Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Lightweight Speech-Conditioned Upper-Face Animation for Virtual Agents via Emotion–Liveness Composition

HWANG YOUN KIM[1], (Member, IEEE), GHAZANFAR ALI[2], JEONGHA LEE[3], (Member, IEEE), AND JAE-IN HWANG[4], (Member, IEEE)

[1]AI-Robotics, University of Science and Technology, Daejeon 34113, South Korea
[2]Department of AI, Gachon University, Gyeonggi-do 13120, South Korea
[3]AI-Robotics, University of Science and Technology, Daejeon 34113, South Korea
[4]Intelligence and Interaction Research Center, Korea Institute of Science and Technology, Seoul 02792, South Korea

Corresponding author: Jae-In Hwang (e-mail: hji@kist.re.kr).

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (RS-2026-25516375, Development of Core Technologies for Structuring and Synchronization of Unstructured Multimodal Sensory Data) and the Korea Institute of Science and Technology (KIST) Institutional Program (No. 26E0212, Development of Hyper-Space Technology for Overcoming the Limits of Time and Space).

ABSTRACT Generating conversational facial expressions in real time still faces significant limitations, and most studies focus on lip-sync animation with limited attention to upper-face movements such as eyebrow and eye-region motion. Nevertheless, generating detailed facial movements using deep learning models requires considerable computational time and cost. In this paper, we propose a hybrid upper-face animation framework that can be applied to various virtual agents with near-real-time computational performance. Unlike full-face facial-animation models, the proposed model directly generates only nine upper-face blendshape channels corresponding to eyebrow and eye-region movements, while lip motion is generated by an external real-time lip-sync module. The proposed framework consists of an emotion-conditioned upper-face generator and a liveness generator. Emotional upper-face animation is generated by a lightweight Transformer-based regressor conditioned on acoustic features, emotion category, emotion intensity, and speaker style, while a separately trained generative motion prior uses speech features as weak temporal and prosodic cues to produce natural liveness variations.

The proposed generators contain 4.18M parameters in total, approximately $1/30$ the number of parameters of EmoFace and $1/150$ that of EmoTalk, and achieve a CPU inference time of 13.16 ms for approximately 3-s audio clips. Liveness-motion fusion increased the motion amplitude ratio from 0.724 to 0.997. This indicates that the proposed method can generate upper-face motion with an amplitude close to that of real facial motion. Furthermore, in a user study with 36 participants, the proposed method achieved the highest mean Animacy rating and showed no significant post-hoc differences from EmoFace.

INDEX TERMS Blendshape animation, conversational facial animation, virtual agent facial animation, HCI, deep learning, augmented/virtual/mixed realities.

I. INTRODUCTION

Conversational facial animation is an essential component of virtual agents and digital humans. It helps establish rapport with users and enhances social presence. Previous studies have demonstrated the importance of facial expressions as nonverbal cues in interactive and conversational systems [1], [2]. However, generating realistic talking-face animation with low computational cost remains challenging for interactive virtual agent applications. Furthermore, some deep learning methods generate 3D mesh vertices or high-dimensional facial geometry, which can lead to high computational cost and long processing times [3], [4]. These limitations raise the question of how speech-driven facial animation can be effectively designed for conversational agents. Recent studies have emphasized accurate lip synchronization, but lip and mouth movements are not sufficient to achieve natural conversational facial behavior. In real conversations, upper-face expressions, such as eyebrow movements and subtle

changes around the eyes, contribute to conveying emotional nuance and speech intent [5], [6]. To address these issues, we propose a lightweight hybrid framework for speech-driven upper-face animation. The proposed framework decomposes facial animation into lip synchronization and upper-face animation. The generated upper-face movements are combined with audio-driven lip motions produced by an existing real-time lip-synchronization API. The proposed upper-face animation module consists of two main components: an emotion-conditioned upper-face generator and a CVAE-GAN-based liveness generator. The main contributions of this study are as follows. First, we propose a facial-animation framework that combines an existing real-time lip-synchronization system with an upper-face motion generator. The proposed method predicts only nine upper-face motion channels, whereas EmoTalk [7] predicts 52 blendshape coefficients and EmoFace [8] predicts 174 MetaHuman controller values. The two proposed generators contain 4.18M parameters in total and achieve a CPU inference time of 13.16 ms for approximately 3-s audio clips. Second, we introduce an emotion–liveness composition method that separately models emotion-related motion and natural liveness variation before combining them. The emotion-conditioned generator produces controllable upper-face expressions, while the liveness generator produces natural upper-face movements that cannot be fully determined from audio features alone. With liveness-motion fusion, the motion amplitude ratio increased from 0.724 to 0.997, close to the reference value of 1.0. Third, a user study with 36 participants showed that the proposed method achieved the highest mean Animacy rating and did not significantly differ from EmoFace.

II. RELATED WORK

A. PARAMETRIC, BLENDSHAPE, AND ACTION-UNIT REPRESENTATIONS

Early facial-animation systems represented motion through manually specified geometric parameters, phoneme-to-viseme rules, and reusable animation units [9], [10]. These approaches remain attractive in real-time applications because they expose compact and predictable controls. JALI, for example, provides animator-oriented viseme and expressive controls for lip synchronization [11]. However, the range and temporal variation of such systems are bounded by the rules and motion units designed in advance.

Modern systems commonly represent facial motion as dense mesh vertices, blendshape coefficients, or Facial Action Coding System (FACS) Action Units (AUs). Dense vertices preserve geometric detail but are tied to a specific topology and introduce a high-dimensional output space. Blendshapes provide a compact rig-specific parameterization, whereas FACS describes visible facial movements using semantically interpretable AUs [5]. OpenFace makes AU occurrence and intensity estimates available from ordinary video, allowing AU trajectories to be used for data construction and evaluation [12]. Chen et al. further showed that speech-related AUs can be predicted from audio and used as localized control signals in talking-head generation [13]. Their work concentrates primarily on speech-related mouth motion; in contrast, we generate upper-face AU trajectories and combine them with an independent lip-animation module. This representation isolates eyebrow and eyelid behavior from character-specific mesh topology, although a fixed AU-to-rig mapping is still required for rendering.

B. SPEECH-DRIVEN FACIAL ANIMATION AND UPPER-FACE AMBIGUITY

Deep speech-driven animation replaces handcrafted mappings with learned temporal models. Karras et al. jointly learned facial pose and emotion from audio [3], while Pham et al. mapped acoustic features to head rotation and facial AU or blendshape activations using a recurrent architecture [14]. VOCA learned a subject-conditioned mapping from speech to 3D facial deformation and generalized across speakers and languages [15]. FaceFormer later used an autoregressive Transformer and self-supervised speech features to model longer temporal context [4].

The upper face presents a different problem from phonetic articulation. Lip and jaw motion are strongly constrained by speech content, whereas blinks and many brow movements are only weakly or inconsistently determined by the acoustic signal. They may nevertheless correlate with prosody and emphasis, so treating all upper-face behavior as completely audio-independent is also inaccurate. MeshTalk explicitly separated audio-correlated and audio-uncorrelated facial components and showed that such disentanglement helps avoid static upper-face animation while preserving speech-related motion [6].

A further limitation of deterministic pointwise regression is that several facial trajectories may be plausible for the same utterance. Minimizing a L_1 or L_2 loss against one reference can therefore favor an average trajectory with reduced motion amplitude. Recent systems address this one-to-many mapping through learned motion priors. CodeTalker predicts discrete codes from a VQ-VAE motion codebook [16]; FaceDiffuser uses a conditional diffusion process and supports both vertex and blendshape representations [17]; and Yang et al. formulate speech-driven facial motion as a probabilistic problem and evaluate both speech fidelity and output diversity [18]. These works motivate probabilistic modeling for upper-face motion, where natural variation should not be collapsed into a single deterministic sequence.

C. EMOTION- AND INTENSITY-CONTROLLABLE FACIAL ANIMATION

Emotional animation requires the model to preserve speech-related motion while controlling affective expression. EmoTalk introduces an emotion-disentangling encoder and cross-reconstruction strategy to separate emotional information from speech content [7]. EMOTE uses a temporal variational motion prior together with decoupled speech and emotion losses, enabling explicit emotional control while maintaining speech-related articulation [19]. Chen et al.

