

Audio-Driven Speech Animation with Text-Guided Expression

Sunjin Jung¹ , Sewhan Chun² , Junyong Noh¹ 

¹KAIST, Visual Media Lab
²NAVER Cloud

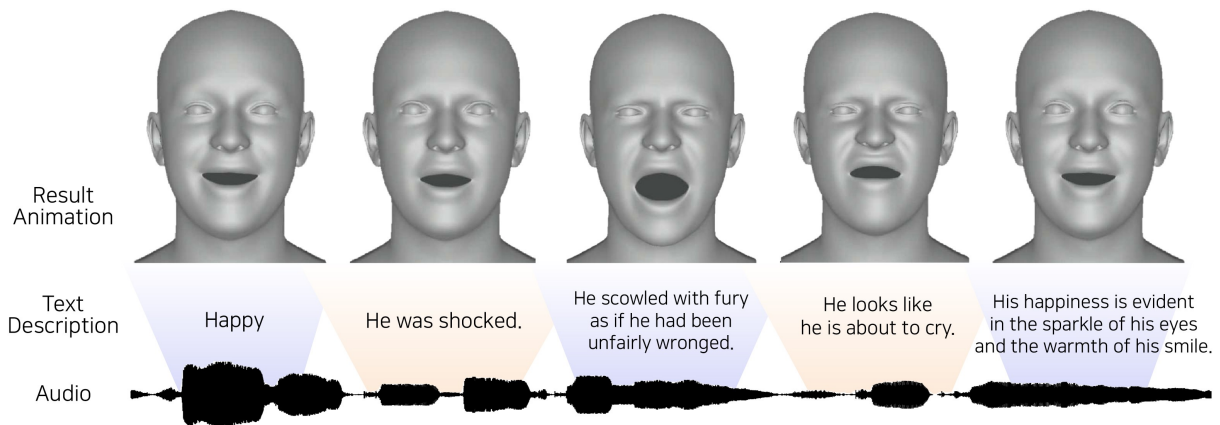


Figure 1: Our approach can produce expressive speech animation based on both audio and text description. The mouth movements are determined by the speech audio input, while the facial expression is generated using the text information. The text can encompass various forms of natural language descriptions, including emotional words for the character or detailed sentences describing the facial expressions.

Abstract

We introduce a novel method for generating expressive speech animations of a 3D face, driven by both audio and text descriptions. Many previous approaches focused on generating facial expressions using pre-defined emotion categories. In contrast, our method is capable of generating facial expressions from text descriptions unseen during training, without limitations to specific emotion classes. Our system employs a two-stage approach. In the first stage, an auto-encoder is trained to disentangle content and expression features from facial animations. In the second stage, two transformer-based networks predict the content and expression features from audio and text inputs, respectively. These features are then passed to the decoder of the pre-trained auto-encoder, yielding the final expressive speech animation. By accommodating diverse forms of natural language, such as emotion words or detailed facial expression descriptions, our method offers an intuitive and versatile way to generate expressive speech animations. Extensive quantitative and qualitative evaluations, including a user study, demonstrate that our method can produce natural expressive speech animations that correspond to the input audio and text descriptions.

CCS Concepts

• **Computing methodologies** → **Animation; Neural networks;**

1. Introduction

The significance of speech animation for 3D characters has become increasingly pronounced across various digital media applications, including films, animations, games, mixed reality, and virtual assistants. Effective speech animation requires precise synchronization of lip movements with sounds and the integration of expressive facial motions. Humans naturally observe both lip movements and

facial expressions together to interpret the message that the speaker tries to convey. Therefore, any disharmony between the lip movements and facial expressions can significantly deteriorate the realism of the resulting animation and emotional engagement of the user/viewer with character. To achieve a high level of believability and immersive experience, both accurate lip sync and nuanced facial expressions are essential in 3D character speech animation.

© 2024 The Authors.

Proceedings published by Eurographics - The European Association for Computer Graphics. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

There are various methods for generating realistic speech animation. Traditionally, the animation generation based on key-framing or performance capture has been a popular choice. While it is well-known that key-framing can produce high-quality animation results, this approach is labor-intensive and requires expertise from skilled artists. Performance capture automates the process by mimicking the motion of the performer and transplanting it to the target character. This technique can create highly realistic animations at the cost of expensive equipment and a controlled environment, making it less accessible to general users. Furthermore, both key-framing and performance capture require revisiting and non-trivial refinement of the work when adjustments are needed after the animation is completed, such as changing the dialogue.

Advances in neural networks have enabled data-driven facial animation using intuitive inputs such as speech audio, as exemplified by many recent attempts [TKY*17, ZXL*18, CBL*19, FLS*22, XXZ*23]. Unlike traditional techniques, neural network-based approaches can automate much of the animation process, reducing the need for manual intervention. These approaches take speech audio as input and generate facial animations for a 3D character, aiming to synchronize lip movements with the speech. While these methods primarily focus on generating accurate lip movements, they often neglect the synthesis of accompanying emotional expressions essential for creating truly immersive and engaging speech animations.

Another line of research has explored generating expressive facial animations for audio-driven speech animation [CCK*21, PWS*23, DCT*23, LLZP24, LLSL24]. Facial expressions are vital in conveying emotions, as people can make different expressions while saying the same words depending on their emotional state. To address this, two main approaches have emerged depending on how to provide emotional features as additional conditioning inputs for the facial expressions. The first approach extracts emotional features from the audio, and generates facial expressions using the features [CCK*21, PWS*23, LLSL24]. The second approach involves providing one of a few emotion class labels as a condition to generate the corresponding facial expression [DCT*23, LLZP24]. Despite the reported successes, the first approach is limited to a specific expression per audio input, and similarly the resulting expression from the second approach is confined to one of the emotion classes used during training.

In this paper, we propose a novel method to generate audio-driven expressive speech animation from text descriptions. Text modality offers an intuitive and straightforward way to control user-desired facial expressions. Using the text description, corresponding facial expressions can be generated in harmony with lip movements controlled according to the input audio. To effectively control the lip movements and facial expressions based on both audio and text inputs, we introduce a two-stage approach for generating expressive speech animations. First, an auto-encoder-based network is trained to disentangle content and expression features from the 3D animation data. This disentangled representation allows us to derive accurate lip sync directly from the audio while simultaneously generating accompanying expressions from the text input. Second, two transformer-based networks are trained to predict the content and expression features from the audio and text

descriptions, respectively, to produce the final expressive speech animation. Given the difficulty in obtaining paired datasets of audio, text, and 3D facial animation required for training our system, we introduce a method to create pseudo training pairs by utilizing a talking head video dataset. After reconstructing 3D facial animations from 2D videos, we utilize Dynamic Time Warping (DTW) in the first stage to temporally align the facial animations with the same speech content but different expressions, as well as animations with different speech content but the same expression. In the second stage, we generate paired text descriptions for the 3D facial animations by leveraging Large Language Models (LLMs) and the ground truth 2D video data.

Our contributions can be summarized as follows:

- We propose a novel method to generate audio-driven speech animation that accompanies with user-desired facial expressions. The expressions are produced from text descriptions unseen during training.
- By disentangling content and expression features from the animation data, our method can generate accurate lip movements and facial expressions from audio and text inputs, respectively.
- By creating pseudo training pairs of audio, text and animation, we can effectively train our models with a limited data.

