Journal of Multimedia Information System VOL. 13, NO. 1, March 2026 (pp. 1-12): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2026.13.1.1

Personalized Speech-Driven Emotional 3D Talking Face Animation via Context–Emotion Decoupling

Seongmin Lee[1*]

Abstract

When people speak with emotions, facial motions are coupled with both context-driven lip motions and emotion-driven facial expressions. However, existing speech-driven emotional 3D facial generation methods aim to generate directly emotional talking faces from speech, without decoupling these motions. It increases the ambiguity during training regarding whether the current facial motion is related to the context or emotion, hindering effective learning. To decouple them effectively, we introduce the contrastive emotion-face loss that learns the mapping between emotions of speech and expressions in talking lip motion. In addition, since individuals have unique emotional expression styles, we enable personalized emotional expression by utilizing person-specific embeddings. By optimizing the person-specific embeddings, the proposed method can generate emotional talking faces personalized to the target subject. At last, the proposed method generates head motions that align with the context and emotion of speech while maintaining diversity. Through extensive experiments, it is demonstrated that our method achieves 7.18% performance improvement over state-of-the-art methods by animating emotional 3D talking facial animation.

Key Words: 3D Talking Face, Motion Decoupling, Personalized Emotion, Head Animation.

I. INTRODUCTION

Talking 3D face generation has become an increasingly attractive technology in growing demand of applications such as virtual reality (VR) and augmented reality (AR) communications, telepresence systems, the film industry, and computer games [1-3]. Dynamic facial motions, including emotional expressions and head motions, are core factors in effectively conveying a user's emotions, enhancing their immersion in such virtual communications. Such a dynamic and realistic 3D talking face animation has great potential in applications leveraged by intimacy with virtual avatars, such as chatbots, education, healthcare, and counseling. This is because humans are particularly sensitive to the subtle nuances when they see the faces of human-like objects such as 3D characters.

Facial expression and head motion serve as nonverbal social cues, playing an essential role in conveying contextual and emotional intentions to the audience. In absence of those visual dynamics in speech-driven 3D facial animation, a user might feel uncomfortable resulting in an uncanny valley effect. Consequently, a lack of such natural visual stimulus would make users perceive a degradation in visual quality during virtual communication. Therefore, to make the content more realistic, it is essential to animate talking faces with emotional expressions. The emotionless talking face is generated by using the CodeTalker [10], which is a state-of-the-art speech-driven 3D face animation method, and the emotional talking face is generated from our method. It shows that animating emotionless talking faces produces similar outputs even from speeches with different emotions. It causes emotional inconsistency between heard speech and the visualized face, leading to deterioration in perceived quality and diminishing user immersion.

Animating emotional talking faces is a challenging task compared to animating emotionless talking faces. This is because the emotionless talking faces are clearly derived from the context of the speech by using the viseme, which represents the distinct units of visual information that correspond to how sound is formed by mouth, but the emotional talking face animation should consider both context and emotion of the speech. For the emotional talking face animation, several methods have been proposed. 3DTalkEmo [20] and FaceXHuBERT [11] use the emotional label as a condition to the network to synthesize the emotional expressions. They require a manual process to provide the emotional condition. Such a

Manuscript received December 30, 2025; Revised January 27, 2026; Accepted February 14, 2026. (ID No. JMIS-25M-12-087)
Corresponding Author (*): Seongmin Lee, +82-44-863-9266, sm.lee@hanbat.ac.kr
[1]Department of Artificial Intelligence Software, Hanbat National University, Sejong, Korea, sm.lee@hanbat.ac.kr

cumbersome process serves as a significant obstacle regarding accessibility, hindering its application in real-world scenarios. EmoTalk [12] and SDETalk [14] extracts contextual and emotional representations from the speech and generates emotional talking faces by combining them. These methods implicitly learn the mapping from speech representations to emotional talking faces without the explicit separation of context-driven and emotion-driven facial motions.

However, when people talk, the lip movement for articulating context-driven pronunciation and the facial expressions for conveying emotions are inherently coupled. Due to the coupled facial motions from context and emotion, previous methods often struggle to accurately learn whether a current facial motion is more relevant to the context or emotion. This ambiguity might lead to the generation of inaccurate lip motions and unnatural facial expressions. To address this problem, a clearer distinctive characterization of facial motions based on the contextual and emotional aspects of the speech signal is needed. For an accurate and natural emotional talking face animation, we explicitly decouple the context-driven and emotion-driven facial motions according to the contextual and emotional representations of speech. We propose a two-stage approach that explicitly decouples and generates context-driven and emotion-driven facial motions. It enables the network to explicitly learn the context-face and emotion-face mappings during the training. In addition, we introduce the contrastive emotion-face loss inspired by contrastive learning [13,19] for an accurate context-emotion decoupling. The proposed contrastive emotion-face loss maximizes the feature similarity of correct pairs between speech emotion speech and facial expressions while minimizing the feature similarity of incorrect pairs. It enables the network to generate natural expressions that correspond to the current emotional states of speech.

To improve the immersion of the 3D talking face, we also propose to generate head motion simultaneously correlated with the context and emotion representations. Head movement is an important social cue in communication to effectively convey the speaker's intent by maximizing the expressiveness of emotions [6]. For example, nodding the head up and down can indicate positive reactions, while shaking the head from side to side can signify negative reactions. Thus, by incorporating head motion, the realism of the animated talking faces would be enhanced, improving the user experience. Unlike facial expressions, head motions should be consistent with emotions in a broader context, but they do not always move the same way in response to a speech. To establish the head motion diversity, we present a head motion synthesizer that synthesizes the diverse head motions based on the reparameterization trick.

Furthermore, for more effective emotional expression in a test phase, we present a personalized emotional tuning scheme that represents personalized emotional expressions for individuals not used during training. For this, we assume that the minimal audio-visual dataset of target subjects is available. Using this minimal dataset, the proposed personalized emotional tuning allows the network to learn the emotional expression style of the target subject based on the pivotal tuning inversion [5]. The proposed method ensures robustness in emotional talking face generation by preventing overfitting to minimal datasets and effectively learns the personalized emotional expression styles.

II. RELATED WORKS

2.1. Emotionless 3D Talking Face Animation

With the recent advancements in deep learning, learning-based automatic talking facial animation generation without the need for complex equipment has been gaining popularity. VOCA [4] extracts the speech context features from a pre-trained DeepSpeech model [21] and learns the mapping from context features to facial motion, producing only animations of facial lips. MeshTalk [9] adopts a long short-term memory (LSTM) architecture to address temporal face generation from speech sequences and learns a categorical latent space that disentangles the information of facial upper and lower motions. By disentangling upper and lower faces, it produces neutral talking faces with blinking eyes. FaceFormer [8] adopts the transformer architecture to address long-term speech sequences. It autoregressively predicts the next facial motion from previous speech and facial motion. Codetalker [10] employs the vector quantized variational autoencoder (VQ-VAE) and introduces discrete motion priors to generate speech-driven talking faces.