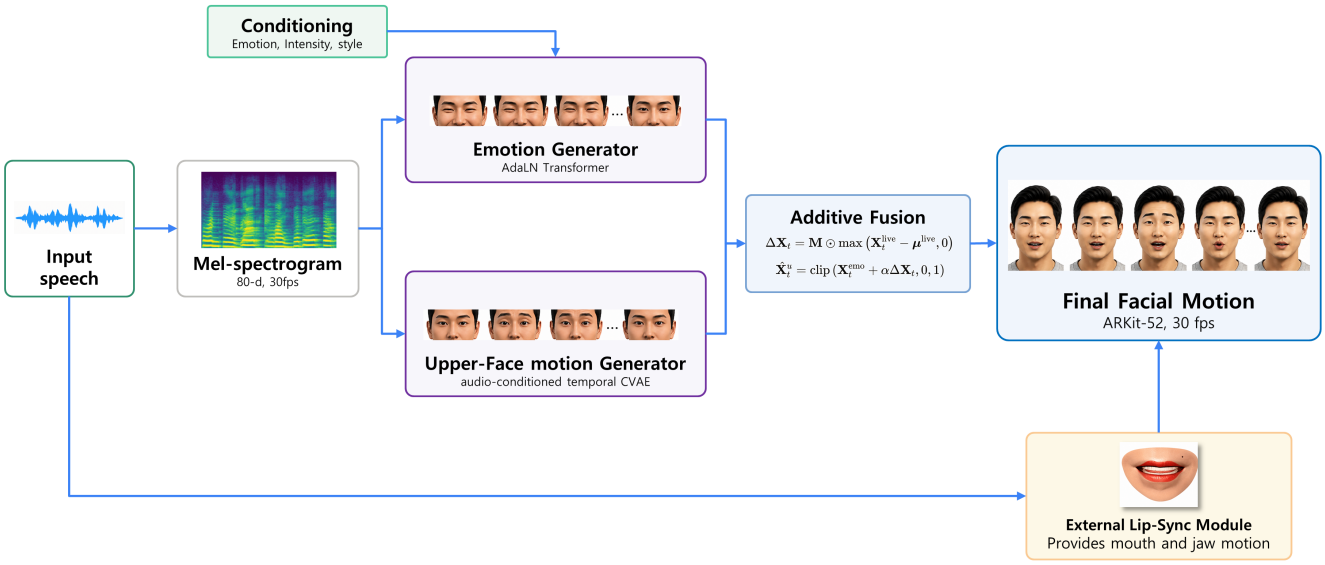

FIGURE 1. Overview of the proposed speech-driven facial animation framework. Emotional upper-face motion and natural motion variation are fused and combined with external lip synchronization.

directly condition facial animation on emotion type and a controllable intensity value [20]. EmoFace employs separate speech and emotion encoders to predict facial controller trajectories and additionally models blinks and eye movements for a MetaHuman-oriented rig [8].

More recent work has increased the granularity or stochasticity of emotional control. DEITalk models time-varying emotional intensity [21]. ProbTalk3D uses a VQ-VAE-based probabilistic formulation conditioned on audio, identity, emotion class, and emotion intensity [22]. EmoVOCA trains emotional 3D talking-head generators with explicit emotion and intensity inputs [23]. DEEPTalk learns a probabilistic cross-modal emotion embedding and a hierarchical VQ-VAE motion prior for diverse emotional facial motion [24]. NVIDIA Audio2Face-3D provides a real-time, deployment-oriented framework with retargeting support and is useful as a practical system comparison [25].

D. VARIATIONAL AND ADVERSARIAL SEQUENCE GENERATION

Variational autoencoders learn a regularized latent distribution from which multiple outputs can be sampled [26], while generative adversarial networks train a generator against a discriminator that distinguishes generated samples from observations [27]. VAE–GAN formulations combine an explicit latent encoder with an adversarially learned realism criterion [28]; Wasserstein objectives with gradient penalties are frequently used when additional adversarial-training stability is needed [29]. In speech-driven talking-face generation, Vougioukas et al. used multiple discriminators to encourage frame quality, audio-visual synchronization, and natural facial behaviors including blinks and eyebrow movements [30].

III. METHOD

A. OVERALL FRAMEWORK

Our framework synthesizes facial motion from an audio feature sequence $\mathbf{A}_{1:T}$ through two components: an emotion-conditioned upper-face generator and a liveness generator, as illustrated in Fig. 1. First, the emotion-conditioned generator G_e produces emotion-specific upper-face motion from the speech signal, emotion label, emotion intensity, and speaker identity. The liveness generator G_l generates audio-conditioned motion variations that improve the naturalness and dynamics of the upper facial animation sequence. Previous talking-face studies have shown that upper-face movement is weakly correlated with the speech signal. Thus, instead of predicting a single animation sequence through deterministic regression, we train a CVAE-GAN to learn a plausible upper-face motion prior from weak audio cues and a stochastic latent variable. The overall process is formulated as follows:

$$\hat{\mathbf{X}}_{1:T}^{u,\text{emo}} = G_e(\mathbf{A}_{1:T}, \mathbf{e}, \mathbf{i}, \mathbf{s}_{\text{id}}), \tag{1}$$

$$\begin{aligned}\mathbf{z}_{1:K} &\sim p_\theta(\mathbf{z}_{1:K} \mid \mathbf{A}_{1:T}) \\ &= \mathcal{N}\big(\boldsymbol{\mu}_p(\mathbf{A}_{1:T}), \text{diag}\big(\boldsymbol{\sigma}_p^2(\mathbf{A}_{1:T})\big)\big),\end{aligned} \tag{2}$$

$$\hat{\mathbf{X}}_{1:T}^{u,\text{live}} = G_l(\mathbf{A}_{1:T}, \mathbf{z}_{1:K}), \tag{3}$$

$$\boldsymbol{\mu}^{\text{live}} = \frac{1}{T}\sum_{\tau=1}^{T} \hat{\mathbf{X}}_\tau^{u,\text{live}}, \tag{4}$$

$$\Delta \hat{\mathbf{X}}_t^{u,\text{live}} = \max\Big(\hat{\mathbf{X}}_t^{u,\text{live}} - \boldsymbol{\mu}^{\text{live}}, \mathbf{0}\Big). \tag{5}$$

$$\hat{\mathbf{X}}_t^u = \mathrm{clip}\left(\hat{\mathbf{X}}_t^{u,\mathrm{emo}} + \alpha\,\Delta\hat{\mathbf{X}}_t^{u,\mathrm{live}}, \mathbf{0}, \mathbf{1}\right), \tag{6}$$

$$K = \left\lceil \frac{T}{10} \right\rceil, \qquad t = 1, \ldots, T. \tag{7}$$

Here, G_e denotes the emotion-conditioned facial-motion generator, whereas G_l denotes the CVAE-GAN-based upper-face liveness generator. $\hat{\mathbf{X}}_{1:T}^{u,\mathrm{emo}}$ denotes the emotion-conditioned base motion, and $\hat{\mathbf{X}}_{1:T}^{u,\mathrm{live}}$ denotes the upper-face liveness motion generated from the upsampled latent sequence and frame-level acoustic features. $\hat{\mathbf{X}}_t^u$ is the final upper-face motion. $\mathbf{e}$ and $\mathbf{i}$ denote the emotion label and emotion intensity; $\mathbf{s}_{\mathrm{id}}$ is an actor-level style embedding used to condition the emotion-conditioned generator; and α is a scaling factor applied to the positive mean-centered liveness residual $\Delta\hat{\mathbf{X}}_t^{u,\mathrm{live}}$.

Following MeshTalk's temporal convolutional audio encoder [6], both generators take 80-dimensional log-Mel spectrograms extracted from 16-kHz speech as input. The audio encoders consist of four temporal convolutional layers with a kernel size of 5 and GELU activations. The emotion generator produces a 64-dimensional acoustic representation, which is projected to the 128-dimensional Transformer hidden space, whereas the liveness generator produces a 128-dimensional acoustic representation.

B. LIVENESS MOTION GENERATOR

Speech is correlated with movements of the jaw and lips, whereas upper-face movements, including those of the eyebrows and eye region, are less directly related to speech. Consequently, the same speech signal can correspond to different plausible upper-face movements, which cannot be adequately represented by a deterministic mapping. To model this uncertainty, we apply a temporal VAE to learn the probability distribution of speech-conditioned upper-face motion. The temporal latent variables capture motion variations that cannot be fully predicted from speech. By sampling latent sequences from the learned audio-conditioned prior, the model can generate diverse and natural liveness motions.

As illustrated in Fig. 2, the liveness motion generator consists of a posterior encoder, an audio-conditioned prior network, and a temporal motion decoder. The audio encoder transforms the normalized log-Mel spectrogram into a frame-aligned acoustic representation $\mathbf{A}_{1:T}$.

The posterior encoder and the audio-conditioned prior network each use two temporal convolutional layers with a hidden dimension of 256 and a kernel size of 5. The temporal latent representation has 128 dimensions and a temporal resolution of 3 Hz, obtained by applying a stride of 10 to the 30-fps motion sequence. The motion decoder subsequently generates upper-face movements using four temporal convolutional layers with a hidden dimension of 256 and a kernel size of 5.