2. Related Work

2.1. Audio-Driven Facial Animation Generation

Recent research on generating speech animation from audio has been actively pursued. A deep-learning-based approach [TKY*17] was proposed to estimate the shape and appearance parameters of the Active Appearance Model (AAM) corresponding to input audios. VisemeNet [ZXL*18] was introduced to predict viseme and coarticulation rig parameters from audio by training LSTM-based models. These studies focused on generating lower face movements while neglecting upper face expressions. MeshTalk [RZW*21] was proposed to generate full facial animations by leveraging a categorical latent space to reconstruct upper face expressions independently of the audio and to generate accurate lip movements according to the audio. Several studies [CBL*19, FLS*22, XXZ*23] utilized one-hot conditional input to apply various speaking styles to the facial animations. To produce personalized speaking styles, Imitator [THA*23] learns identity-specific details from a short reference video. Although these methods can generate full facial animations with specific speaking styles, generating emotional expressions in speech animation remains a challenge.

Several studies have focused on generating expressive facial animations for audio-driven speech animation. An end-to-end convolutional network [KAL*17] was trained to produce a full facial animation by learning additional trainable latent emotion features to control the expression. After training, the facial expression can be achieved by adjusting this trained emotion feature vector. Other studies extract these emotion features from the input audio [CCK*21, PWS*23, LLSL24]. For instance, Emotalk [PWS*23] takes audio, emotion level, and personal speaking style as inputs to generate emotional speech animation. To extract the emotion latent feature, they disentangle emotion and content latent features from the raw speech audio.

Another approach to producing expressive speech animation uses emotion class labels as a condition [DCT*23, LLZP24]. EMOTE [DCT*23] was proposed to train the model with disentanglement losses using predicted facial animations given audios and emotion labels. Although these studies enable the generation of emotional speech animation, the resulting expressions are confined to one of the emotion classes used during training. Additionally, facial expressions cannot be always classified into one emotion, because people often show mixed emotions simultaneously in one facial expression. Furthermore, there are some facial expressions that do not result from emotions. Our method takes text descriptions as input, enabling to generate user-desired expressions with lip movements synchronized to an input audio.

2.2. Text-Driven Content Generation

Text is a powerful and intuitive medium to generate visual content, making it an invaluable tool in various fields. Contrastive Language-Image Pre-training (CLIP) [RKH*21] is a large-scale visual-textual embedding model. The embedded rich semantic latent representations enable the generation and editing of 2D images, videos, and even 3D models and animations based on text descriptions. Many previous studies [PWS*21, NDR*21, FSW22, KY22, CBK*22, KKY22, GPM*22] leverage CLIP to synthesize and manipulate images from text. For example, StyleCLIP [PWS*21] is an image manipulation method that combines the capabilities of StyleGAN [KLA*20] with CLIP. StyleGAN-NADA [GPM*22] introduces a text-driven generative model that shifts images to new domains. In the field of 3D character animation, text-to-motion generation is actively studied [YJYO22, HZP*22, TGH*22]. CLIP-Actor [YJYO22] synthesizes textured human motion from text descriptions. AvatarCLIP [HZP*22] generates text-driven motion using CLIP-guided losses with rendered poses. MotionClip [TGH*22] generates 3D human motion from text descriptions by aligning the human motion manifold to CLIP space. In this work, we also leverage the semantic power of CLIP space to generate audio-driven facial animations from text descriptions. Similar to our work, ExpCLIP [ZWYW24] utilizes text descriptions to generate audio-driven facial animations. Unlike their resulting expressions that are often influenced by the emotion of the input audio, the content and expression features of our outputs are disentangled. This allows us to control the facial expressions more accurately according to the text description.

3. Overview

Our goal is to generate expressive speech animations from audio and text description. Specifically, we aim to derive accurate lip movements from the audio while controlling the full facial expressions via text descriptions. To this end, we propose a novel two-stage approach. The first stage learns disentangled representations of content and expression features from facial animation data by training an auto-encoder. The decoder of the auto-encoder combines the disentangled features into a coherent facial animation. In the second stage, we freeze this pre-trained decoder while training separate transformer-based networks to generate the content and expression representations from audio and text inputs, respectively.

Taking advantage of the disentangled representation enables deriving accurate lip movements from an audio while controlling the full facial expressions via natural language text descriptions during the animation synthesis process. In the following sections, we first describe how to obtain training data in Section 4, and then explain each stage in detail in Sections 5 and 6.

4. Data Preparation for Training

4.1. Facial Animation Data

To train the networks in both stages, high-quality facial animation data are required. To collect 3D emotional speech animations, we leverage a comprehensive talking head video dataset, MEAD [WWS*20]. MEAD is a high-quality audio-visual dataset that includes emotional talking head videos of 60 actors and actresses speaking in English. The dataset covers eight different emotions (angry, disgust, contempt, fear, happy, sad, surprise, and neutral) at three intensity levels, excluding neutral emotion. From the MEAD dataset videos, we reconstructed 3D facial animations using EMOCA [DBB22], which provides accurate 3D facial reconstructions following the FLAME [LBB*17] structure. Although our method can predict FLAME parameters for face representation, we opted to use vertex positions, resulting in more accurate mouth shapes and expressions. Consequently, the networks in both stages are trained with the reconstructed facial animation data and output the vertex positions of a 3D face model.

4.2. Audio and Text Data

While the auto-encoder of the first stage is trained using only facial animation data, the networks in the second stage require a dataset comprising 3D facial animation, audio, and text. To obtain audio and text descriptions paired with the facial animation, we utilize the videos used to reconstruct the 3D facial animations. While extracting audio data from the video sequences is straightforward, collecting text data describing facial expressions poses a significant challenge. To overcome this, we leverage the visual understanding capabilities of Large Language Models (LLMs) [GOO23]. For each video sequence, we randomly selected five continuous 15-frame segments and asked two questions for each segment, collecting a total of 10 sentences per video sequence. First, we prompted the LLMs to describe the facial expressions of the speaker in the video. Second, we asked the LLMs to describe the speaker's emotions.

The text derived from the speaker's face can be influenced by the speaker's inherent facial characteristics. For example, a person with upturned lip corners may appear happy, or a person with deep furrows between the brows may seem angry, even in the neutral state. To mitigate such speaker-specific influences, we utilized the ground truth emotion labels from the MEAD dataset. Specifically, we collected various sentences by prompting the LLMs to describe the emotional representations of given labels. Additionally, we observed that training the network solely on sentence-type texts did not yield satisfactory results when word-type texts were given. Therefore, we also collected similar words for each ground truth emotion label to ensure that the model could handle both sentences and individual words effectively. Examples of the collected sentences and words are included in the supplementary material.