2.2. Emotional 3D Talking Face Animation

To address the emotional discrepancy between audio and visual, several studies aim to produce emotional talking faces from speech. Chen et al. [15] 3DTalkEmo, and FaceXHuBERT utilized the one-hot emotion condition instead of directly extracting the emotional state from the audio to generate the emotional talking faces. Thus, these methods require a manual process to provide the emotional condition. Providing the emotional conditions manually to the network serves as a significant obstacle to automatic face generation, hindering its application in real-world scenarios. EmoTalk decomposed emotional speech into contextual and emotional representations based on the method of Ji et al. [16]. These methods aim to generate emotional 3D talking faces by learning the direct mapping

Journal of Multimedia Information System VOL. 13, NO. 1, March 2026 (pp. 1-12): ISSN 2383-7632 (Online)
https://doi.org/10.33851/JMIS.2026.13.1.1

from the speech to the emotional talking faces without the consideration of decoupling the context-driven and emotion-driven shapes. It leads to an increase in ambiguity during the training, resulting in inaccurate lip motion and unnatural expressions. To address this, we propose a two-stage approach with contrastive emotion-face loss that decouples the facial shapes for context and emotion. Additionally, the proposed method generates both emotional talking faces and head movements simultaneously, providing users with a heightened sense of immersion. This will make it more applicable and versatile for various 3D applications.

2.3. Personalized Facial Animation

Recently, personalization has become an important issue in computer vision. This approach is in line with image manipulation approaches that learn from limited data of a target subject to generate personalized 2D facial images of the target subject. To generate personalized images from limited data of the target subject, a pre-trained generative model is employed, followed by a small amount of fine-tuning. These methods aim to learn the personalized static style, such as eyes, nose, skin color, and hairstyle, to generate the static personalized image. Another approach aims to generate audio-driven personalized talking facial images by learning the unique facial changes that appear when the target subject speaks [17]. Although these approaches have been successful in generating personalized 2D images, they cannot be directly applied in 3D application fields. To address this limitation, we propose a personalized emotional tuning that enables the network can effectively animate personalized 3D talking faces even for unseen subjects during training. We employ the person-specific emotional embeddings trained from the several subjects in the training set as pivots. The person-specific emotional embeddings for a new subject are optimized within a latent space spanned from pivots to effectively capture personalized emotional expressions. A small dataset of a new subject is used for the optimization. Consequently, our approach can produce personalized emotional talking 3D faces for new identities not used in training through personalized tuning.

III. METHOD

We propose an end-to-end personalized speech-driven emotional 3D talking face and head generation framework. Fig. 1 illustrates the overall architecture of the proposed network. In addition, we propose a personalized emotional tuning to animate personalized emotional talking faces of unseen subjects during the training.

3.1. Context Encoder

The goal of the context encoder is to extract the contextual representations from the given speech input $s_{1:M}$. To do this, we employ the state-of-the-art pre-trained speech model, wav2vec2.0 [18]. By leveraging this model, which was pre-trained on a massive, large-scale speech dataset, we can effectively mitigate the limited linguistic diversity of our training set and ensure more robust feature extraction. After the wav2vec2.0 module, we add the linear interpolation layer to sample the contextual representation as a multiple of frame number in face animation. This is because the context representation sampled from the speech might have a different sampling rate f_s compared to that of face animation f_v. Through the linear interpolation layer, we resample the contextual representations to have the data length of kN, where $k = \left\lceil \frac{fc}{fv} \right\rceil$. Thus, the output of the

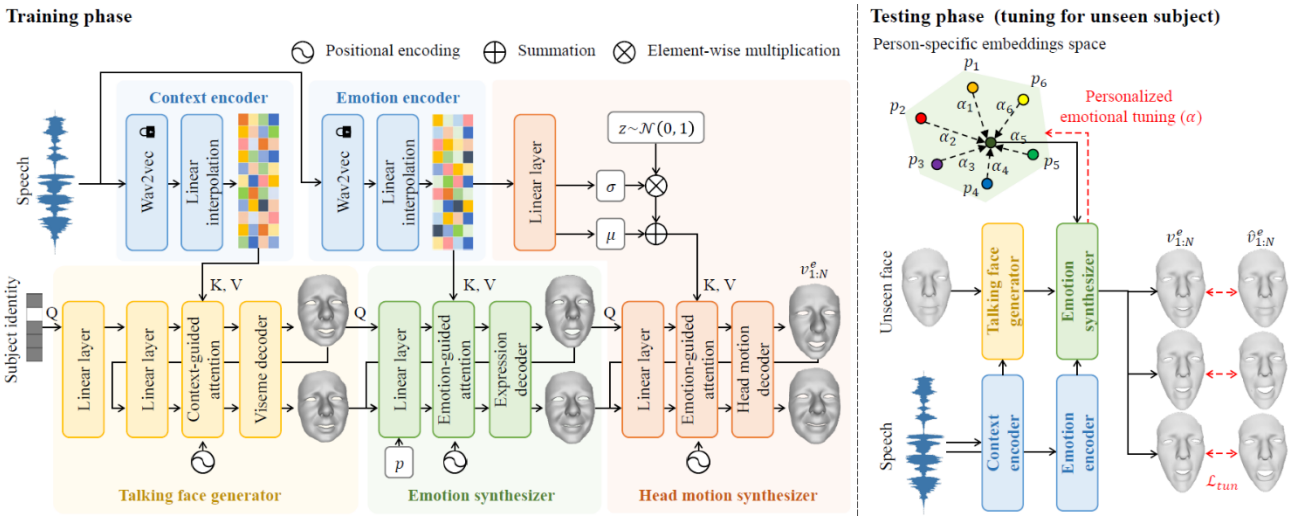

Fig. 1. An overall framework of the proposed speech-driven emotional talking 3D face and head generation (right) and personalized emotional tuning (left).

context encoder $c_{1:kN}$ can be represented as follows:

$$c_{1:kN} = \mathrm{Interp}_{M\to kN}\big(E_c(s_{1:M})\big),$$

where E_c is the pre-trained context encoder and $\mathrm{Interp}_{M\to kN}$ is the linear interpolation operator from M samples to kN samples. We freeze the context encoder during the training.