During training, the posterior encoder receives the acoustic representation and the ground-truth upper-face motion as inputs and predicts the mean and log-variance that parameterize the posterior distribution $q_\phi(\mathbf{z}_{1:T_z}^l \mid \mathbf{X}_{1:T}^u, \mathbf{A}_{1:T})$. Using the acoustic representation, the audio-conditioned prior network similarly predicts the mean and log-variance that parameterize the prior distribution $p_\theta(\mathbf{z}_{1:T_z}^l \mid \mathbf{A}_{1:T})$. The sampled latent sequence is upsampled to the motion frame rate and decoded with the frame-level acoustic representation to generate the upper-face liveness motion.

1) Discriminator

The discriminator takes the nine-channel motion sequence as input and does not use audio features. The temporal resolution is reduced by a factor of four using two stride-2 convolutional layers, and the final convolution outputs real–fake scores over local temporal windows.

2) Training Objective

For the liveness generator, the position, velocity, and acceleration reconstruction losses are computed as L_1 distances between the predicted and ground-truth motion sequences and their first- and second-order temporal differences. The temporal latent distribution is regularized using a KL-divergence loss between the approximate posterior and the audio-conditioned prior. We apply a small free-bits allowance [31] and gradually increase the KL weight during training [32] to encourage the model to utilize the temporal latent variables.

In addition, a prior-matching loss is used to train the audio-conditioned prior to approximate the posterior distribution:

$$\mathcal{L}_{\mathrm{prior}}^l = D_{\mathrm{KL}}\left(\mathrm{sg}[q_\phi] \,\|\, p_\theta\right), \tag{8}$$

where q_ϕ and p_θ denote the approximate posterior and the audio-conditioned prior, and $\mathrm{sg}[\cdot]$ denotes the stop-gradient operation.

For adversarial training, the discriminator distinguishes real upper-face motion sequences from motions generated using samples from the prior distribution. The liveness generator is optimized using the combined reconstruction, KL-divergence, prior-matching, and adversarial objectives:

$$\mathcal{L}_{G_l} = \mathcal{L}_{\mathrm{rec}}^l + \beta(s)\mathcal{L}_{\mathrm{KL}}^t + \lambda_{\mathrm{prior}}\mathcal{L}_{\mathrm{prior}}^l + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}}^G. \tag{9}$$

C. EMOTION-CONDITIONED UPPER-FACE GENERATOR

In this study, we adopt a regression-based model to generate emotion-specific facial movements from temporal acoustic information and emotional conditions. As shown in Fig. 3, the proposed generator takes frame-level acoustic features, emotion labels, emotion intensity, and speaker style as inputs. These conditioning variables are encoded into a single condition vector and injected into each Transformer block through AdaLN-Zero modulation [33]. Based on the inputs, the model predicts facial movements that are temporally aligned with the speech signal and consistent with the intended emotion. Subsequently, the generated emotion-related

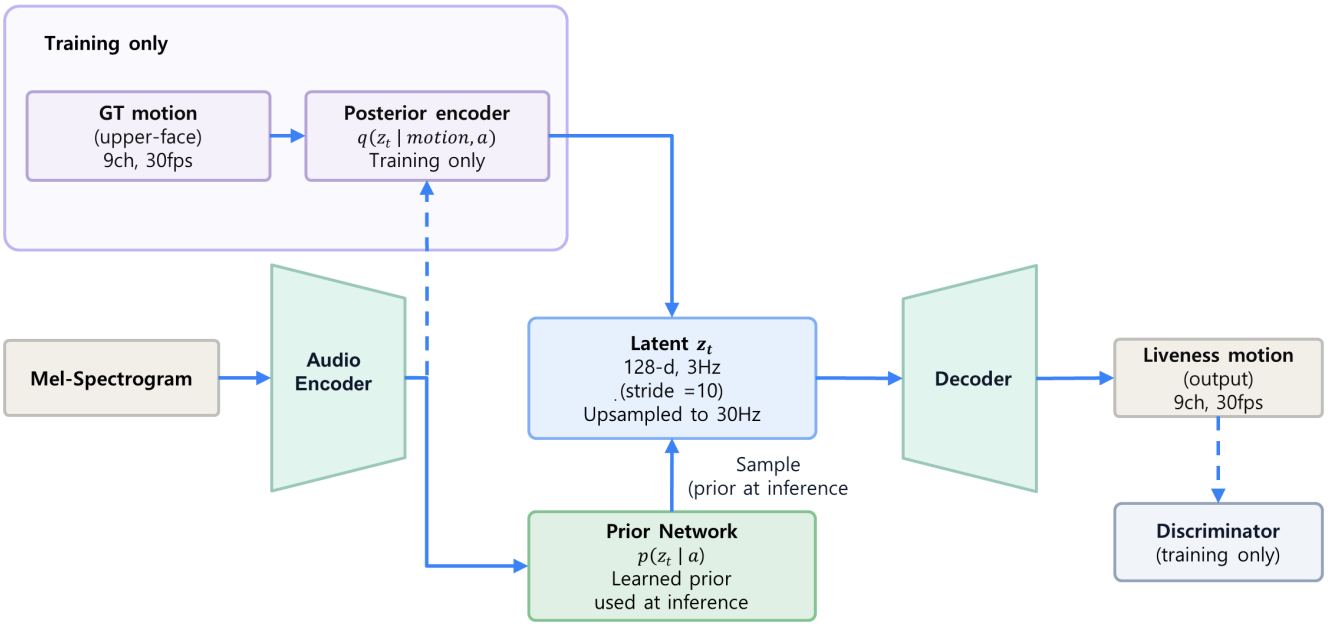

FIGURE 2. The posterior encoder is used only during training, whereas the learned audio-conditioned prior generates the latent sequence during inference. The 3-Hz latent sequence is upsampled to 30 Hz and decoded together with frame-level acoustic features to produce nine-channel upper-face liveness motion. The discriminator is used only during adversarial training.

motion sequence is fused with subtle liveness movements to produce the final facial animation.

The acoustic representation is projected into a 128-dimensional hidden space and then processed by four AdaLN-Zero Transformer blocks, each with four attention heads. Emotion, emotion intensity, and speaker style are represented using embeddings of dimensions 24, 16, and 24, which are concatenated into a single conditioning vector. The Transformer output is further refined by two temporal convolutional layers with kernel sizes of 5 and 3. Finally, a 1×1 convolutional motion head generates the upper-face emotion motion sequence. The emotion-conditioned generator predicts emotion-related movements around the eyebrows and eyelids. Its output consists of AU01, AU02, AU04, AU07, and an eye-widening feature, which are converted into nine upper-face motion channels for facial retargeting. The AU05 (Upper Lid Raiser) estimates did not sufficiently capture eye-widening variations in the MEAD dataset. Therefore, we replaced AU05 with a geometric eye-widening feature derived from MediaPipe facial landmarks.

1) Eye-widening feature

Eye widening was quantified separately for the left and right eyes using geometric descriptors based on the normalized distance between the upper and lower eyelids.

As illustrated in Fig. 4, the horizontal eye width w_t was measured as the distance between the inner and outer eye-corner landmarks, and the vertical eyelid opening h_t was measured as the distance between the upper- and lower-eyelid landmarks.

$$r_t^{\text{wide},L} = \frac{h_t^L}{w_t^L + \epsilon}, \qquad r_t^{\text{wide},R} = \frac{h_t^R}{w_t^R + \epsilon}. \tag{10}$$

where h_t^L and h_t^R denote the average vertical openings of the left and right eyes, and w_t^L and w_t^R denote their horizontal widths.

To reduce the effect of individual differences in neutral eye shape, we subtracted each subject's mean neutral baseline, calculated from the neutral-expression videos, and set negative residuals to zero.

$$\mu_i^{\text{neu},L} = \frac{1}{N_i^{\text{neu}}} \sum_{t \in \mathcal{N}_i} r_{i,t}^{\text{wide},L}, \tag{11}$$

$$\mu_i^{\text{neu},R} = \frac{1}{N_i^{\text{neu}}} \sum_{t \in \mathcal{N}_i} r_{i,t}^{\text{wide},R}, \tag{12}$$

$$d_{i,t}^{\text{wide},L} = \max\left(r_{i,t}^{\text{wide},L} - \mu_i^{\text{neu},L}, 0\right), \tag{13}$$

$$d_{i,t}^{\text{wide},R} = \max\left(r_{i,t}^{\text{wide},R} - \mu_i^{\text{neu},R}, 0\right). \tag{14}$$

where $\mathcal{N}_i$ denotes the set of valid frames in the neutral-expression videos of subject i, and $N_i^{\text{neu}} = |\mathcal{N}_i|$.

The positive residuals were subsequently divided by the 99th percentile computed from the training set and clipped to the range [0, 1].