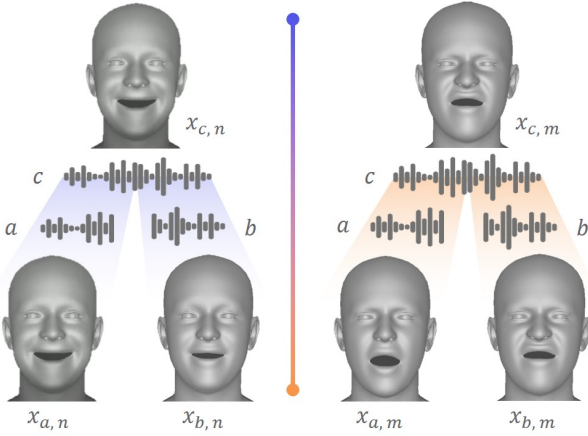


Figure 2: Four animation sequences with identical length for training the content-expression disentanglement network

5. Stage 1: Content and Expression Disentanglement

To effectively control facial expressions during speech animation, we disentangle content and expression features from a facial animation. Expressions influence both the upper face and lower face; the lip corners rise when smiling. Therefore, it is crucial to disentangle the latent representations that affect content and expression, instead of simply dividing the face into upper and lower regions. Unlike previous studies on producing emotional speech animations [PWS*23, DCT*23], we propose a novel method for disentangling content and expression using facial animation data itself. This disentanglement mechanism ensures to produce accurate mouth shapes corresponding to the input audio and appropriate facial expressions based on the text description, allowing our models to predict these features with high precision from both audio and text data.

5.1. Crossing Pairs of Facial Animation Sequences

To learn the disentangled representations of content and expression from facial animation data, we require crossing pairs of facial animation sequences with the same speech content but different expressions, as well as sequences with different speech content but the same expression. Unfortunately, obtaining real-world data of such pairings with high precision is extremely challenging, as it is nearly impossible for a performer to utter the same phrase in the identical timing while conveying different expressions.

To address this issue, we leverage Dynamic Time Warping (DTW) [BC94] to temporally align facial animation data with the same spoken content but different facial expressions. While previous work [JZW*21] applied DTW to audio, we warp the facial animation sequences, particularly focusing on the lip vertices. This enables us to obtain crossing pairs with different expressions in perfectly synchronized speech articulation timings. Specifically, given two animation sequences $X_{c,n}$ and $X_{c,m}$ speaking the same content c with different expressions n and m , the pair of animation sequences

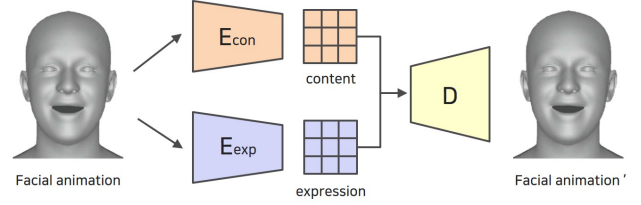


Figure 3: First stage: learning the disentangled representations of content and expression from facial animation data

$(X_{c,n}, X_{c,m})$ will have an identical length after aligning the speech content timings via DTW.

However, this approach alone cannot provide crossing pairs of different speech content under the same facial expression. By utilizing an animation dataset in which each sequence maintains consistent emotional expressions without dramatic changes, we split each aligned sequence pair with the same content in half, treating each split as a pseudo pair with the same expression but different speech content. As a result, four types of pseudo training pairs with uniform timing are collected: $(X_{a,n}, X_{b,n}, X_{a,m}, X_{b,m})$, where a and b represent spoken content split from content c as shown in Figure 2. These four crossing pairs are used for training the content-expression disentanglement network.

5.2. Training Disentanglement Auto-encoder

As shown in Figure 3, we employ a transformer-based auto-encoder architecture to learn the disentangled representations of speech content and expression from facial animation data. The encoder consists of two branches: a content encoder, E_{con} , which extracts content features, and an expression encoder, E_{exp} , which yields expression features. The decoder, D , takes the concatenation of these features to reconstruct the facial animation sequence.

The auto-encoder is trained using four crossing pairs $(X_{a,n}, X_{b,n}, X_{a,m}, X_{b,m})$ from the pseudo training dataset, with a loss function composed of four terms: self reconstruction loss, cross reconstruction loss, content loss, and expression loss. The self reconstruction loss, \mathcal{L}_{self} , is computed by measuring the L_2 distance between the reconstructed facial animation and the original facial animation, as follows:

$$\mathcal{L}_{self} = \|D(E_{con}(X_{a,n}), E_{exp}(X_{a,n})) - X_{a,n}\|_2^2 + \|D(E_{con}(X_{b,m}), E_{exp}(X_{b,m})) - X_{b,m}\|_2^2. \quad (1)$$

A cross reconstruction loss, \mathcal{L}_{cross} , is calculated by swapping the expression features between pairs. For example, taking the content feature $E_{con}(X_{a,n})$ from $X_{a,n}$ and the expression feature $E_{exp}(X_{b,m})$ from $X_{b,m}$ to reconstruct a pseudo ground truth animation $X_{a,m}$. The cross reconstruction loss is defined as follows:

$$\mathcal{L}_{cross} = \|D(E_{con}(X_{a,n}), E_{exp}(X_{b,m})) - X_{a,m}\|_2^2 + \|D(E_{con}(X_{b,m}), E_{exp}(X_{a,n})) - X_{b,n}\|_2^2. \quad (2)$$

The third and fourth loss terms involve triplet losses applied to the

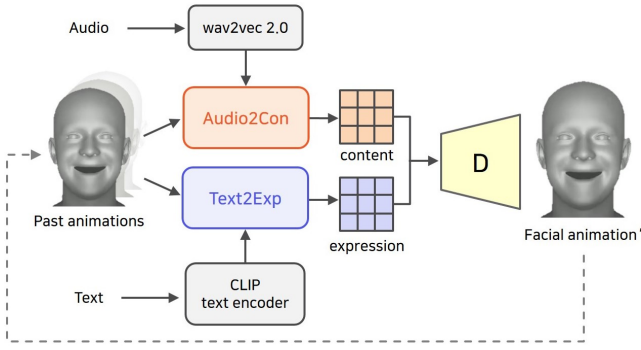


Figure 4: Overview of our system (second stage)

embedding features of content and expression. These losses aim to ensure that animation sequences with similar content or expression are closely grouped together in the latent space, while sequences with dissimilar content or expression are pushed apart. Specifically, given three animation sequences comprising an anchor, a positive, and a negative sequence, the triplet loss is designed to minimize the distance between the anchor and the positive sequence while maximizing the distance between the anchor and the negative sequence, with a specified margin value. The margin value guarantees a minimum distance difference between the anchor-positive pair and the anchor-negative pair, ensuring the negative sequence is at least a margin distance farther from the anchor than the positive sequence. This enforces a clear separation between the distributions of the animation sequences with similar content/expression and sequences with dissimilar content/expression, enhancing the model’s ability to distinguish between them. The content loss, \mathcal{L}_{con} , is represented as follows:

$$\mathcal{L}_{con} = \sum \max(0, \|E_{con}(X_{anc}) - E_{con}(X_{pos})\|_2^2 - \|E_{con}(X_{anc}) - E_{con}(X_{neg})\|_2^2 + \mathcal{M}), \quad (3)$$

where facial animation sequences X_{anc} and X_{pos} have the same content and X_{neg} has different content. The margin value \mathcal{M} is set to 1 in our experiment. Similarly, the expression loss, \mathcal{L}_{exp} , is represented as follows:

$$\mathcal{L}_{exp} = \sum \max(0, \|E_{exp}(X_{anc}) - E_{exp}(X_{pos})\|_2^2 - \|E_{exp}(X_{anc}) - E_{exp}(X_{neg})\|_2^2 + \mathcal{M}), \quad (4)$$

where facial animation sequences X_{anc} and X_{pos} represent the same nuanced expression and X_{neg} shows a different expression.