3.2. Talking Face Generator

The talking face generator is designed based on the transformer to learn the context-driven facial shape, ignoring the emotions of the speech. The one-hot label of subject identity or previous facial shape v_{t-1}^n is used as a query, and the contextual representation $c_{1:kN}$ is used as key and value. For the first frame, a one-hot label of subject identity is used as a query. Using the context-guided attention between the contextual representation of speech and face, the network effectively learns the viseme mapping. Then, the encoded viseme features are decoded to the emotionless talking face through a viseme decoder, which consists of a single linear layer. Thus, the neutral talking face generator G_n is defined as follows:

$$v_{t-1}^n = G_n(v_{t-1}^n \oplus p, c_{1:kN}, B_A),$$

where $t = 1,\dots,N$ is the frame index of face animation, v_t^n is the output neutral talking face of t-th frame. B_A is the alignment bias mask proposed in FaceFormer [9]. The alignment bias mask B_A enables the transformer architecture to effectively learn the query-key attention by aligning the query (i.e., talking 3D faces) and key (i.e., contextual representations). The alignment bias mask $B_A(1 \le i \le t,\ 1 \le j \le kN)$ is defined as follows:

$$B_A(i,j) = \begin{cases} 0, & ki \le j \le k(i+1) \\ -\mathrm{inf}, & \text{otherwise} \end{cases}$$

3.3. Emotion Encoder

Similar to the context encoder, we employ the pre-trained speech emotion recognition (SER) model because the learning-based speech model has achieved superior emotion recognition accuracy. The state-of-the-art SER method [22], which is based on the wav2vec2.0 architecture, is employed as the emotion encoder. Given a speech input $s_{1:M}$, emotional representation $e_{1:M}$ is embedded through the emotion encoder. Same with the contextual representations, we interpolate emotional representations as $e_{1:kN}$ to synchronize the emotional representations and expressions.

$$e_{1:kN} = \mathrm{Interp}_{M\to kN}\big(E_e(s_{1:M})\big),$$

where E_e is the pre-trained emotion encoder with a large-scale emotional speech dataset. We freeze the emotion encoder during the training.

3.4. Emotion Synthesizer

We design the transformer-based emotion synthesizer to take input as emotionless talking faces corresponding to the contextual representation and emotional representations, simultaneously. For the query of the network, we use the neutral face of the current frame v_t^n and the emotional face of the previous frame v_t^e. This is because the current neutral face is used for the viseme mapping of the current contextual representation and the previous emotional face is used for expression consistency. The emotional representation $e_{1:kN}$, which is encoded through the emotion encoder, is used for the key and value of the network. Using emotion-guided attention, the network learns the mapping between emotional representations and facial expressions. At last, a single linear layer is added to decode the expressions for the emotional faces. In addition, since facial expressions for emotions can vary among individuals, a person-specific embedding is provided to the emotion synthesizer to represent the personalized emotional expression style. Person-specific embeddings p are defined as learnable vectors specific to each person in the training set. Thus, emotional talking faces v_t^e is estimated through the emotion synthesizer G_e as follows:

$$v_t^e = G_n(v_t^n \oplus v_{t-1}^e \oplus p, e_{1:kN}, B_A),$$

where v_t^e is the emotional face in t-th frame, p is the person-specific embedding.

3.5. Head Motion Synthesizer

Unlike emotional expression, head motion has a high diversity because head motion can be very different even from the same emotional speech. To ensure the diversity of the head motion, we employ the reparameterization trick. Through a single linear layer, the emotional representation is encoded into mean and variance. For each frame, diversified emotional representations are computed by multiplying Gaussian noise $N(0,1)$ by the encoded variance and adding the encoded mean. It is used as the key and value in the head motion synthesizer. For the query, we use the current emotional face v_t^e, the previous head rotation R_{t-1}, and the previous head translation T_{t-1}. The current emotional face is used to guide the correlation between expression and head motion, and the previous head motion is used for the head motion consistency. This allows the network to concurrently consider the speaker's emotional state and the continuity of head movement, thereby enabling the generation of natural head motions. For the emotional correlation, the emotional representation $e_{1:kN}$

is used for the key and value of the network. At the end of the transformer architecture, a single linear is added to decode the rotation matrix and translation vector of head motion. Thus, the head motion is estimated as follows:

$$[R_t, T_t] = G_h(v_t^E \oplus R_{t-1} \oplus T_{t-1}, e_{1:kN}, B_A),$$

where G_h is the head motion synthesizer, R_t and T_t are the rotation matrix and translation vector of the emotional head in t-th frame. Finally, the emotional talking face v_t with head motion in t-th frame is computed as follows:

$$v_t = v_t^e R_t + T_t$$

3.6. Network Training

3.6.1. Reconstruction Loss

The proposed method is composed of two stages to decouple context-driven and emotion-driven facial shapes during training. To accurately decouple these, we separately design the loss functions for the context-driven emotionless face and emotion-driven expressions. In addition, as the proposed method generates head motions, head motion loss is used in training. Thus, the reconstruction loss L_{rec} is defined as the summation of the losses for emotionless talking faces L_n, emotional talking faces L_e, head rotations L_r, and translations L_t, as follows:

$$\begin{aligned} L_n &= \|v_{1:N}^n - \bar{v}_{1:N}^n\|_2, \\ L_e &= \|v_{1:N}^e - \bar{v}_{1:N}^e\|_2, \\ L_r &= \|R_{1:N} - \bar{R}_{1:N}\|_2, \\ L_t &= \|T_{1:N} - \bar{T}_{1:N}\|_2, \\ L_{rec} &= L_n + L_e + \lambda_r L_r + \lambda_t L_t, \end{aligned}$$

where $v_{1:N}^n$, $vv_{1:N}^e$, $R_{1:N}$, and $T_{1:N}$ are the sequence set of the neutral faces, emotional faces, head rotation, and head translation, respectively, and $\bar{v}_{1:N}^n$, $\bar{v}_{1:N}^e$, $\bar{R}_{1:N}$, and $\bar{T}_{1:N}$ are the ground truth sequences. λ_r and λ_t are the balance factors between losses and we set $\lambda_r = \lambda_t = 0.01$ in our experiments. For this loss, it is required to obtain ground truth neutral talking faces synchronized with emotional talking faces. The dataset we have used contains instances of the same sentence spoken with and without emotion. However, since the presence or absence of emotion results in variations in the length of the speech, it cannot be used due to synchronization issues. To address this, based on the superior performance in emotionless talking face generation, we used the CodeTalker, which is the state-of-the-art emotionless talking 3D face generation method, to achieve neutral talking faces.