2) Training Objective

For the emotion-conditioned generator G_e, we employ position, velocity, and acceleration losses. These losses are com-

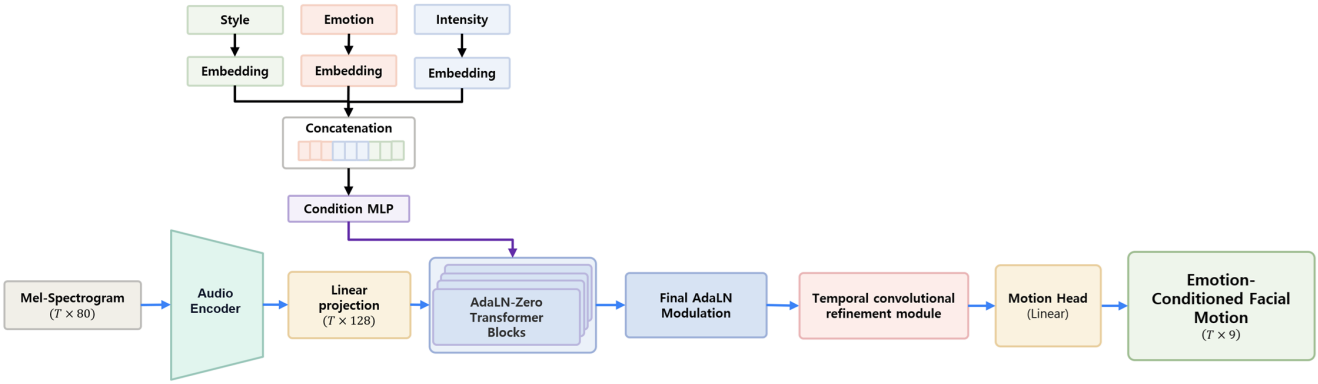

FIGURE 3. Architecture of the emotion-conditioned upper-face generator using AdaLN-Zero Transformer blocks and temporal convolutional refinement.

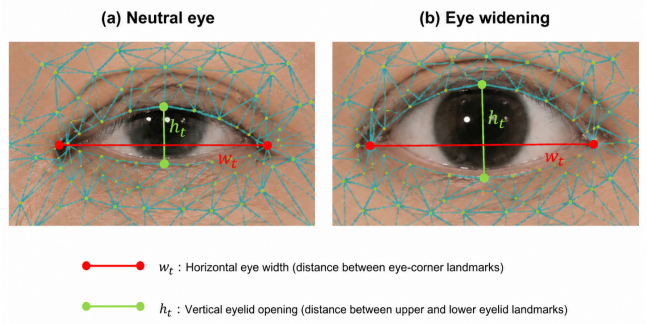

FIGURE 4. Geometric measurement of eye widening.

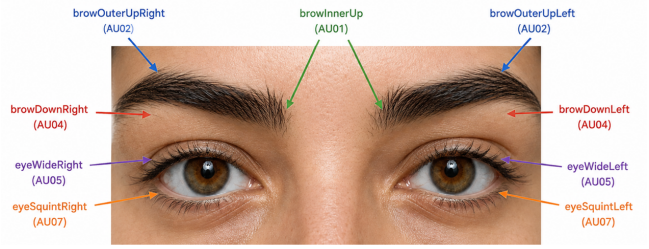

FIGURE 5. Nine ARKit-compatible upper-face motion channels and their corresponding Facial Action Units. Left and right refer to the anatomical sides of the face.

puted using L_1 distances between the predicted and ground-truth motion sequences and their first- and second-order temporal differences. The overall objective of G_e is

$$\mathcal{L}_{G_e} = \mathcal{L}^e_{\text{pos}} + \lambda^e_{\text{vel}}\mathcal{L}^e_{\text{vel}} + \lambda^e_{\text{acc}}\mathcal{L}^e_{\text{acc}}. \tag{15}$$

D. TRAINING DATA USAGE

We treat emotional upper-face expressions and natural upper-face movements as two distinct components. The emotion-conditioned generator is trained on MEAD, whereas the liveness generator is trained separately on BEAT. MEAD is a talking-face video dataset featuring eight emotion categories at three intensity levels [34]. The speech corpus consists of common, generic, and emotion-related sentences, enabling the model to learn from acoustic variation, emotion category, emotion intensity, and speaker style. MEAD is used to train G_e, which generates emotional upper-face animation. We extracted five upper-face Action Unit features from the MEAD videos using OpenFace and converted them into facial motion sequences. To match the ARKit 52 blendshape representation [35] used for virtual-agent animation, four of the five features were mapped to separate left and right channels (e.g., browDownLeft and browDownRight), except for AU01 (inner-brow raising), which was represented by the ARKit channel browInnerUp. Previous evaluations of automated AU detection systems reported lower detection performance for AU05 than for other AUs [36]. To improve the estimation of eye widening, we derived the corresponding motion channels from MediaPipe [37] facial landmarks. The resulting nine upper-face motion channels and their corresponding Facial Action Units are illustrated in Fig. 5. BEAT is a multimodal motion-capture dataset containing synchronized speech, facial motion, and full-body motion sequences [38]. Its facial motion is represented using 52 ARKit blendshape channels. We selected nine upper-face channels from the 52 ARKit blendshape channels. These channels are identical to the output channels of the emotion-conditioned generator and were used to train G_l.

IV. QUANTITATIVE EVALUATION

Upper-face movements are not fully determined by the speech signal alone. Even for the same utterance, different upper-face movements may be produced depending on speaker style, emotional state, and natural behavioral variation. Thus, it is difficult to determine whether an upper-face animation is superior or more natural based on a lower reconstruction error relative to a single reference sequence. In particular, a deterministic regression model may generate weak and static movements compared with real facial motion by producing

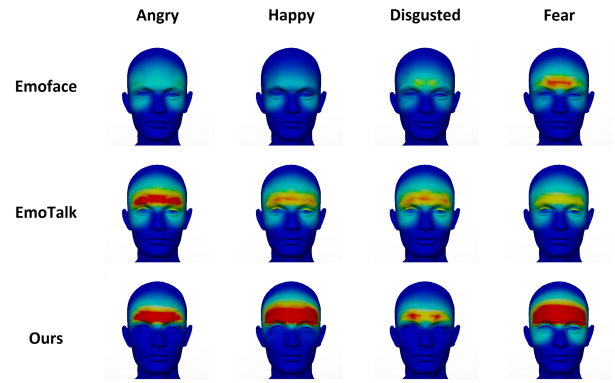

FIGURE 6. Spatial distribution of upper-face motion generated by EmoFace, EmoTalk, and the proposed method across four emotion conditions. Warmer colors indicate regions with greater facial deformation or motion magnitude. A CC0-licensed 3D face model from Filmic Worlds [39] was used for visualization.

near-average predictions to reduce positional error.

Thus, the quantitative evaluation focuses on four aspects. First, audio–visual synchronization is evaluated to determine whether it is maintained when lip movements are generated by an external real-time module. Second, the generated upper-face movements are analyzed to determine whether their motion magnitude and speed are comparable to real facial movements, rather than excessively static or exaggerated. Third, the integration of emotional and liveness-related movements is assessed to determine whether it compensates for the lack of motion energy in the emotion-conditioned generator's output. Finally, the computational efficiency of the proposed module for real-time interactive agents is evaluated in terms of the number of model parameters and inference time. Based on these considerations, we report reconstruction-based metrics, motion statistics, emotion-preservation measures, and computational efficiency rather than relying on a single metric.

A. EXPERIMENT SETUP

1) Datasets

For component-level evaluation, the emotion-conditioned generator was evaluated on the held-out MEAD test set, whereas the liveness generator was evaluated on the held-out BEAT test set. Both datasets were split at the speaker level to ensure that speakers in the test set did not appear in the training set. For the system-level comparison, we used RAVDESS [40] as an external evaluation dataset. Four RAVDESS emotion conditions were selected: angry, happy, fearful, and disgusted. The same emotion set was used in the user study to maintain consistent conditions between the quantitative and perceptual evaluations. All compared methods were evaluated using the same speech utterances and rendering conditions. EmoFace produces 174 MetaHuman facial controller values, whereas EmoTalk produces 52 blendshape coefficients. For quantitative comparison, the same nine upper-face channels were selected from their outputs.