The disentanglement auto-encoder is trained by minimizing the total loss function:

$$\mathcal{L}_{first} = \lambda_{self} \mathcal{L}_{self} + \lambda_{cross} \mathcal{L}_{cross} + \lambda_{con} \mathcal{L}_{con} + \lambda_{exp} \mathcal{L}_{exp}, \quad (5)$$

where λ_{self} , λ_{cross} , λ_{con} , and λ_{exp} represent the weights for each loss. By optimizing this loss, the auto-encoder learns to disentangle content and expression feature in a self-supervised manner using facial animation data.

6. Stage 2: Audio-Text Driven Speech Animation Generation

Following the training of the disentanglement auto-encoder in the initial stage, we utilize the pre-trained decoder to train two new networks, namely Audio2Con and Text2Exp, in the subsequent stage. An overview of our system is depicted in Figure 4. The Audio2Con and Text2Exp predict the content and expression features from the audio and text description, respectively. These predicted features are concatenated and used as input to the pre-trained decoder, generating an expressive speech animation aligned with the provided audio and text input.

6.1. Training Audio2Con and Text2Exp

Using the collected pseudo-paired dataset of audio, text, and 3D facial animation, we train Audio2Con and Text2Exp in the second stage. The audio features are extracted using wav2vec 2.0 [BZMA20], a pre-trained speech model based on a large-scale corpus. For the text features, we utilized a CLIP text encoder [RKH*21], which effectively embeds text representations. Both Audio2Con and Text2Exp are transformer decoder-based models trained auto-regressively. They take the previously generated animations conditioned on the audio and text features as input to predict the content and expression features for the current frame. These predicted features are passed through the frozen decoder, which was pre-trained in the first stage, to generate expressive speech animations.

As shown in Figure 5, Audio2Con and Text2Exp are trained with two loss functions: facial animation loss and feature embedding loss. The facial animation loss, \mathcal{L}_{face} , is the primary loss term that computes the mean squared error between the generated facial animation \hat{X}^t and the ground truth animation X^t at frame t :

$$\mathcal{L}_{face} = \|\hat{X}^t - X^t\|_2^2 \quad (6)$$

$$\hat{X}^t = D(\text{Audio2Con}(X^{1:t-1}, \mathcal{A}), \text{Text2Exp}(X^{1:t-1}, \mathcal{T})),$$

where \mathcal{A} and \mathcal{T} represent the audio and text feature, respectively.

Additionally, we employ a feature embedding loss, \mathcal{L}_{emb} , which encourages the predicted content and expression features to match the corresponding disentangled representations from the content and expression encoders pre-trained in the first stage. Specifically, we calculate the $L1$ distance between the predicted content and expression features and the corresponding disentangled features from the generated facial animation \hat{X}^t , as follows:

$$\mathcal{L}_{emb} = \|\text{Audio2Con}(X^{1:t-1}, \mathcal{A}) - E_{con}(\hat{X}^t)\|_1 + \|\text{Text2Exp}(X^{1:t-1}, \mathcal{T}) - E_{exp}(\hat{X}^t)\|_1. \quad (7)$$

Audio2Con and Text2Exp are trained by minimizing the total loss function:

$$\mathcal{L}_{second} = \lambda_{face} \mathcal{L}_{face} + \lambda_{emb} \mathcal{L}_{emb}, \quad (8)$$

where λ_{face} and λ_{emb} represent the weights for each loss term.

7. Experiments

In this section, we first provide implementation details of our training setup. Subsequently, we present the results generated by our

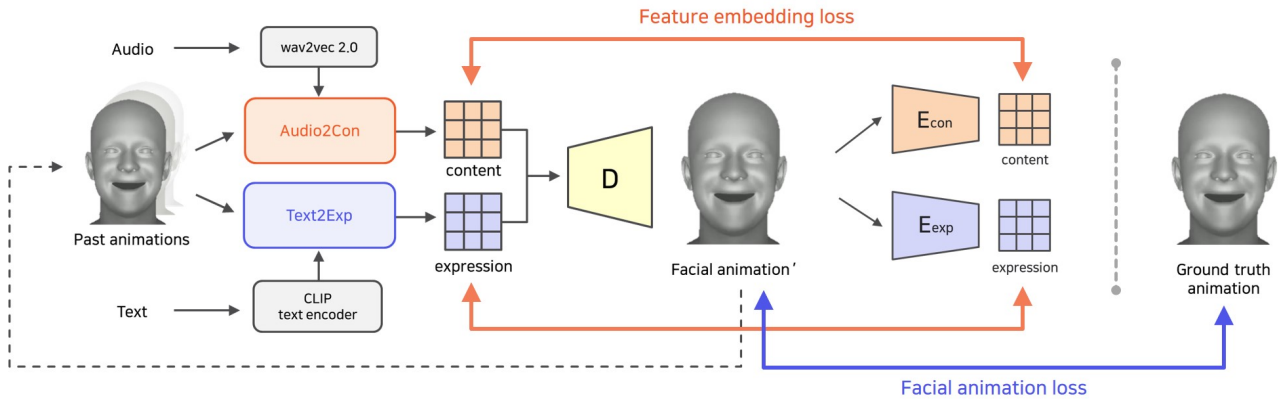


Figure 5: Training in the second stage. *Audio2Con* and *Text2Exp* are trained with two losses: facial animation loss and feature embedding loss.

method from both audio and text descriptions. The resulting animations are available in the supplementary video. We conducted an ablation study to validate the design choices underpinning our training methodology. Finally, we compared our method with state-of-the-art methods through both qualitative and quantitative analyses, including a user study.

7.1. Implementation Details

We employed the emotional talking head video of an actor (M003) from the MEAD dataset. The reconstructed facial animation data consists of a total of 66,666 frames, amounting to 37 minutes at 30 FPS. We split the data into two groups; 85% training set and 15% test set. In the first stage, we trained a transformer-based auto-encoder with four layers. This training process required 1 hour and 2000 epochs, using a learning rate of $1e-5$. In the second stage, we trained two transformer decoder-based models, *Audio2Con* and *Text2Exp*, each of which has four layers. The training required 23 hours over 2000 epochs with a learning rate of $1e-4$. For both stages, the Adam optimizer [KB14] was used with hyperparameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$ on a single Nvidia Tesla V100 GPU. The loss weights were configured as follows: $\lambda_{self} = 1$, $\lambda_{cross} = 1$, $\lambda_{con} = 1e-6$, $\lambda_{exp} = 1e-6$, $\lambda_{face} = 1$, and $\lambda_{emb} = 1e-4$. The disparity in the scale of the weights is due to the small scale of the 3D face model, which results in minimal differences between the predicted and ground truth animations. Consequently, the weights for reconstruction are relatively large compared to those for embedding losses. The inference process is identical to the second stage, generating facial animations from the provided audio and text descriptions. This process operates at 80.3 FPS, ensuring that the pipeline is capable of real-time performance.

7.2. Results of Our Method

Figures 1 and 6 show the animations generated from given audio and text descriptions. The expressions can be synthesized from various types of text descriptions, ranging from single emotion words to complete sentences. For example, the audio source utilized to

test the model used for the results presented in Figure 6 was from a test set, M003_003 in the MEAD dataset. Although the original emotion label of the input audio was "angry" with the maximum emotional intensity of 3, the resulting expressions were generated according to the input text descriptions, demonstrating the model's ability to disentangle content and expression features effectively. Notably, all the results exhibit similar lip movements corresponding to the input audio, ensuring that both the lip movements and facial expressions are accurately generated from the input audio and text description, respectively.