3.6.2. Contrastive Emotion-Face Loss

We present contrastive emotion-face loss to effectively learn the mapping between different domains: emotional representation and facial expression. We define contrastive emotion-face loss as the cross entropy between speech features, which includes both context and emotion, and the facial shape differences based on the presence or absence of emotion. Fig. 2 shows the details of the proposed contrastive emotion-face loss. For the speech features, we concatenate the contextual and emotional representations. This is because contextual and emotional lie in different latent spaces unlike the emotionless and emotional faces, which lie on the same vertex space. Then, speech features and facial shape differences are fed to the linear layer to transfer them into the embedding space for measuring similarity. The speech features $e_{1:N}^d$ is linearly interpolated with N samples matched with the frame rate of the face animation. The proposed contrastive emotion-face loss L_{con} is defined as follows:

$$\begin{aligned} L_{con} = &\sum_{i=1}^{N} CE\left(y(v_i^d), \frac{\exp(v_i^d \cdot e_i^d)}{\sum_{j=1}^{N} \exp(v_i^d \cdot e_j^d)}\right) \\ &+ \sum_{i=1}^{N} CE\left(y(e_i^d), \frac{\exp(v_i^d \cdot e_i^d)}{\sum_{j=1}^{N} \exp(v_i^d \cdot e_j^d)}\right) \end{aligned}$$

where $CE(\cdot,\cdot)$ is the cross-entropy function, y is the one-hot encoding matrix.

3.6.3. Head Motion Diversity Loss

To encourage the head motion synthesizer G_h to produce natural and diverse head movements, we explicitly regularize G_h by presenting the diversity loss. Following the diverse image generation task [8, 25], we define the diversity loss L_{div} as follows:

$$L_{div} = \frac{\left\|R_{z_1,1:N} - R_{z_2,1:N}\right\|_2 + \left\|T_{z_1,1:N} - T_{z_2,1:N}\right\|_2}{\|z_1 - z_2\|_2}$$

where the pair of $(R_{z_1,1:N}, R_{z_2,1:N})$ and $(T_{z_1,1:N}, T_{z_2,1:N})$ are generated rotation matrices and translation vectors by

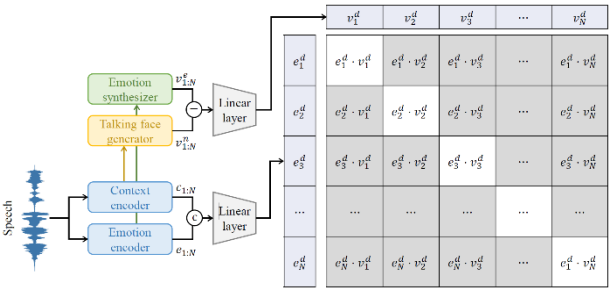

Fig. 2. Framework for calculating the proposed contrastive emotion-face loss. White boxes represent the correct pairs and gray boxes represent the incorrect pairs.

using two different emotional representations, which are diversified using Gaussian noise z_1, z_2 sampled differently. By maximizing the distance between two head motions with respect to the distance between latent noise z_1 and z_2, the proposed network focuses more on the variation of the head movement. In the absence of the motion diversity loss, mode collapse can readily occur within the head motion synthesizing network, leading the network to produce similar head motions for all speeches.

In summary, the overall loss for training the entire framework is defined as follows:

$$L = L_{rec} + \lambda_{con} L_{con} + \lambda_{div} L_{div},$$

where λ_{con} and λ_{div} are the balance factors between losses. We set the $\lambda_{con} = 1$ and $\lambda_{div} = 0.01$ in all our implementations.

3.7. Personalized Emotional Tuning

Fig. 2 (right) shows the framework of the proposed personalized emotional tuning. Personalized emotional tuning is to leverage a small dataset from a target individual, not previously used during training, to generate emotional expressions of unseen subject. Thus, from a small set of unseen speech $\hat{s}_{1:M}$ and emotional faces $\hat{v}_{1:M}$, the network is optimized to learn the subtle nuance of facial expressions corresponds to the target subject's emotional state. Here, as the proposed network is trained to produce personalized emotional expressions using person-specific embeddings, we optimize the person-specific embedding to represent the personalized emotional expression from the unseen small dataset rather than optimize the entire network. Thus, the objective for the personalized emotional tuning is defined as follows:

$$p^* = \underset{p}{\operatorname{argmin}} L_e(v^e_{1:N}, \hat{v}^e_{1:N}, p),$$

where p is the person-specific embedding and $v^e_{1:N}$ is the estimated emotional faces from the emotion synthesizer. In the network training, the person-specific embedding p is encoded from the one-hot identity label. In contrast, in the personalized emotional tuning, we directly optimize the person-specific embedding p instead of encoding from the one-hot identity label.

However, optimizing the network strictly on a limited dataset of target subjects can not only introduce artifacts but also make it easy to overfit to that particular dataset [2]. To address this, we present the latent constraint tuning scheme. We observed that the person-specific embedding is well-positioned within the latent space because it is trained using a relatively large dataset. Fig. 3 visualizes the results of

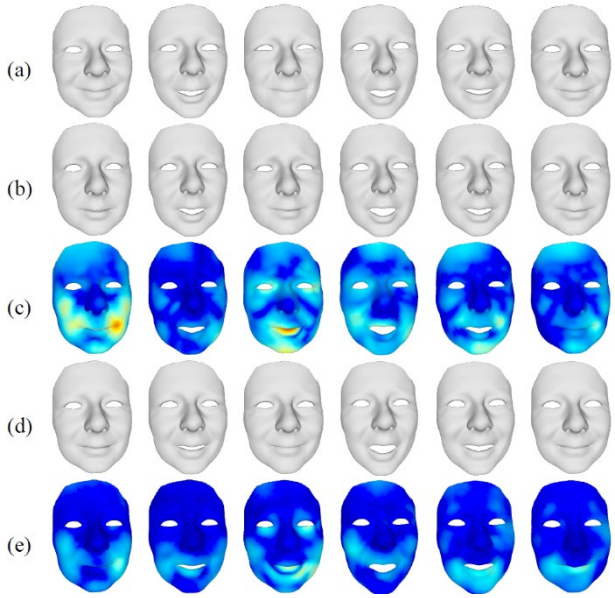

Fig. 3. Comparison of talking faces according to the utilization of person-specific embeddings during the training. (a) is the ground truth faces, (b) is the network output trained without person-specific embeddings, and (d) is the network output trained with person-specific embeddings. (c) and (e) are the vertex displacements of (b) and (d) with regard to the (a), respectively.

talking faces, comparing cases with and without the utilization of person-specific embeddings. The result shows that using person-specific embedding assists the network in training more accurately on an individual's emotional expressions. Thus, it confirms that the person-specific embedding is well-positioned within the personalized emotional latent space.