2) Implementation Details

Both generators were trained with the AdamW optimizer [41] at a learning rate of 2×10^{-4}. The emotion-conditioned generator was trained for 40 epochs with a batch size of 16. The velocity and acceleration loss weights were set to $\lambda^{e}_{\mathrm{vel}} = 1.0$ and $\lambda^{e}_{\mathrm{acc}} = 0.2$. The liveness generator was trained for 60 epochs with a batch size of 64. The velocity and acceleration loss weights were set to $\lambda^{l}_{\mathrm{vel}} = 1.5$ and $\lambda^{l}_{\mathrm{acc}} = 1.25$. The prior-matching and adversarial loss weights were set to $\lambda_{\mathrm{prior}} = 1.0$ and $\lambda_{\mathrm{adv}} = 0.01$. The KL weight was increased from 0 to 0.02 over 10,000 steps after an initial 500-step warm-up, with free bits set to 0.02 per temporal latent dimension. The discriminator was trained using a binary cross-entropy objective. Ground-truth motion was used as real samples, while motion generated from the audio-conditioned prior was used as fake samples. Both the generator and discriminator were updated once per training step.

Both generators were trained on an NVIDIA RTX A6000 GPU. For the final system, the fusion scaling coefficient was fixed at $\alpha = 0.5$.

B. COMPONENT-LEVEL BEHAVIOR OF THE TWO GENERATORS

TABLE 1. Quantitative evaluation of the proposed generators on their respective held-out test sets.

Generator	MAE ↓	Velocity MAE ↓	Acceleration MAE ↓	Velocity ratio → 1
Emotion generator	0.1475	0.0327	0.0241	0.286
Liveness generator	0.1290	0.0153	0.0104	1.372

Table 1 presents the results of independently evaluating each generator on its respective speaker-independent held-out test set. The emotion-conditioned generator was evaluated on the MEAD test set, whereas the liveness generator was evaluated on the BEAT test set. As the two generators were trained on different datasets and designed for different generation objectives, the results should not be compared directly between them.

The emotion-conditioned generator achieved an MAE of 0.1475. However, its velocity ratio was only 0.286, indicating that the generated motion magnitude was approximately 28.6% of the reference motion. This suggests that the generator can reproduce the overall emotional expression but tends to generate weaker temporal facial movements. The liveness generator achieved a velocity ratio of 1.372, indicating that its average frame-to-frame motion magnitude was approximately 37.2% greater than that of the reference motion. Accordingly, a scaling coefficient α was used to control the contribution of the liveness motion and prevent exaggerated movements during fusion.

TABLE 2. Comparison of upper-face fusion strategies at $\alpha = 0.50$ on the RAVDESS evaluation set ($n = 352$).

Method	Emotion preservation MAE ↓	Mean abs. vel. ratio → 1	Clipping ↓
Emotion only	–	0.743	0.0%
Direct addition $\mathbf{R}_t = \mathbf{L}_t$	0.0640	1.214	25.3%
Signed residual $\mathbf{R}_t = \mathbf{L}_t - \bar{\mathbf{L}}$	0.0245	1.095	35.4%
Positive residual $\mathbf{R}_t = \max(\mathbf{L}_t - \bar{\mathbf{L}}, 0)$	**0.0153**	**1.014**	**10.0%**

Emotion preservation MAE is measured against the emotion-only output. The velocity ratio compares mean absolute frame-to-frame motion with the reference motion. Clipping denotes the proportion of pre-clipping values outside $[0, 1]$.

TABLE 3. Audio–visual synchronization on RAVDESS (100 clips per condition). Values are mean (SD). Real video is a reference point, not a competing condition: SyncNet is trained on photographic faces, so absolute scores on rendered virtual characters are not comparable to it.

Method	LSE-D ↓	LSE-C ↑	Offset
Real video	*7.775 (1.27)*	*3.863 (1.74)*	*−2.09 (1.53)*
EmoTalk [7]	8.372 (0.69)	0.935 (0.38)	−3.85 (2.75)
EmoFace [8]	8.007 (0.79)	0.749 (0.33)	−0.84 (6.00)
Oculus LipSync [42]	**7.506 (0.61)**	**1.145 (0.31)**	**−1.19** (1.22)

Table 2 reports the results of comparing three simple fusion methods for efficiently integrating the emotion and liveness motion sequences without increasing computational complexity. All methods use the same outputs from the two generators and differ only in the fusion operation. $\mathbf{L}_t = \hat{\mathbf{X}}_t^{u,\text{live}}$ denotes the liveness motion at frame t, and $\bar{\mathbf{L}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{L}_t$ denotes its temporal mean. The emotion preservation MAE measures the absolute difference between the emotion-conditioned output and the fused result, with lower values indicating better preservation of the original emotional expression. The velocity ratio indicates the motion velocity relative to the reference motion, while the clipping ratio represents the proportion of output values that exceed the valid range $[0, 1]$ before clipping. Direct addition showed the largest deviation from the emotion-conditioned output and frequent clipping. The signed residual method also exhibited a high clipping ratio, making it less suitable for stable fusion. In contrast, the positive residual method achieved the lowest emotion preservation MAE (0.0153), a velocity ratio closest to the reference value (1.014), and the lowest clipping ratio (10.0%). Therefore, it was adopted in the final system because it minimizes deviation from the emotion-conditioned output while maintaining motion dynamics close to the reference motion, without requiring an additional trainable fusion model.

Table 3 summarizes the lip-synchronization results from 100 RAVDESS clips for each condition. In the proposed pipeline, the proposed generators produce upper-face motion, and an external real-time lip-sync module generates lip motion. Lip synchronization is evaluated using LSE-D and LSE-C from SyncNet [43], [44]. The external module achieved the best LSE-D and the highest LSE-C among the animated conditions, indicating that its use does not degrade temporal synchronization. The original RAVDESS videos are included only as a reference.

TABLE 4. Quantitative comparison on the RAVDESS evaluation set. MACE measures the maximum channel-wise reconstruction error, while MAE measures the overall reconstruction error. Motion amplitude and velocity ratios are optimal near 1.

Method	MACE ↓	MAE ↓	Velocity MAE ↓	Motion amplitude ratio → 1	Velocity ratio → 1
EmoFace	0.588	0.195	0.012	2.252	0.767
EmoTalk	**0.359**	**0.160**	0.013	0.986	**1.071**
Ours(Emotion)	0.668	0.250	**0.011**	0.724	0.584
Ours(Full)	0.673	0.261	0.013	**0.997**	0.787

TABLE 5. Comparison of model size and inference efficiency based on model forward-pass time and real-time factor (RTF). For the proposed method, inference time is reported as the sum of the emotion- and liveness-generator forward-pass times, excluding the external lip-sync module.

Method	Parameters (M)	GPU Time (ms/clip)	CPU Time (ms/clip)	GPU RTF	CPU RTF
EmoFace	126.77	11.38	101.30	0.0033	0.0284
EmoTalk	640.57	63.90	484.73	0.0177	0.1342
Emotion generator	1.45	3.69	6.10	–	–
Liveness generator	2.74	2.98	7.05	–	–
Ours (two-model sum)	**4.18**	**6.67**	**13.16**	**0.0018**	**0.0036**

Figure 6 illustrates the spatial distribution of upper-face motion under the four emotion conditions. The proposed method generated clear emotion-dependent spatial patterns, consistent with the quantitative motion analysis.

For the system-level comparison on RAVDESS, the metrics reported in Table 4 are used to evaluate reconstruction error and motion characteristics. Inspired by the emotional vertex error (EVE) used in EmoTalk, MACE measures the maximum absolute reconstruction error among the selected upper-face channels at each frame, averaged over the evaluation sequence, whereas MAE measures the overall absolute reconstruction error over all frames and channels. The emotion-conditioned generator produced lower motion amplitude and temporal velocity than the reference. After adding the liveness component, the motion amplitude and velocity ratios increased to 0.997 and 0.787. Among the compared methods, EmoTalk achieved the lowest MACE and a velocity ratio closest to 1, whereas the proposed full model achieved the motion amplitude ratio closest to that of the reference videos. The MACE-based metric measures only frame-level errors relative to a single standardized reference sequence extracted using OpenFace and is therefore interpreted as a supplementary indicator.

TABLE 6. Mean user ratings and Friedman test results for the four facial animation methods ($N = 36$).

Measure	EmoFace	Ours	EmoTalk	Ours (Emotion Only)	p
Emotional Expression	2.87	3.01	2.12	**3.06**	$< .001$
Anthropomorphism	2.31	2.46	1.91	**2.52**	.057
Animacy	2.56	**2.85**	2.18	2.72	$< .001$
Likeability	2.73	**2.75**	2.35	2.68	.379
Overall Average	2.62	**2.77**	2.14	2.75	.009

Ratings were collected using a five-point Likert scale. Each measure represents the mean of three questionnaire items. The p-values were obtained using Friedman tests. Bold values indicate the highest mean in each row.