Figure 7 shows an animation produced by interpolating the text features from "Happy" to "Disgusted". All five faces were captured at the same moment of pronouncing the same phoneme. The transition of facial expressions from happy to disgusted demonstrates that our method can effectively interpolate text descriptions. This capability makes our system versatile and adaptable, enabling it to generate a wide range of facial expressions from various text descriptions.

7.3. Ablation Study

To validate the design choices made for the training of our system, we conducted ablation studies. We compared the resulting animations of our full model with those generated by five ablated versions both quantitatively and qualitatively. The first variant model ("w/o cross loss") was trained without the cross reconstruction loss at the first stage. For the second variant model ("w/o triplet loss"), the auto-encoder of the first stage was trained without the triplet losses applied to the embedding features of content and expression. The third variant model ("w/o embedding loss") was trained without the feature embedding loss at the second stage. The fourth variant model ("End-to-end learning") was trained directly from the second stage without training the first stage. Because the model did not use the pre-trained decoder, all three networks, including *Audio2Con*, *Text2Exp*, and the decoder of the first stage, were trained together. For the last variant model ("Decoder fine-tuning"), the de-

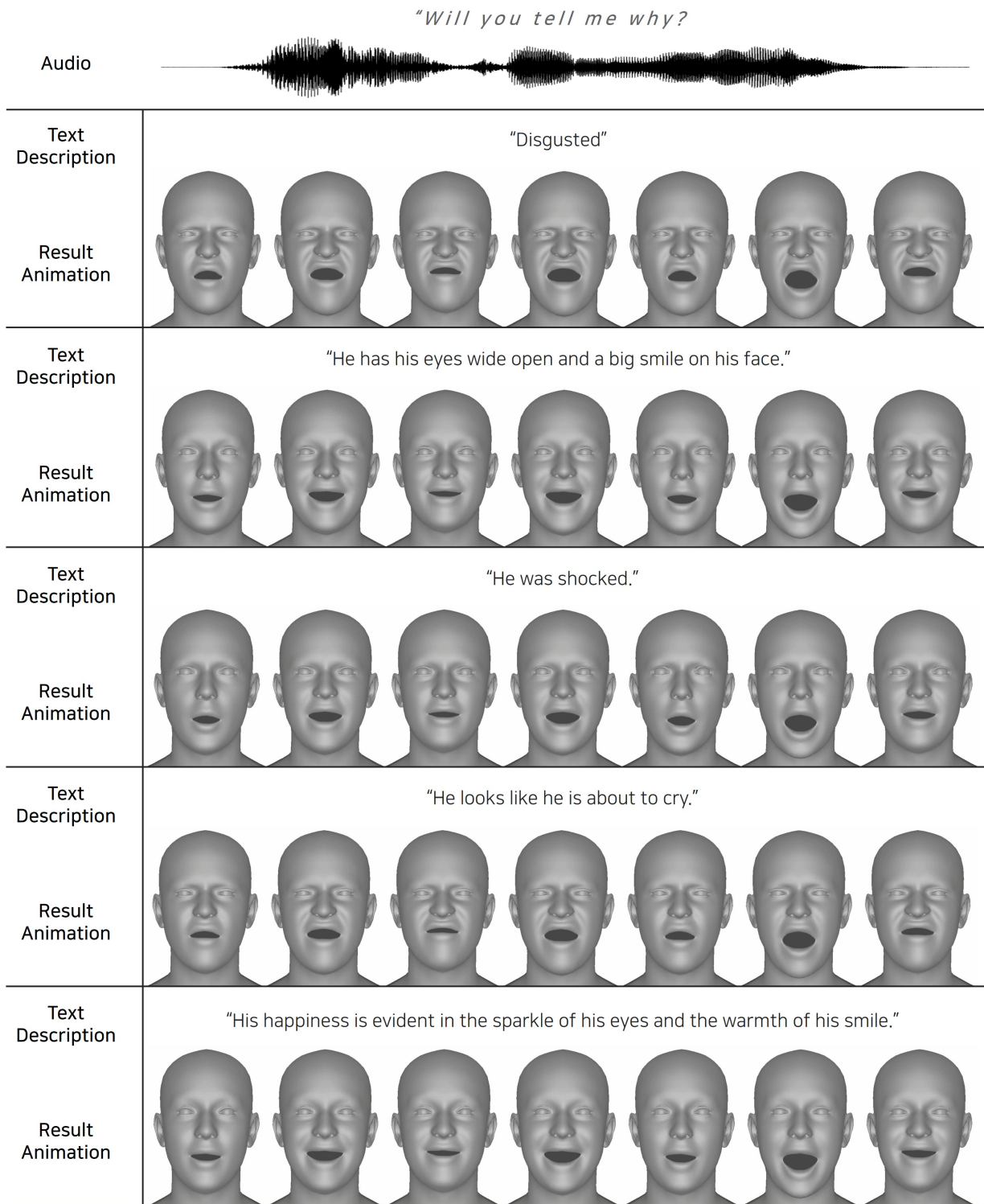


Figure 6: Resulting animations produced from a given audio and various test descriptions

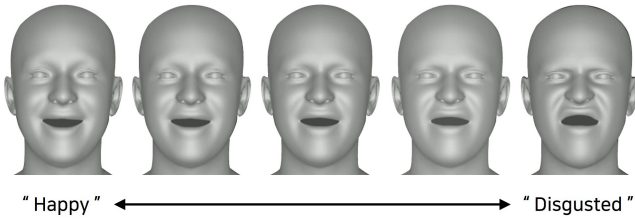


Figure 7: Transition of facial expressions from "Happy" to "Disgusted".

Table 1: Comparison of vertex errors ($\times 10^{-4}$ mm) produced by the variant models and our model

| Methods | MAX | | MSE | |
|---------------------|---------------|---------------|---------------|---------------|
| | Lip | Face | Lip | Face |
| w/o cross loss | 6.535 | 7.2563 | 2.7663 | 0.3450 |
| w/o triplet loss | 8.685 | 9.6641 | 3.7524 | 0.4538 |
| w/o embedding loss | 7.4054 | 8.1153 | 3.2498 | 0.3696 |
| End-to-end learning | 7.3601 | 8.0444 | 3.1837 | 0.3637 |
| Decoder fine-tuning | 8.3314 | 9.2620 | 3.6347 | 0.4327 |
| Ours | 6.3778 | 7.0267 | 2.6984 | 0.3131 |

coder was fine-tuned from the pre-trained weights, and all three networks were jointly trained in the second stage.

Quantitative evaluation. To quantitatively evaluate the facial animations generated by the variant models and our model, we calculated the vertex errors between the produced animations and the ground truth animations. Table 1 shows the measured errors using two metrics: the maximal L2 error ("MAX") and the mean squared error ("MSE"). The maximal L2 error, a conventional metric used in audio-driven speech animation generation methods [FLS*22, XXZ*23, RZW*21], was computed as the average value of the maximum errors of the vertices at each frame. We also computed the mean squared error to show the overall average vertex errors. Because we generate lip movements from an input audio and expressions from a text, we separately calculated the errors for the lip vertices and the full face. As shown in Table 1, our model ("Ours") produced the lowest errors for both metrics, indicating that our design choices improve the overall quality of the resulting facial animations, including accurate lip movements and facial expressions.