Based on this observation, we find that the well-trained person-specific embeddings can be regarded as the pivots that represent the personalized emotional expression. Thus, using the pre-trained person-specific embeddings p as pivots and fixed generators, we estimate the person-specific embedding p^* of the target subject. Instead of directly optimizing the person-specific embedding p^* of the target subject, we optimize a weight factor α that represents how far the target subject's embedding deviates from the pre-trained embeddings p The objective L_{tun} for personalized emotional tuning is newly defined as follows:

$$L_{sum}(\alpha) = \left(\sum_i \alpha_i - 1\right)^2,$$

$$L_{tun} = L_e\left(v^e_{1:N}, \hat{v}^e_{1:N}, \sum_i \alpha_i p_i\right) + L_{sum}(\alpha),$$

$$\alpha^* = \underset{\alpha}{\operatorname{argmin}} L_{tun},$$

where L_{sum} is the penalty term to restrict the sum of α_i to 1. Thus, the person-specific embedding p^* of an unseen

subject is computed as follows:

$$p^* = \sum_i \alpha_i^* p_i.$$

This ensures that during the optimization process, the convexity of the embedding is preserved, thereby minimizing artifacts and yielding more natural outcomes.

IV. EXPERIMENTAL RESULTS

4.1. Experimental Details

4.1.1. Dataset

We utilize the BIWI dataset [7] for network training and evaluation because the BIWI is the only publicly available dataset that includes emotional speech and corresponding facial scans. VOCASET [4] and Multiface [23] datasets do not contain the emotional information. BIWI dataset contains synchronized audio and facial scan pairs of 14 subjects where each subject uttered 40 phonetically balanced English sentences twice, with neutral and emotional expressions. Thus, totally synchronized 1,120 speeches and 3D facial scans are included in the BIWI dataset. The 3D facial scans are captured at 25 fps with 23,370 vertices. In our experiments, we split the BIWI dataset into a training set of 384 sentences spoken by six subjects, a validation set of 48 sentences spoken by six subjects, and a test set of 56 sentences. The test set is split into four subsets: BIWI-SN, BIWI-SE, BIWI-UN, and BIWI-UE. BIWI-SN and BIWI-SE set each contain 24 neutral and emotional sentences spoken by six seen subjects, who are the same subjects in the training set. BIWI-UN and BIWI-UE set each contain 64 neutral and emotional sentences spoken by eight unseen subjects, who are not included in the training set. In each BIWI-UN and BIWI-UE set, 6 emotional speeches are used for the personalized emotional tuning and 32 neutral and 26 emotional speeches are used for the evaluation.

4.1.2. Implementation Details

We use the sampling rate for the speech f_s as 49Hz and for the facial animation f_v as 25 fps. In all transformer architecture, we used $N_d = 2$ blocks with 4 attention heads, 128 hidden nodes, and layer normalization. We experimentally observed that the increasing number of attention blocks and heads does not improve performance. We used an absolute positional encoding. For the training, it is required that both neutral and emotional talking faces. In the BIWI dataset, faces of neutral speech and emotional speech are included, but they are not aligned. To make the alignment, we employ CodeTalker to obtain the ground truth of neutral talking faces from emotional. Since this method neglects the emotional condition in talking face generation, neutral talking faces correspond to emotional speech can be obtained. We used the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The learning rate is initialized as 10^{-4} and decayed after 100 epochs with a factor of 0.1. The training is converged after 100 epochs using a GPU of NVIDIA 3080 (10 GB).

4.2. Qualitative and Quantitative Evaluation

For the comparison, we use two neutral talking face generation methods, FaceFormer [8] and CodeTalker [10], and two emotional talking face generation methods, FaceXHuBERT [11] and EmoTalk [12]. Fig. 4 visualizes the qualitative evaluation on BIWI-SE and BIWI-UE datasets. We also visualize the ground truth data for an accurate comparison. The results from CodeTalker display significant discrepancies on the truth surface, as they generate talking faces by neglecting emotional expressions even in emotional speeches. In contrast, FaceXHuBERT and EmoTalk demonstrate the capability to generate emotional talking faces, as evident by the smiling lips in Fig. 4(a) and opening mouth largely in Fig. 4(b). However, all these methods do not generate head movement, which is an important factor in increasing emotional delivery. The proposed method generates both emotional expression and head movement from the speech. For the accurate comparison in expressing emotions, we also visualize our results without head motions. It is shown that the proposed method generates emotional expression more accurately than the comparison methods, especially in smiling lips and largely opening mouth.

For the quantitative evaluation, we summarize the L_2 vertex errors in lip and entire facial regions in Table 1. Vertex error provides a precise, reproducible, and quantitative measure of how closely the generated 3D mesh aligns with the ground truth. In this study, we do not employ perceptual metrics such as mean opinion scores (MOS), as these evaluations can vary significantly based on viewers' cultural backgrounds or individual interpretations of emotions. Such subjectivity poses a challenge for maintaining a consistent and objective quality evaluation; therefore, we prioritize quantitative geometric metrics to ensure reproducible results.

The result shows that the emotional talking face generation methods [8,10] including ours show lower vertex error than the neutral talking face generation methods [11-12]. In addition, the result shows that the proposed method outperforms in both lip and facial vertex error than the comparison methods. Especially in the talking face generation of unseen subjects, it is demonstrated that the proposed method with personalized