C. MODEL SIZE AND INFERENCE EFFICIENCY

Table 5 compares the parameter count and inference cost of each method. The proposed emotion and liveness generators contain a total of 4.18M trainable parameters, which is approximately 30 times fewer than EmoFace and 150 times fewer than EmoTalk. The reported parameter count and inference time exclude the external lip-sync module. Inference time was measured on a dual-socket Ubuntu server equipped with two AMD EPYC 9124 16-core processors and an NVIDIA RTX A6000 GPU. The proposed model achieved a CPU RTF of 0.0036, approximately 1/8 of the RTF of EmoFace and 1/37 of that of EmoTalk. In addition, the proposed method achieved an RTF of 0.0080 (SD $= 0.0009$) on a Windows 11 desktop equipped with an AMD Ryzen Threadripper 2950X processor. Even when execution was restricted to a single thread, it maintained an RTF of 0.0146 (SD $= 0.0005$).

V. USER STUDY

All participants in the two user studies reported below provided informed consent before participation, and both studies received an exemption determination from the Institutional Review Board of the Korea Institute of Science and Technology (KIST-202607-HR-001).

To evaluate upper-face emotional expression in a virtual character, short video clips were generated using only the nine upper-face motion channels. The area below the nose was covered so that participants evaluated only the upper-face expression. Fig. 7 shows example video clips used in the user study. The clips used RAVDESS speech samples from four emotional conditions: happy, disgusted, angry, and fearful. Four emotion conditions were selected to keep the session length manageable under the within-subject design, in which each participant evaluated all four methods: Ours, Ours (Emotion Only), EmoFace, and EmoTalk. The questionnaire was adapted from the Godspeed Questionnaire Series (GQS) [45], with minor modifications, and included three items for each of the four measures. The user study was conducted with 40 adults ranging in age from their 20s to their 60s who were recruited through Prolific. A cyclic Latin square design was used to reduce potential order effects. Four participants who failed the attention-check question were excluded, and the remaining 36 participants were included in the final analysis. Table 6 presents the mean ratings and Friedman test results. Table 7 presents the significant post-hoc comparisons.

TABLE 7. Significant post-hoc comparisons following the Friedman tests using two-sided Wilcoxon signed-rank tests with Holm correction [46].

Measure	Higher-rated method	Lower-rated method	Holm-adjusted p
Emotional Expression	EmoFace	EmoTalk	.001
	Ours	EmoTalk	.004
	Emotion Only	EmoTalk	$< .001$
Animacy	Ours	EmoTalk	.005
	Emotion Only	EmoTalk	.021
Overall Average	EmoFace	EmoTalk	.009
	Ours	EmoTalk	.008
	Emotion Only	EmoTalk	.008

Significant differences among the methods were found for Emotional Expression, Animacy, and the Overall Average, but not for Anthropomorphism or Likeability. All significant post-hoc differences were found in comparisons with EmoTalk. The proposed method received significantly higher ratings than EmoTalk for Emotional Expression, Animacy, and the Overall Average, but showed no significant differences from EmoFace. In particular, it achieved the highest Animacy rating (2.85). These results suggest that the proposed method can generate upper-face animation sequences with quality comparable to EmoFace using fewer parameters, including both emotional expressions and natural movements. No significant differences were found between EmoFace and the proposed method in the post-hoc comparisons. Although the full model received the highest mean Animacy rating (2.85), it did not differ significantly from the emotion-conditioned generator. However, EmoTalk generates emotional expressions from acoustic features, making it difficult to interpret its lower ratings as evidence of inferior quality. EmoTalk's audio-driven approach is useful for automatically reflecting emotions inherent in speech, but it may be less suitable for conversational virtual agents that require explicit emotion control.

A. ADDITIONAL PERCEPTUAL EVALUATION

For an additional perceptual evaluation, 20 participants evaluated full-face animations, and Table 8 summarizes the mean ratings and Friedman test results. The proposed method used nine upper-face channels and 17 mouth-related blendshape channels mapped from Oculus LipSync outputs. In the manually specified mouth-expression condition, mouthSmileLeft/Right were set to 0.3 for the happy condition, whereas mouthFrownLeft/Right were set to 0.3 for angry, fearful, and disgusted conditions. For disgust, mouthLowerDownLeft/Right were additionally set to 0.2. EmoTalk used all 52 ARKit blendshape outputs, while EmoFace was retargeted from 174 MetaHuman controller values to 52 blendshape channels. All methods were compared under the same rendering conditions.

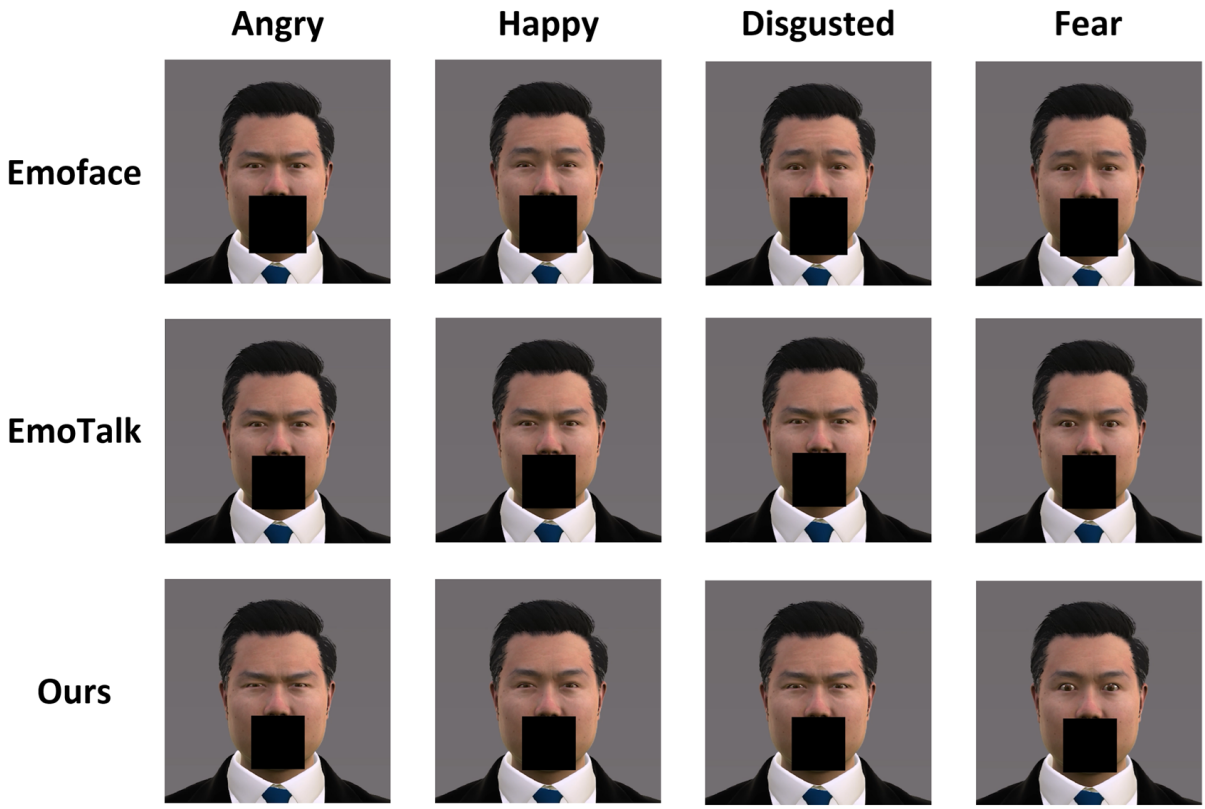

FIGURE 7. Examples used in the user study. Each column shows an emotion condition, and each row shows a method. The mouth region was masked so that participants evaluated only the upper-face expression.

TABLE 8. Mean user ratings and Friedman test results for the four facial-animation methods ($N = 20$).

Measure	Ours (Manual Mouth)	EmoTalk	Ours	EmoFace	p
Emotional Expression	**3.63**	2.82	3.18	3.07	**.033**
Anthropomorphism	**2.80**	2.02	2.53	2.30	**.037**
Animacy	**3.03**	2.57	2.87	2.70	.061
Likeability	**3.42**	2.78	2.85	2.92	**.006**
Overall Average	**3.22**	2.55	2.86	2.75	.122

Ratings were collected using a five-point Likert scale. Each perceptual measure represents the mean of three questionnaire items. The p-values were obtained using Friedman tests. Bold values indicate the highest mean in each row.

TABLE 9. Significant post-hoc comparisons following the Friedman tests using two-sided Wilcoxon signed-rank tests with Holm correction.