Qualitative evaluation. Figure 8 shows two facial animations (a) and (b) generated by five variant models and our model. We used an audio (M003_001 in the MEAD dataset) from the test set for both animations. The only difference is the input text descriptions used to generate the facial animations: "Angry" for (a) and "Surprised" for (b). Note that only animation (a) has a ground truth reference, because the emotion label of the test audio was angry. The facial expressions generated by all models appear to be angry in animation (a) because both the emotion of the test audio and the text description are angry. For the facial expressions in anima-

Table 2: Comparison with state-of-the-art methods by measuring the vertex errors ($\times 10^{-4}$ mm)

| Methods | MAX | | MSE | |
|---------------------|---------------|---------------|---------------|---------------|
| | Lip | Face | Lip | Face |
| FaceFormer [FLS*22] | 7.5200 | 8.3395 | 3.3220 | 0.3911 |
| CodeTalker [XXZ*23] | 9.5152 | 10.3834 | 4.2759 | 0.4813 |
| ExpCLIP [ZWYW24] | 9.5500 | 10.8637 | 4.1220 | 0.5074 |
| Ours | 6.3778 | 7.0267 | 2.6984 | 0.3131 |

tion (b), only the results of the variant model ("w/o cross loss") and our full model ("Ours") show a surprised expression. Between the two, our model generated more natural expressions and accurate lip shapes than the model trained without the cross reconstruction loss, as demonstrated by the lowest vertex errors reported in Table 1. The variant model ("w/o triplet loss") shows a happy expression, indicating that the model struggled to differentiate embedding features in the latent space accurately. The other three variant models produced angry expressions that reflect the emotion of the given audio, instead of the intended surprised expression. This indicates that all the components of our model are crucial for generating accurate expressions and lip shapes corresponding to the input text descriptions and audio.

7.4. Comparison with State-of-the-Art Methods

We conducted both qualitative and quantitative comparisons between our method and state-of-the-art methods. This section primarily focuses on comparisons with three prior methods: FaceFormer [FLS*22], CodeTalker [XXZ*23], and ExpCLIP [ZWYW24]. In the supplementary video, readers can see the comparison results with these three methods as well as with other previous studies [KAL*17, TKY*17, CBL*19, RZW*21]. Because the target characters and datasets differ, we trained and tested the previous methods using the same dataset as our system. For FaceFormer and CodeTalker, we used the publicly available official implementations [†]. Because the implementation of ExpCLIP is not publicly available, we implemented ExpCLIP to the best of our understanding based on the paper. The implementation details of ExpCLIP are provided in the supplementary material.

Quantitative evaluation. To quantitatively compare the quality of the facial animations generated by state-of-the-art methods and our method, we calculated the vertex error. Table 2 presents the averaged vertex error, including the maximal L2 error and mean squared error, for both the lip region and full face region over the test dataset. Because ExpCLIP and our model require an additional text input to generate a facial animation, we used the original emotion labels as the text descriptions. Our model ("Ours") produced the lowest vertex errors for both metrics. The lowest lip vertex error of our model demonstrates its superior lip sync ability compared to previous methods [FLS*22, XXZ*23] that focus more on

[†] FaceFormer: <https://github.com/EvelynFan/FaceFormer>
CodeTalker: <https://github.com/Doubiiu/CodeTalker>

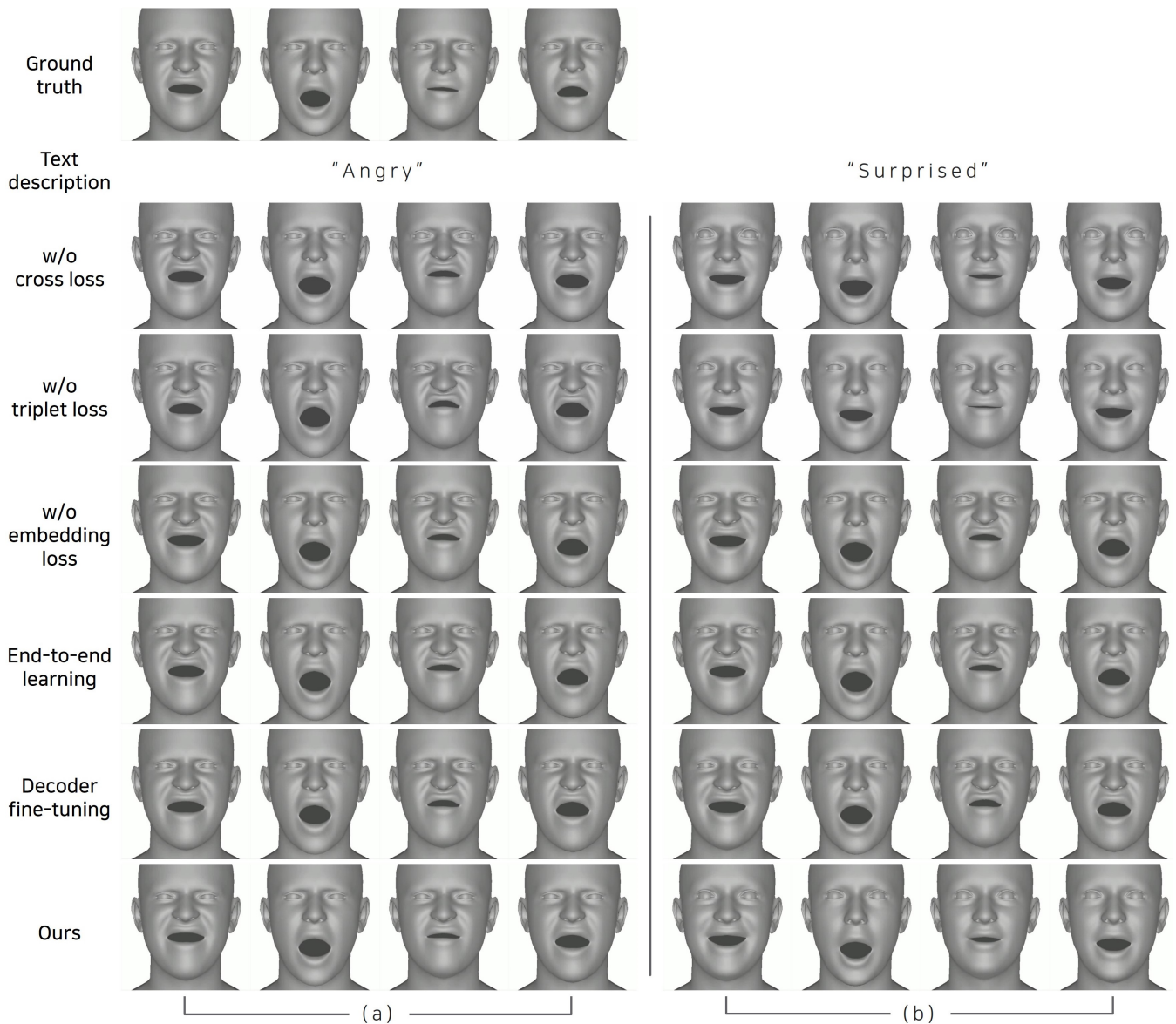


Figure 8: Comparison of facial animations generated by five variant models and our model, using the same audio with text descriptions "Angry" for (a) and "Surprised" for (b).

the accurate reconstruction of lip shapes than the reconstruction of facial expressions. Additionally, the lowest vertex error for the full face region indicates that our model can generate more expressive facial animations compared to ExpCLIP, which aims to create expressive speech animation. The superior performance of our model is due to the effective disentanglement strategy. While the previous studies directly generate full facial vertex positions from audio [FLS*22, XXZ*23] and text descriptions [ZWYW24], simultaneously creating both lip shapes and expressions, our method predicts content and expression features from audio and text descriptions, respectively, resulting in more accurate lip shapes and

expressive facial animations compared to the state-of-the-art methods.