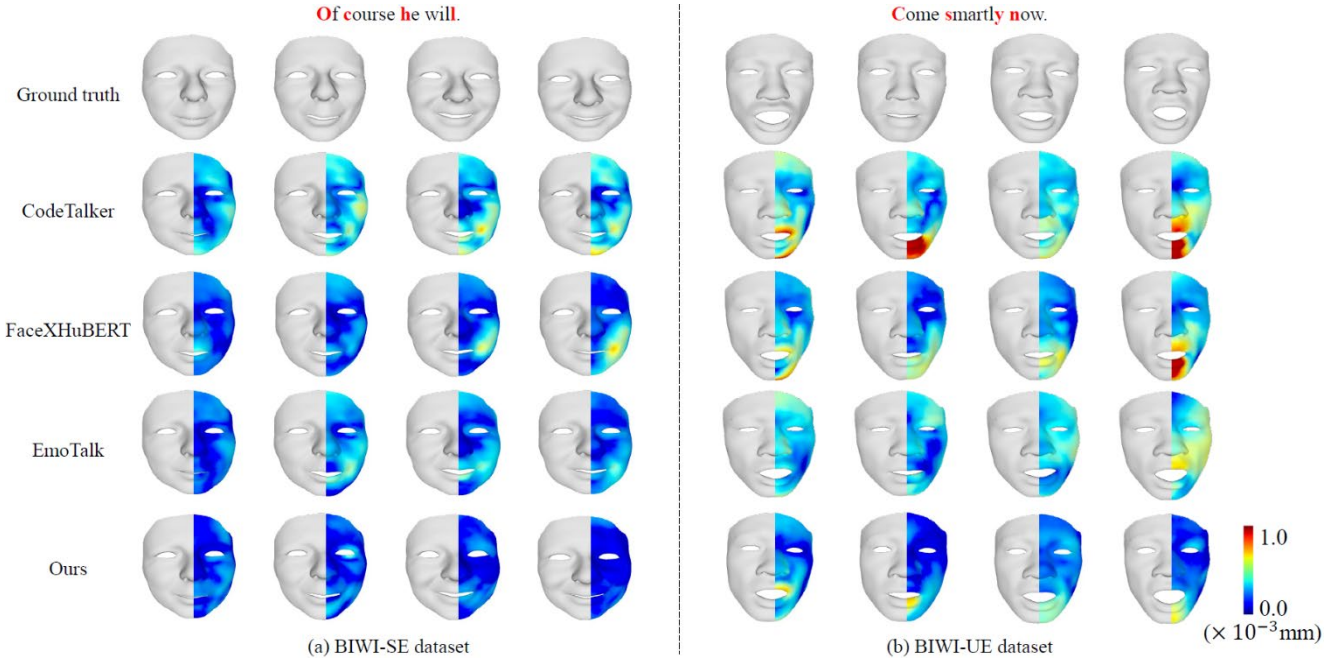

Fig. 4. Results of 3D talking face generation of ours and comparison methods on BIWI dataset: (a) BIWI-SE and (b) BIWI-UE dataset.

Table 1. Comparison of L_2 vertex error (lip/face) on neutral and emotional speech on BIWI dataset (10^{-4} mm).

Method\ Dataset	BIWI-SN	BIWI_SE	BIWI-UN	BIWI-UE
FaceFormer	5.608 / 4.354	7.926 / 5.514	11.350 / 6.136	12.745 / 7.756
CodeTalker	**4.806 / 3.336**	6.025 / 4.324	10.096 / 5.785	10.706 / 6.263
FaceXHuBERT	5.553 / 3.936	5.307 / 3.932	10.963 / 6.310	10.201 / 6.106
EmoTalk	5.045 / 3.694	5.328 / 3.647	9.786 / 6.068	9.987 / 6.320
Ours (w/o tuning)	4.828 / 3.355	**4.926 / 3.514**	9.745 / 5.756	9.760 / 5.954
Ours (w tuning)	4.828 / 3.335	**4.926 / 3.514**	**5.298 / 3.866**	**5.416 / 4.151**

emotional tuning significantly outperforms the other comparison methods. Note that the speech data used in the personalized emotional tuning is not used for the evaluation. Thus, it is confirmed that the proposed method not only accurately generates emotional talking faces from speech but also accurately produces personalized emotional talking faces for subjects who are not used in the training through personalized emotional tuning. Fig. 5 visualizes the generated talking heads from neutral and emotional speeches. The left side represents the generated results from neutral speech and the right side represents the generated results from emotional speech in the BIWI dataset. The 14 categories of emotion labels are measured in the BIWI dataset. The speech used in this experiment has the highest for sadness among multiple emotional categories. In the neutral speech, the result shows that the network outputs from the neutral face generator and emotion synthesizer have similar results. In contrast, in emotional speech, the network outputs from the neutral face generator and emotion synthesizer are significantly different. The results from the emotion synthesizer have more emotional expression than that of the neutral face generator. In addition, the proposed network generates more large head movements from the emotional speech compared to the neutral speech, as shown in Fig. 5(d). This result represents the overlap result of the neutral face generator and head motion synthesizer.

4.3. Evaluation on Personalized Emotional Tuning

By manipulating the person-specific embeddings, the proposed model can synthesize new talking styles. We select two person-specific embeddings p_i and p_j and interpolate new person-specific embedding as $p_{new} = \omega p_i + (1-\omega)p_j$ using linear interpolation. For the newly synthesized talking style, we plot the lip distance between upper and lower lips across frames in Fig. 6. The result shows that the synthesized styles have a smooth translation of lip distance between the two reference styles. It confirms that the person-specific embedding effectively represents the talking styles and is also practical for generating talking faces of an unseen subject. The proposed method can generate accurate new talking styles using personalized emotional tuning. To evaluate this, we visualize the generated talking faces of unseen subjects according to the

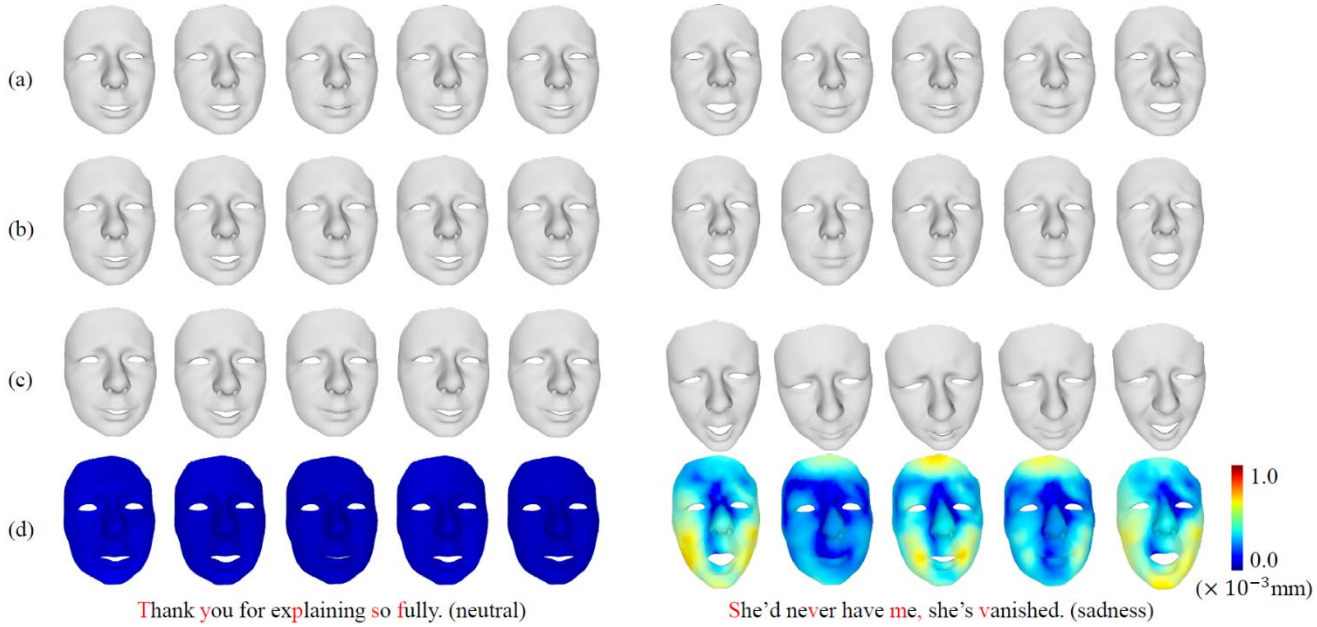

Fig. 5. Visualization results of talking heads from neutral (left) and emotional (right) speeches. (a) is the output computed from the talking face generator, (b) is the output computed from the emotion synthesizer, (c) is the final output with head motion, and (d) is the difference map between (a) and (b).