Measure	Higher-rated method	Lower-rated method	Holm-adjusted p
Emotional Expression	Ours (Manual Mouth)	EmoTalk	.047
Anthropomorphism	Ours (Manual Mouth)	EmoTalk	.019
Likeability	Ours (Manual Mouth)	EmoFace	.049

Pairwise post-hoc comparisons were conducted using two-sided Wilcoxon signed-rank tests with Holm correction for multiple comparisons. The results are presented in Table 9. Significant differences were found between Ours (Manual Mouth) and EmoTalk for Emotional Expression and Anthropomorphism. For Likeability, a significant difference was found between Ours (Manual Mouth) and EmoFace. This result suggests that not only upper-face motion but also mouth expressions can contribute to perceptual quality. The current proposed system separately processes upper-face emotional motion and lip synchronization using an external module. In future work, emotion-related mouth features, such as mouth-corner movements and mouth shapes, should be integrated to achieve more natural coordination between emotional expression and lip synchronization.

VI. ABLATION

We evaluated the liveness generator in terms of deterministic versus latent-variable modeling, continuous VAE versus discrete VQ-VAE [47] representations, and the use of adversarial training. A diverse range of plausible upper-face motion sequences can be generated from the same speech input, making it difficult to assess generation quality using reconstruction error alone. Therefore, in addition to MACE and MAE, we evaluated the temporal characteristics of the generated motion

TABLE 10. MACE (Maximum Absolute Channel Error) captures the largest channel-wise reconstruction error, while MAE (Mean Absolute Error) measures the overall reconstruction error. The HF (high-frequency) and velocity ratios are optimal near 1.

Variant	MACE ↓	MAE ↓	HF ratio → 1	Velocity ratio → 1
Ground truth	-	-	*1.000*	*1.000*
Deterministic regression	0.374	0.188	0.637	**1.424**
VQ-VAE latent	0.426	0.198	0.495	0.499
w/o adversarial training	0.391	0.194	0.785	1.845
Full	**0.327**	**0.170**	**0.983**	1.650

TABLE 11. Comparison of conditioning strategies for the emotion generator on the MEAD held-out test split. MACE and MAE are frame-wise errors against the reference; the velocity ratio compares the frame-to-frame motion magnitude and is best near 1.

Method	MACE ↓	MAE ↓	Velocity ratio → 1
AdaLN	**0.561**	**0.165**	**0.469**
Concat.	0.578	0.169	0.407

using the velocity ratio and HF ratio. The frequency-band boundaries were determined from the spectral distribution of 1,000 BEAT test clips. The cumulative reference power reached 95% at 3.53 Hz; therefore, 5 Hz was selected as a practical upper bound that fully covers the dominant motion-frequency range. For the lower bound, 68% of the spectral power was concentrated below 1 Hz. To increase sensitivity to faster temporal changes, this low-frequency component was excluded by setting the lower bound to 1.5 Hz.

Table 10 presents the ablation results. The VQ-VAE variant showed the highest MACE and the lowest velocity ratio, suggesting that the discrete latent representation may have limitations in modeling continuous temporal motion. Without adversarial training, the HF ratio decreased from 0.983 to 0.785, and the velocity ratio increased from 1.650 to 1.845. The full model achieved the lowest MACE and MAE and an HF ratio closest to the reference value. We compared two conditioning strategies for the emotion generator: injecting the emotion and intensity conditions into each Transformer block through AdaLN-Zero modulation and concatenating them with the input features. The evaluation was conducted on the MEAD test split, which was not used for training or validation. Table 11 presents the results. AdaLN-Zero achieved MACE and MAE values of 0.561 and 0.165, which were lower than those of the concatenation-based Transformer. The velocity ratio of AdaLN-Zero was also closer to the ground-truth value.

VII. DISCUSSION

A. IMPLICATIONS OF SEPARATING LIP-SYNC AND UPPER-FACE GENERATION

The results support designing lip-sync and upper-face generation as separate modules rather than using a full-face animation model for interactive virtual agents. Although lip movements were handled by an external real-time module, the proposed system achieved comparable audio–visual synchronization performance (Table 3). The upper-face generator used only 4.18M parameters, corresponding to approximately $1/30$ of EmoFace and $1/150$ of EmoTalk. However, the comparison uses the total parameter count, which includes the Wav2Vec 2.0 [48] components of EmoFace and EmoTalk. No significant post-hoc differences were observed between the proposed method and EmoFace. These results suggest that a substantial reduction in model size is possible without a significant perceptual difference from EmoFace.

The proposed system processed each audio clip substantially faster than its duration on the CPU, suggesting that it can operate without GPU resources. However, the system requires a complete audio clip before generating the corresponding motion sequence.

B. DESIGN OF THE TWO-GENERATOR ARCHITECTURE

The quantitative evaluation results support the use of separate emotion-conditioned and liveness generators. The emotion-conditioned regression model reproduced the intended expressions. However, it generated less motion than the reference. The same tendency was observed with different conditioning methods. To address this limitation, the liveness generator was added to make the overall motion magnitude closer to the reference. The mean-centered positive residual method preserved the emotional expression while compensating for the lack of motion. In contrast, directly adding the outputs of the two generators distorted the emotional expression and caused output saturation.

C. ABLATION STUDY

The ablation study suggests that adversarial training contributes to preserving the temporal characteristics of upper-face motion. Without adversarial training, motion energy within the defined high-frequency band decreased, although the overall motion magnitude did not. This result shows that motion magnitude alone is insufficient to assess the temporal structure of a generated sequence. The VQ-VAE-based liveness model showed reductions in both overall motion and motion energy within the defined high-frequency band, suggesting that the discrete latent representation may limit the generation of subtle temporal transitions.

D. LIMITATIONS

The results of the proposed system should be interpreted considering the following limitations. First, the proposed system performs inference at the clip level. It generates the complete facial animation sequence after receiving an audio clip as input. This approach is suitable for pipelines that play the audio and facial animation simultaneously. However, it does

not allow the generated animation to be modified immediately during playback. Second, eye blinking and head rotation were not included in the output of the proposed system. Eye blinking is weakly related to speech and emotion conditions, and both blinking and head rotation can substantially affect frame-to-frame motion statistics. Therefore, these components were excluded from the present evaluation. However, both are essential for conversational virtual agents and should be addressed in future work. Third, the upper-face movements used in this study were estimated from video data rather than directly captured using motion-capture equipment. The reference motion was extracted using OpenFace action units. The training targets for the emotion-conditioned generator were also derived from OpenFace features, and some components were computed using MediaPipe landmarks. Therefore, the reported errors, ratios, and other metrics should be interpreted as relative differences under the same extraction conditions rather than as measures of absolute motion accuracy. Finally, the frame-to-frame variation remained relatively small compared with the variation in the reference motion, although the overall motion magnitude was restored to a similar level. The positive residual method may have contributed to this result, which suggests that the current fusion strategy may not fully preserve temporal motion. Therefore, a key next step is to develop a more advanced fusion method that preserves both emotional expression and temporal motion.

E. FUTURE WORK

First, the current system generates only nine upper-face motion channels. Future work should expand the output channels to include more detailed facial movements, such as eye blinking and head rotation. This extension would enable the system to model upper-face and head movements together and generate more natural animation for interactive virtual agents.

Second, a more advanced extraction method is needed to build accurate and consistent training targets efficiently. The current approach uses OpenFace action units and MediaPipe landmarks, which may introduce estimation errors during feature extraction. Therefore, future work should improve landmark-based calculation methods or develop a new method for extracting facial movements from video data.

Lastly, the system should be extended to generate mouth shapes that convey emotion in addition to using an external lip-sync module. Mouth movements are also important for emotional expression. Therefore, a fusion method is needed to combine speech-driven lip movements with emotion-expressive mouth shapes.

VIII. CONCLUSION

In this study, we proposed a lightweight speech-driven upper-face animation framework. Compared with existing full-face generation frameworks, the proposed system processes lip motion and upper-face motion separately and directly generates nine upper-face channels corresponding to eyebrow and eye-region movements. The upper-face animation module consists of an emotion-conditioned generator and a liveness generator, whose outputs are fused using a positive mean-centered residual method. This design preserves emotion-related facial expressions and compensates for insufficient upper-face motion. In addition, the modular architecture enables the generated upper-face animation to be combined with an external real-time lip-synchronization module. This architecture offers an efficient approach to talking-face animation. In the evaluation, integrating the output of the liveness generator with that of the emotion-conditioned generator increased the upper-face motion amplitude ratio from 0.724 to 0.997, bringing it close to the reference motion amplitude observed in the RAVDESS data. Among the evaluated fusion methods, the positive residual method showed the lowest error relative to the emotion-conditioned output and a velocity ratio of 1.014, which was closest to 1. In addition, the two proposed generators contain only 4.18M parameters in total. For approximately 3-s speech inputs, the system achieved a CPU real-time factor (RTF) of 0.0036, which demonstrates lower computational cost than the other methods evaluated. In the user study, the proposed method achieved the highest mean Animacy rating and showed no significant difference from EmoFace. In the additional perceptual evaluation, emotion-related mouth shapes were manually specified in addition to the generated upper-face motion and lip-synchronization output. Significant differences were observed in some perceptual measures, which suggests that natural emotional facial expression may require consideration of not only upper-face motion but also emotion-related mouth movements. Overall, this study suggests that emotional upper-face expressions and natural facial motion can be generated using lightweight models, while lip synchronization can be handled separately by an external real-time lip-sync module. This modular design therefore provides a practical alternative for real-time talking-face animation systems. In future work, the framework should be extended to model additional facial and head movements such as eye blinks and head rotation. Emotion-related mouth shapes and mouth-corner movements should also be integrated more naturally with the upper-face animation.