Qualitative evaluation. Figure 9 shows a comparison of the animations produced by ExpCLIP and our model. For the test audio source, we used an audio (M003_011) from the test set in the MEAD dataset and another (dia3_utt8) from a new dataset, the Multimodal EmotionLines Dataset (MELD) [PHM*18], which contains emotional dialogues. Our model generated facial expressions that correspond to the text descriptions more accurately compared to those produced by ExpCLIP. The results from ExpCLIP often show uncorrelated expressions to the given texts. Notably,

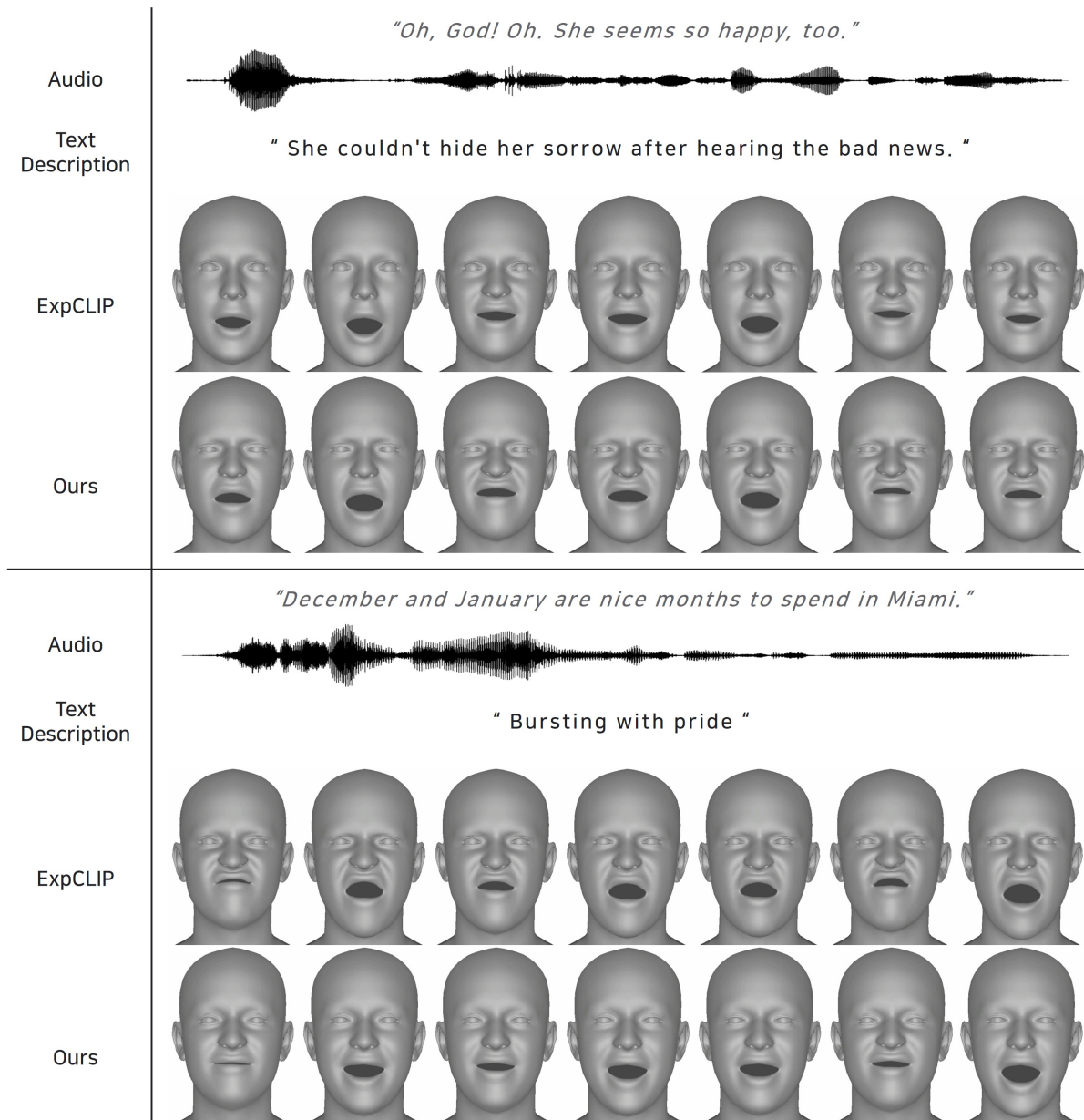


Figure 9: Comparison between the animations produced by ExpCLIP and our model given the same audio and text descriptions as input. The audio source in the upper figure is from the MELD dataset, and the audio source in the lower figure is from the MEAD dataset.

the facial expressions generated by ExpCLIP were frequently influenced by the emotion of the given audio. For example, in the lower figure, ExpCLIP produced a disgust expression, which was the original emotion label of the test audio. Our model created appropriate facial expressions from the text descriptions, demonstrating that disentangling content and expression features can enable the control of facial expressions using text inputs. For additional comparisons with FaceFormer, CodeTalker, ExpCLIP, and our model, please refer to the supplementary video.

User Study. Because ExpCLIP is the only previous method ca-

pable of generating facial expressions from text descriptions, we conducted a user study to compare the quality of animations produced by our model and ExpCLIP. We generated 11 facial animations for each method using various audios and text descriptions not included in the training data. A total of 20 participants in the age distribution of 24 to 36 took part in the web-based study; 17 had a background in computer science while three did not. Participants watched 11 side-by-side videos in a randomized order, each accompanied by the same text descriptions used to test the models. They were asked to choose the animation they preferred or select

Table 3: Results of the user study. We present the percentage of participants who preferred our model or ExpCLIP for animation quality based on three criteria.

| | Favorability (%) | | | Ours better or equal |
|--------------------|------------------|-------|-------|----------------------|
| | ExpCLIP | Equal | Ours | |
| Lip sync | 38.64 | 38.64 | 22.73 | 61.36 |
| Facial expression | 16.36 | 11.36 | 72.27 | 83.64 |
| Overall preference | 21.36 | 14.55 | 64.09 | 78.64 |

equal if the quality was similar, based on three criteria: lip synchronization with the audio, facial expression representing the text, and overall preference considering both the audio and text inputs.

Table 3 presents the percentage of votes received for each option (ExpCLIP, equal, and ours) out of the total number of votes. More than half of the participants indicated that our results were better or equal to those of ExpCLIP. While ExpCLIP received more votes for lip sync quality, it often generated exaggerated lip movements influenced by the input audio, which we believe more positively appealed to the participants compared to the results of our model. Despite the fewer votes that our results received, the lip movements were faithful to the input audio, as demonstrated by the lip vertex error in the quantitative evaluation. Additionally, our model was significantly preferred for facial expressions and the overall quality of facial animations. This indicates that our model produced more natural facial animations with well-coordinated lip movements and facial expressions in response to the input audio and text descriptions, outperforming the previous method.