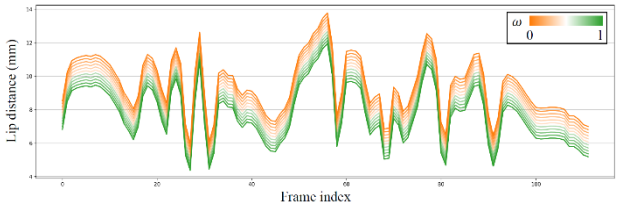

Fig. 6. Lip distance between the upper and lower lips conditioned on different weight ω in linear interpolations between two person-specific embeddings.

personalized emotional tuning in Fig. 7. Here, Fig. 7(a) is the ground truth faces, Fig. 7(b) is the network output without personalized emotional tuning, and Fig. 7(d) is the network output after the personalized emotional tuning. Note that the speech input for talking face generation is selected and is not used in the personalized emotional tuning for a fair comparison. In addition, we intentionally remove the head motions for an accurate qualitative comparison. This result shows that the generated talking faces of the unseen subject do not accurately represent the emotional expression. In contrast, after the personalized emotional tuning, it is demonstrated that the network output has less vertex error with the ground truth faces compared to that without personalized emotional tuning. It shows that the proposed personalized emotional tuning can be effective in generating talking faces of unseen subjects.

Fig. 8 shows the lip distance according to the usage of personalized emotional tuning. Note that the visualized sample in this evaluation is one that was not used in either the network training or the personalized emotional tuning.

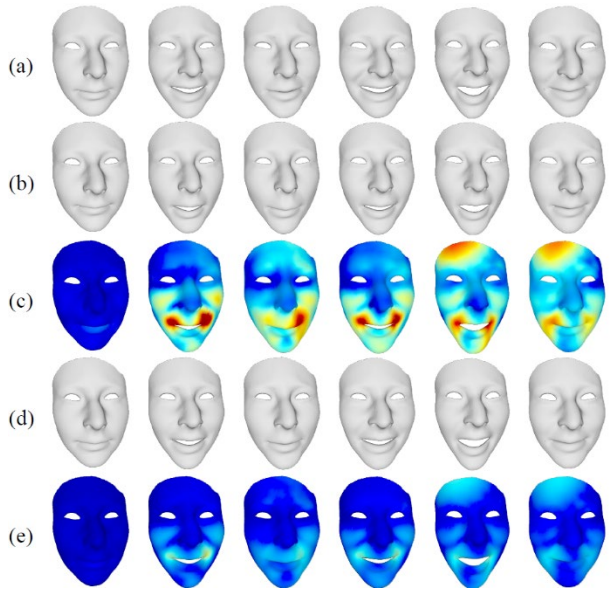

Fig. 7. Comparison of talking faces according to the personalized emotional tuning. (a) is the ground truth faces, (b) and (d) are the network output without and with the personalized emotional tunning, respectively. (c) and (e) represent the vertex displacements of (b) and (d) with regard to the (a).

This result shows that the synthesized talking faces without personalized emotional tuning exhibit a large discrepancy from the ground truth lip distance. In contrast, the synthesized talking faces with personalized emotional tuning show similar lip distance to the ground truth. It confirms that the proposed personalized emotional tuning is effective in representing personalized talking styles. This demonstrates that the proposed method can be beneficial in generating personalized talking content.

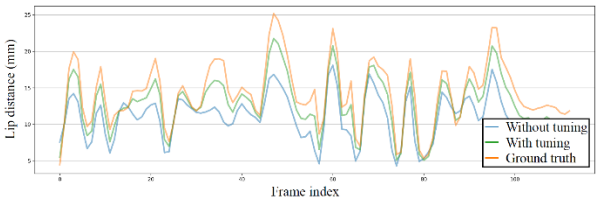

Fig. 8. Lip distance between the upper and lower lips according to the personalized emotional tuning.

4.4. Evaluation on Head Motion

Humans move their heads differently when even speaking the same sentence. Thus, the diversity of the head motions is important to increase the naturalness of the 3D facial animation. To evaluate the diversity, we visualize the head movements across the time in Fig. 9. It visualizes the generated head movements from neutral and emotional speech inputs, separately. Generated head movements from the neutral speech are demonstrated in Fig. 9(a) and head movements from the emotional speech are demonstrated in Fig. 9(b). The head movements are visualized along the x, y, and z directions, separately. For a fair comparison, we use the same sentence for this evaluation. We repeat the head motion generation 20 times to accurately evaluate the head motion diversity. This result shows that the proposed method can generate diverse head motions from the speech. In particular, it is shown that the proposed method generates more dynamic and diverse head motions from the emotional speech than the neutral speech. This demonstrates that the proposed network is able to more effectively attempt emotion transmission by generating dynamic head movements in emotional speech. Furthermore, it is shown that the proposed method emphasizes generating head movements in the y direction than the x and z directional movements. This is a semantically natural result because the y direction represents the primary vertical nodding motion of the head.

4.5. Ablation Study

Based on the fact that context-driven lip motion and emotion-driven expressions are coupled in the emotional talking face, we proposed a two-stage training approach that generates context-driven lip motions first and then adds emotion-driven expressions. To verify the proposed approach, we conduct the ablation study according to the generation strategies. Table 2 summarizes the L_2 vertex errors of lip and entire facial regions according to generation strategies: single-stage, two-stage, two-stage with contrastive emotion-face loss. The single-stage method is designed by excluding the neutral talking face loss L_n. The two-stage with contrastive emotion-face loss L_{con} is the proposed method. The result shows that employing a two-stage approach has performance improvement than results of using single-stage generation. In addition, the results show that using the proposed contrastive emotion-face loss in a two-stage approach achieves the lowest vertex error. It confirms that the proposed two-stage method effectively decouples the context-driven lip motions and emotion-driven expressions than the single-stage method.

