DEEPTalk: Dynamic Emotion Embedding for Probabilistic Speech-Driven 3D Face Animation

Jisoo Kim[1*], Jungbin Cho[1*], Joonho Park[2], Soonmin Hwang[1], Da Eun Kim[2], Geon Kim[2], Youngjae Yu[1†]

[1]Yonsei University
[2]GIANTSTEP Inc.
{jisoo6687, whwjdqls99, smsm0307, yjy}@yonsei.ac.kr
{joonho.park, daeun.kim, geon.kim}@giantstepcorp.com

Abstract

Speech-driven 3D facial animation has garnered lots of attention thanks to its broad range of applications. Despite recent advancements in achieving realistic lip motion, current methods fail to capture the nuanced emotional undertones conveyed through speech and produce monotonous facial motion. These limitations result in blunt and repetitive facial animations, reducing user engagement and hindering their applicability. To address these challenges, we introduce DEEPTalk, a novel approach that generates diverse and emotionally rich 3D facial expressions directly from speech inputs. To achieve this, we first train DEE (Dynamic Emotion Embedding), which employs probabilistic contrastive learning to forge a joint emotion embedding space for both speech and facial motion. This probabilistic framework captures the uncertainty in interpreting emotions from speech and facial motion, enabling the derivation of emotion vectors from its multifaceted space. Moreover, to generate dynamic facial motion, we design TH-VQVAE (Temporally Hierarchical VQ-VAE) as an expressive and robust motion prior overcoming limitations of VAEs and VQ-VAEs. Utilizing these strong priors, we develop DEEPTalk, a talking head generator that non-autoregressively predicts codebook indices to create dynamic facial motion, incorporating a novel emotion consistency loss. Extensive experiments on various datasets demonstrate the effectiveness of our approach in creating diverse, emotionally expressive talking faces that maintain accurate lip-sync. Our project page is available at https://whwjdqls.github.io/deeptalk.github.io/

Introduction

Speech-driven 3D facial motion generation has a wide range of applications, encompassing avatar animation for game or cinematic productions, virtual chatbots, and immersive virtual meetings within virtual reality environments. Despite substantial advancements in accurate lip synchronization with speech, as demonstrated by recent research (Richard et al. 2021; Fan et al. 2022; Xing et al. 2023; Stan, Haque, and Yumak 2023), most of these methods still produce blunt and unexpressive facial expressions. Since facial expressions are crucial for conveying nonverbal cues, this limitation significantly reduces their effectiveness in scenarios requiring realistic interactions, such as interactions with non-player characters in games or emotionally responsive virtual chatbots. Therefore, it is essential to focus on enhancing the full range of facial expressions, not just the lip movements.

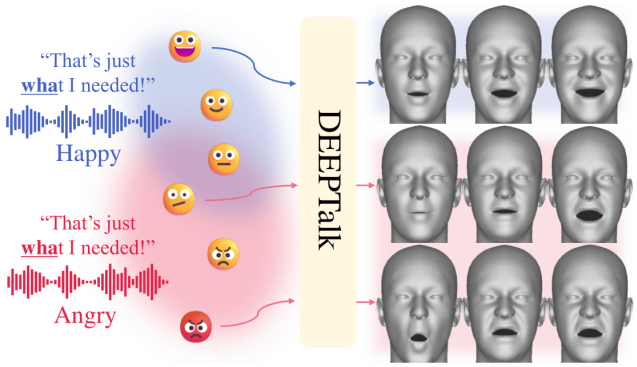

Figure 1: Overview of DEEPTalk. Starting with an emotional speech input (left), we extract probabilistic emotion embeddings (depicted as blobs), and sample from these embeddings to generate diverse emotional facial animations aligned with the input speech (right).

Previous studies on talking faces have generated expressions using either emotion labels (Daněček et al. 2023; Gan et al. 2023; Pan et al. 2023; Ji et al. 2021) or reference expressions (Ji et al. 2022; Tan et al. 2024; Ma et al. 2023a; Liang et al. 2022). Using emotion labels provides expressive but limited outcomes, while reference expressions offer more diversity at the cost of needing numerous expressive references. Moreover, both approaches fail to capture vocal nuances, often leading to misalignment between facial expressions and speech. As shown in Figure 1, the phrase "That's just what I needed!" can carry different meanings based on the emotion it conveys (e.g., happiness or anger), emphasizing the importance of facial expressions that reflect the intended sentiment. Without this alignment, expressions may appear unnatural from spoken words, contributing to uncanny valley effect (Wang, Lilienfeld, and Rochat 2015).

*These authors contributed equally.
†Corresponding author

Therefore, the most effective approach to generating non-verbal facial expressions from speech is to leverage the rich information embedded within the speech, which simultaneously conveys the speaker's intentions and emotions. Central to this process is prosody, encompassing non-linguistic elements such as pitch, speed, volume, and tone variations. Prosody is crucial due to its intricate link with facial expressions as shown in (Cvejic, Kim, and Davis 2010)(Pell 2005).