8. Conclusion

In this work, we propose a novel method for generating expressive speech animations of a 3D face model driven by both audio and text descriptions. Unlike previous approaches that relied on pre-defined emotion categories, our method leverages text descriptions to generate facial expressions unseen during training, thereby overcoming limitations related to specific emotion classes. To achieve this, we introduce a two-stage approach. First, we pre-train an auto-encoder that can disentangle content and expression features from the facial animation. In the second stage, we train transformer-based networks that can predict these features from audio and text inputs. These predicted features are then decoded to generate the final expressive speech animations using the pre-trained decoder in the first stage. We also introduce an efficient way to train our system with a limited unpaired dataset, by collecting pseudo training pairs of audio, text, and 3D facial animation data. Extensive quantitative and qualitative evaluations, including a user study, demonstrate that our model can effectively generate expressive speech animations corresponding to the input audio and text descriptions whose quality is much higher compared to that of the results produced by previous studies. By accommodating diverse forms of natural language, such as emotion words or detailed facial expression descriptions, our approach offers an intuitive and versatile solution for generating audio-driven expressive speech animations.

8.1. Limitations and Future Work

Despite the promising results of our method for generating expressive speech animations from audio and text descriptions, our method has several limitations. First, the training dataset was limited to data from a single actor and eight distinct emotions, which restricts the range of facial expressions that our model can produce. Expanding the dataset to include more actors and actresses, along with diverse emotional states and facial actions such as winking or puffing cheeks, could significantly enhance the expressiveness and realism of the results. Additionally, incorporating a wider variety of emotional intensities and speaking styles will improve the model's generalizability. Addressing these limitations will further enhance the effectiveness and versatility of our approach in generating realistic and expressive 3D facial animations driven by audio and text descriptions.

Acknowledgements

We thank the anonymous reviewers for their invaluable comments. We also extend our gratitude to all members of the Avatar team at Naver Cloud, with special thanks to Dong-Hyun Hwang and Se Yun Lee for their helpful discussions on the experiments. This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2024-00439499, Generating Hyper-Realistic to Extremely-stylized Face Avatar with Varied Speech Speed and Context-based Emotional Expression)

References

- [BC94] BERNDT D. J., CLIFFORD J.: Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd international conference on knowledge discovery and data mining* (1994), pp. 359–370.
- [BZMA20] BAEVSKI A., ZHOU Y., MOHAMED A., AULI M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems* 33 (2020), 12449–12460.
- [CBK*22] CROWSON K., BIDERMAN S., KORNIS D., STANDER D., HALLAHAN E., CASTRICATO L., RAFF E.: Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Euro-pean Conference on Computer Vision* (2022), Springer, pp. 88–105.
- [CBL*19] CUDEIRO D., BOLKART T., LAIDLAW C., RANJAN A., BLACK M. J.: Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 10101–10111.
- [CCK*21] CHUN S., CHOE D., KANG S., AN S., JO Y., OH I.: Emotion guided speech-driven facial animation. In *SIGGRAPH Asia 2021 Posters*. 2021, pp. 1–2.
- [DBB22] DANECEK R., BLACK M. J., BOLKART T.: EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 20311–20322.
- [DCT*23] DANĚČEK R., CHHATRE K., TRIPATHI S., WEN Y., BLACK M., BOLKART T.: Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers* (2023), pp. 1–13.
- [FLS*22] FAN Y., LIN Z., SAITO J., WANG W., KOMURA T.: Face-former: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18770–18780.

- [FSW22] FRANS K., SOROS L., WITKOWSKI O.: Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems* 35 (2022), 5207–5218.
- [GOO23] GOOGLE: Gemini pro vision. <https://console.cloud.google.com/vertex-ai/publishers/google/model-garden/gemini-pro-vision>, 2023.
- [GPM*22] GAL R., PATASHNIK O., MARON H., BERMANO A. H., CHECHIK G., COHEN-OR D.: Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.
- [HZP*22] HONG F., ZHANG M., PAN L., CAI Z., YANG L., LIU Z.: Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535* (2022).
- [JZW*21] JI X., ZHOU H., WANG K., WU W., LOY C. C., CAO X., XU F.: Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021), pp. 14080–14089.
- [KAL*17] KARRAS T., AILA T., LAINE S., HERVA A., LEHTINEN J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–12.
- [KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [KKY22] KIM G., KWON T., YE J. C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 2426–2435.
- [KLA*20] KARRAS T., LAINE S., AITTALA M., HELSTEN J., LEHTINEN J., AILA T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 8110–8119.
- [KY22] KWON G., YE J. C.: Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 18062–18071.
- [LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.
- [LLSL24] LEE S., LEE J., SONG H., LEE S.: Speech-driven emotional 3d talking face animation using emotional embeddings. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2024), IEEE, pp. 7840–7844.
- [LLZP24] LIU C., LIN Q., ZENG Z., PAN Y.: Emoface: Audio-driven emotional 3d face animation. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)* (2024), IEEE, pp. 387–397.
- [NDR*21] NICHOL A., DHARIWAL P., RAMESH A., SHYAM P., MISHKIN P., MCGREW B., SUTSKEVER I., CHEN M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).
- [PHM*18] PORIA S., HAZARIKA D., MAJUMDER N., NAIK G., CAMBRIA E., MIHALCEA R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).
- [PWS*21] PATASHNIK O., WU Z., SHECHTMAN E., COHEN-OR D., LISCHINSKI D.: Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision* (2021), pp. 2085–2094.
- [PWS*23] PENG Z., WU H., SONG Z., XU H., ZHU X., HE J., LIU H., FAN Z.: Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 20687–20697.
- [RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PMLR, pp. 8748–8763.
- [RZW*21] RICHARD A., ZOLLHÖFER M., WEN Y., DE LA TORRE F., SHEIKH Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 1173–1182.
- [TGH*22] TEVET G., GORDON B., HERTZ A., BERMANO A. H., COHEN-OR D.: Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision* (2022), Springer, pp. 358–374.
- [THA*23] THAMBIRAJA B., HABIBIE I., ALIAKBARIAN S., COSKER D., THEOBALT C., THIES J.: Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 20621–20631.
- [TKY*17] TAYLOR S., KIM T., YUE Y., MAHLER M., KRAHE J., RODRIGUEZ A. G., HODGINS J., MATTHEWS I.: A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–11.
- [WWS*20] WANG K., WU Q., SONG L., YANG Z., WU W., QIAN C., HE R., QIAO Y., LOY C. C.: Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision* (2020), Springer, pp. 700–717.
- [XXZ*23] XING J., XIA M., ZHANG Y., CUN X., WANG J., WONG T.-T.: Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12780–12790.
- [YJYO22] YOUWANG K., JI-YEON K., OH T.-H.: Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *European Conference on Computer Vision* (2022), Springer, pp. 173–191.
- [ZWYW24] ZHONG Y., WEI H., YANG P., WANG Z.: Expclip: Bridging text and facial expressions via semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2024), vol. 38, pp. 7614–7622.
- [ZXL*18] ZHOU Y., XU Z., LANDRETH C., KALOGERAKIS E., MAJI S., SINGH K.: Visemenet: Audio-driven animator-centric speech animation. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–10.