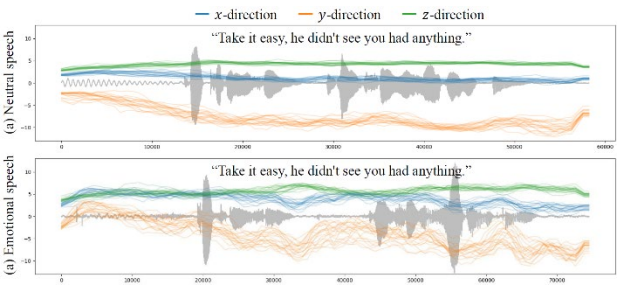

Fig. 9. Head motion visualization results. (a) represents the head motion in neutral speech and (b) represents the head motion in emotional speech. (a) and (b) is spoken with the same sentence.

4.6. Computational Cost

The experimental results show that the proposed method outperforms state-of-the-art speech-driven 3D face animation methods. However, the main bottleneck in our method is the computational cost, comprising 199.95M parameters, generates all the facial vertices instead of blendshape parameters. Despite this complexity, the model demonstrates practical efficiency in our experimental setup using an NVIDIA RTX 3080. Given that an 11-second audio sequence at our target frame rate of 25 fps consists of 275 frames, the total inference time is 6.048 seconds, which corresponds to approximately 21.99 ms per frame. This latency is well within the 40 ms threshold required for real-time playback, demonstrating the practical efficiency of our vertex-based generation approach.

V. CONCLUSION

Unlike existing speech-driven methods, which generate emotionless talking faces or pre-defined limited emotional styles, our method animates emotional talking faces expressed with personalized emotional styles. By proposing the personalized emotional tuning scheme, the proposed method can accurately generate a new emotional talking

Table 2. Ablation study on blation study on L_2 vertex error (lip/face) on BIWI dataset (10^{-4} mm).

Dataset	1-stage	2-stage	2-stage+$\boldsymbol{L_{con}}$
BIWI-SE	5.345 / 3.970	5.143 / 3.626	4.926 / 3.514
BIWI-UE	6.057 / 4.836	5.678 / 4.399	5.416 / 4.151

style, which is unseen in the training dataset. By doing this, the proposed method effectively increases the visual quality and further increases the engagement of users in virtual communication. While the current implementation of the proposed method generates the 3D talking face after receiving the entire audio sequence, it is also capable of delayed real-time synthesis by processing the audio in fixed intervals (e.g., 1-second segments). By adopting this window-based inference strategy, our framework can achieve seamless emotional animation with minimal latency, facilitating its use in interactive applications.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25432454).

REFERENCES

[1] C. Park, J. Cho, J. Kim, S. Lee, J. Kim, and S. Lee, "AVIN-chat: An audio-visual interactive chatbot system with emotional state tuning," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 8763-8766, 2024.

[2] K. Lee, J. Lee, H. Lee, M. Jang, S. Lee, and S. Lee, "Faceclone: Interactive facial shape and motion cloning system using multi-view images," in *Proceedings of the 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, IEEE, 2023, pp. 512-513.

[3] M. Jang, K. Lee, S. Lee, H. Tong, J. Chung, and Y. Ro, et al., "InViTe: Individual virtual transfer for personalized 3D face generation system," in *Proceedings of the Thirty-Third International Joint Conf-erence on Artificial Intelligence*, 2024, pp. 8683-8686.

[4] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10101-10111.

[5] Y. Nitzan, K. Aberman, Q. He, O. Liba, M. Yarom and Y. Gandelsman, et al., "Mystyle: A personalized generative prior," *ACM Transactions on Graphics*, vol. 41, no. 6, pp. 1-10, 2022.

[6] S. R. Livingstone and C. Palmer, "Head movements encode emotions during speech and song," *Emotion*, vol. 16, no. 3, p. 365, 2016.

[7] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool, "A 3-d audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591-598, 2010.

[8] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "FaceFormer: Speech-driven 3D facial animation with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18770-18780.

[9] A. Richard, M. Zollhöfer, Y. Wen, F. De la Torre, and Y. Sheikh, "MeshTalk: 3D face animation from speech using cross-modality disentanglement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp.1173-1182.

[10] J. Xing, M. Xia, Y. Zhang, X. Cun, J. Wang, and T. T. Wong, "CodeTalker: Speech-driven 3D facial animation with discrete motion prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 12780-12790.

[11] K. I. Haque and Z. Yumak, "FaceXHuBERT: Text-less speech-driven expressive 3D facial animation synthesis using self-supervised speech representation learning," in *Proceedings of the 25th International Conference on Multimodal Interaction*, 2023, pp. 282-291.

[12] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, and J. He, et al., "EmoTalk: Speech-driven emotional disentanglement for 3D face animation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 20687-20697.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, and S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 8748-8763.

[14] S. Lee, J. Lee, H. Song, and S. Lee, "Speech-driven emotional 3D talking face animation using emotional embeddings," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 7840-7844.

[15] Y. Chen, J. Zhao, and W. Q. Zhang, "Expressive speech-driven facial animation with controllable emotions," in *Proceedings of the 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, IEEE, 2023, pp. 387-392.

[16] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, and X. Cao, et al., "Audio-driven emotional video portraits," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 14080-14089.

[17] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, and F. Xu, et al., "EAMM: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1-10.

[18] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli,

"wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449-12460, 2020.

[19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, and S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, PMLR, 2021, pp. 8748-8763.

[20] Q. Wang, Z. Fan, and S. Xia, "3D-TalkEmo: Learning to synthesize 3D emotional talking head," *arXiv Prep. arXiv2104.12051*, 2021.

[21] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, and E. Elsen, et al., "Deep speech: Scaling up end-to-end speech recognition," *arXiv Prep. arXiv 1412.5567*, 2014.

[22] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv Prep. arXiv2104.03502*, 2021.

[23] C. H. Wuu, N. Zheng, S. Ardisson, R. Bali, D. Belko, and E. Brockmeyer, et al., "Multiface: A dataset for neural face rendering," *arXiv Prep. arXiv2207.11243*, 2022.

AUTHOR

Seongmin Lee received the BS degree in electronic and electrical engineering from Hongik University, Seoul, Korea, in 2018, and the M.S. and Ph.D. degrees in electrical and electronic engineering with the Multidimensional Insight Laboratory, Yonsei University, Seoul, Korea, in 2024. From 2024 to 2025, he worked as a postdoctoral researcher with the University Industry Foundation of Yonsei University, Seoul, Korea. He is currently an assistant professor with the Department of Artificial Intelligence Software, Hanbat National University, Sejong, Korea. His research interests include computer vision, computer graphics, and deep learning.