ACKNOWLEDGMENT

Figs. 4 and 5 were generated using OpenAI ChatGPT. The authors modified and annotated the figures for technical illustration and explanatory purposes.

REFERENCES

[1] J. Cassell, "Embodied conversational agents: Representation and intelligence in user interfaces," *AI Mag.*, vol. 22, no. 4, pp. 67–83, 2001.
[2] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People*. Cambridge, U.K.: Cambridge Univ. Press, 1996.
[3] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, 2017.
[4] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "FaceFormer: Speech-driven 3D facial animation with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 18770–18780.
[5] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique*

for the Measurement of Facial Movement. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[6] A. Richard, M. Zollhöfer, Y. Wen, F. de la Torre, and Y. Sheikh, "MeshTalk: 3D face animation from speech using cross-modality disentanglement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 1173–1182.

[7] Z. Peng, H. Wu, Z. Song, H. Xu, X. Zhu, J. He, H. Liu, and Z. Fan, "EmoTalk: Speech-driven emotional disentanglement for 3D face animation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 20687–20697.

[8] C. Liu, Q. Lin, Z. Zeng, and Y. Pan, "EmoFace: Audio-driven emotional 3D face animation," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces (VR)*, 2024, pp. 387–397.

[9] F. I. Parke, "Computer generated animation of faces," in *Proc. ACM Annu. Conf.*, vol. 1, 1972, pp. 451–457.

[10] M. M. Cohen and D. W. Massaro, "Modeling coarticulation in synthetic visual speech," in *Models and Techniques in Computer Animation*, N. Magnenat-Thalmann and D. Thalmann, Eds. Berlin, Germany: Springer, 1993, pp. 139–156.

[11] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "JALI: An animator-centric viseme model for expressive lip synchronization," *ACM Trans. Graph.*, vol. 35, no. 4, 2016.

[12] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2016, pp. 1–10.

[13] S. Chen, Z. Liu, J. Liu, Z. Yan, and L. Wang, "Talking head generation with audio and speech related facial action units," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2021.

[14] H. X. Pham, S. Cheung, and V. Pavlovic, "Speech-driven 3D facial animation with implicit emotional awareness: A deep learning approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 80–88.

[15] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 10101–10111.

[16] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 12780–12790.

[17] S. Stan, K. I. Haque, and Z. Yumak, "FaceDiffuser: Speech-driven 3D facial animation synthesis using diffusion," in *Proc. 16th ACM SIGGRAPH Conf. Motion, Interaction Games*, 2023, Art. no. 13.

[18] K. D. Yang, A. Ranjan, J.-H. R. Chang, R. Vemulapalli, and O. Tuzel, "Probabilistic speech-driven 3D facial motion synthesis: New benchmarks, methods, and applications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 27294–27303.

[19] R. Daněček, K. Chhatre, S. Tripathi, Y. Wen, M. J. Black, and T. Bolkart, "Emotional speech-driven animation with content-emotion disentanglement," in *SIGGRAPH Asia Conf. Papers*, 2023, pp. 1–13.

[20] Y. Chen, J. Zhao, and W.-Q. Zhang, "Expressive speech-driven facial animation with controllable emotions," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, 2023, pp. 387–392.

[21] K. Shen, H. Xia, G. Geng, G. Geng, S. Xia, and Z. Ding, "DEITalk: Speech-driven 3D facial animation with dynamic emotional intensity modeling," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 10506–10514.

[22] S. Wu, K. I. Haque, and Z. Yumak, "ProbTalk3D: Non-deterministic emotion controllable speech-driven 3D facial animation synthesis using VQ-VAE," in *Proc. 17th ACM SIGGRAPH Conf. Motion, Interaction Games*, 2024, Art. no. 15.

[23] F. Nocentini, C. Ferrari, and S. Berretti, "EmoVOCA: Speech-driven emotional 3D talking heads," in *Proc. Winter Conf. Appl. Comput. Vis. (WACV)*, 2025, pp. 2859–2868.

[24] J. Kim, J. Cho, J. Park, S. Hwang, D. E. Kim, G. Kim, and Y. Yu, "DEEPTalk: Dynamic emotion embedding for probabilistic speech-driven 3D face animation," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 4, pp. 4275–4283, 2025.

[25] C. Chung, I. Fedorov, M. Huang, A. Karmanov, D. Korobchenko, R. Ribera, Y. Seol, *et al.*, "Audio2Face-3D: Audio-driven realistic facial animation for digital avatars," *arXiv preprint arXiv:2508.16401*, 2025.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[27] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[28] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. 33rd Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 1558–1566.

[29] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein GANs," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[30] K. Vougioukas, S. Petridis, and M. Pantic, "Realistic speech-driven facial animation with GANs," *Int. J. Comput. Vis.*, vol. 128, no. 5, pp. 1398–1413, 2020.

[31] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, "Improved variational inference with inverse autoregressive flow," in *Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[32] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," in *Proc. 20th SIGNLL Conf. Comput. Natural Lang. Learn. (CoNLL)*, pp. 10–21, 2016.

[33] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4195–4205.

[34] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "MEAD: A large-scale audio-visual dataset for emotional talking-face generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 700–717.

[35] Apple Inc., "ARFaceAnchor.BlendShapeLocation," Apple Developer Documentation. [Online]. Available: https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation. [Accessed: Aug. 9, 2026].

[36] S. Namba, W. Sato, M. Osumi, and K. Shimokawa, "Assessing automated facial action unit detection systems for analyzing cross-domain facial expression databases," *Sensors*, vol. 21, no. 12, Art. no. 4222, 2021.

[37] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, *et al.*, "MediaPipe: A framework for building perception pipelines," *arXiv preprint arXiv:1906.08172*, 2019.

[38] H. Liu, Z. Zhu, N. Iwamoto, Y. Peng, Z. Li, Y. Zhou, E. Bozkurt, and B. Zheng, "BEAT: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 612–630.

[39] Filmic Worlds, "Solving Face Scans for ARKit," [Online]. Available: https://filmicworlds.com/blog/solving-face-scans-for-arkit/. [Accessed: Aug. 11, 2026].

[40] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, Art. no. e0196391, 2018.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[42] Meta, "Oculus Lipsync for Unity Development," Meta Horizon Developer Documentation. [Online]. Available: https://developers.meta.com/horizon/documentation/unity/audio-ovrlipsync-unity/. [Accessed: Aug. 9, 2026].

[43] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Asian Conf. Comput. Vis. (ACCV) Workshops*, 2016.

[44] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 484–492.

[45] C. Bartneck, "Godspeed questionnaire series: Translations and usage," in *International Handbook of Behavioral Health Assessment*, Springer, 2023, pp. 1–35.

[46] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Statist.*, vol. 6, no. 2, pp. 65–70, 1979.

[47] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[48] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12449–12460, 2020.

HWANG YOUN KIM is a Ph.D. student at AI-Robotics, University of Science and Technology. He received the Master of Entertainment Technology degree from Carnegie Mellon University in 2015. He has worked as a developer in various fields, including educational games, avatar control systems, and healthcare applications. His current research interests include the integration of virtual humans and deep learning models.

GHAZANFAR ALI is an Assistant Professor at the Department of AI, Gachon University. He previously worked at KIST, KAIST, and various Technology Companies. He completed his Ph.D. from UST in 2023. His research topic primarily focuses on infusing emotional Intelligence into virtual humans in AR/MR scenarios by leveraging deep learning of human behavior. He specializes in developing and supervising AI systems for digital experiences in museums, zoos, botanical gardens, medical education, counseling, and entertainment. Published and presented research at various forums, including SIGGRAPH Asia Realtime Live.

JEONGHA LEE received the B.S. degree in art and technology from Chung-Ang University, South Korea, in 2023. He is currently a PhD student with the University of Science and Technology (UST), KIST school and a Student Research Scientist with the Intelligence and Interaction Research Center, Korea Institute of Science and Technology.

JAE-IN HWANG is a professor at KIST School and a principal research scientist in Intelligence and Interaction Research Center at Korea Institute of Science and Technology. He received the B.S., M.S., and Ph.D. degrees in computer science and engineering at Pohang University of Science and Technology, Korea, in 1998, 2000, and 2007, respectively.