 39(3):1–12, 2020.

[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] C. Zhang, S. Ni, Z. Fan, H. Li, M. Zeng, M. Budagavi, and X. Guo. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[39] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3867–3876, 2021.

[40] W. Zhang, Z. Guo, K. Chen, L. Li, Z. Zhang, Y. Ding, R. Wu, T. Lv, and C. Fan. Prior aided streaming network for multi-task affective analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3539–3549, 2021.

[41] Z. Zhang, L. Li, Y. Ding, and C. Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021.

[42] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4176–4186, 2021.

[43] 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.

[44] 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.