Utilizing this groundwork, we propose a talking head model, DEEPTalk, that leverages both emotion and motion priors to generate diverse emotional 3D facial expressions directly from speech inputs. We first utilize cross-modal contrastive learning to capture the emotional correlation between speech and facial expressions. Unlike (Albanie et al. 2018), which used this correlation primarily for emotion recognition from unlabeled data, we aim to use it to develop a joint embedding space, which we call DEE(Dynamic Emotion Embedding). Given the inherent ambiguity in both speech and facial expressions—where multiple expressions can correspond to a single piece of speech—we utilize probabilistic embeddings (Chun et al. 2021; Chun 2023). These embeddings are designed to model the uncertainty associated with the inputs and facilitate sampling from the probability distribution. This allows the generation of diverse emotional faces from the same speech input as illustrated by the red lines in Figure 1. After constructing a joint embedding space for facial motion and speech, we use it as a strong emotion prior to train an emotional talking head model.

We then aim to build an expressive motion prior that is robust to perceptual losses. Recent studies have demonstrated that by learning motion priors from discrete codebooks, it is possible to generate a diverse range of facial and body motions (Li et al. 2021; Yi et al. 2023; Xing et al. 2023; Ng et al. 2022). However, due to the dynamic nature of emotional talking faces, VQ-VAE(Van Den Oord, Vinyals et al. 2017) alone struggles to capture the entire motion space fully. Recognizing that facial motion encompasses different temporal hierarchies—for instance, the mouth region exhibits high frequencies while the upper face displays lower temporal frequencies—we train a hierarchical discrete motion prior to effectively address these variations effectively.

Building on the aforementioned robust emotion and motion priors, DEEPTalk is specifically engineered to non-autoregressively map emotional speech to our codebook indices. To ensure that generated expressions consistently reflect the input speech emotions, we introduce a novel emotion consistency loss. Training incoporates Gumbel-Softmax (Jang, Gu, and Poole 2016) and differentiable rendering to ensure end-to-end differentiability. Our qualitative and quantitative assessments, including extensive user studies, demonstrate that DEEPTalk excels at generating diverse emotional facial motions while also outperforming lip synchronization.

In summary, our main contributions are as follows: we design a Dynamic Emotion Embedding (DEE) that jointly learns from speech and facial motion sequences through probabilistic contrastive learning. Additionally, we propose a novel temporally hierarchical VQ-VAE (TH-VQVAE) to construct a motion prior that is both expressive and robust. Finally, we train a talking head model, DEEPTalk, leveraging these strong priors to generate diverse and expressive emotional facial motions while also outperforming state-of-the-art models on lip synchronization.

Related Work

Speech-Driven 3D Face Animation. The availability of large 4D mesh datasets synchronized with audio (Richard et al. 2021; Fanelli et al. 2010) has greatly advanced deep learning-based 3D face animation, enabling robust lip synchronization. Extending these advancements, VOCA (Cudeiro et al. 2019) improves realism by generating animations from any speech inputs, employing time convolutions with a speaker identity one-hot vector. FaceFormer (Fan et al. 2022) uses a transformer-based model to generate facial movements auto-regressively. However, despite accurate lip synchronization, the generated upper face remains static as it is less correlated with input speech. MeshTalk (Richard et al. 2021) addresses this by disentangling audio-correlated and uncorrelated movements using a categorical latent space to model upper face dynamics. Similarly, CodeTalker (Xing et al. 2023) utilizes a discrete motion prior with VQ-VAE to reconstruct dynamic facial motions across the entire face. However, these methods generate deterministic motion, overlooking the inherently non-deterministic relationship between facial motion and speech. Therefore, FaceDiffuser (Stan, Haque, and Yumak 2023) addresses the limitation of deterministic models by employing a diffusion model, which is probabilistic in nature, to generate multiple facial motions for a given speech. However, it still falls short in capturing the diverse emotional expressions conveyed through speech, which is crucial for enhancing interactiveness for talking face models.

Emotional 3D Face Animation. Recent studies have underscored the critical role of emotion in creating realistic and expressive 3D facial motions by integrating additional emotional information. Specifically, EMOTE (Daněček et al. 2023) employs one-hot labels to control emotion, producing emotional facial motions through an emotion-content disentanglement loss. Chen et al. (Chen, Zhao, and Zhang 2023) utilized the logits from an emotion classifier applied to reference images as an emotion prior during training, generating emotional facial motion through an emotion-augmented network. However, these methods require explicit control, lacking a direct connection to the emotion conveyed in the actual speech. The work most related to ours is EmoTalk (Peng et al. 2023), which aims to generate emotional blendshapes from audio input only, utilizing an emotion-content disentangling method. However, the training of EmoTalk's emotion-content disentangling encoder is constrained to specific datasets like RAVDESS (Livingstone and Russo 2018), containing the same sentences expressed in multiple emotions. This constraint limits the use of larger, more diverse datasets such as MEAD (Wang et al. 2020), ultimately hindering the model's ability to generalize emotional expressions on other datasets. Unlike previous methods that either disentangle content and emotion from speech or rely on explicit conditioning, our approach establishes a

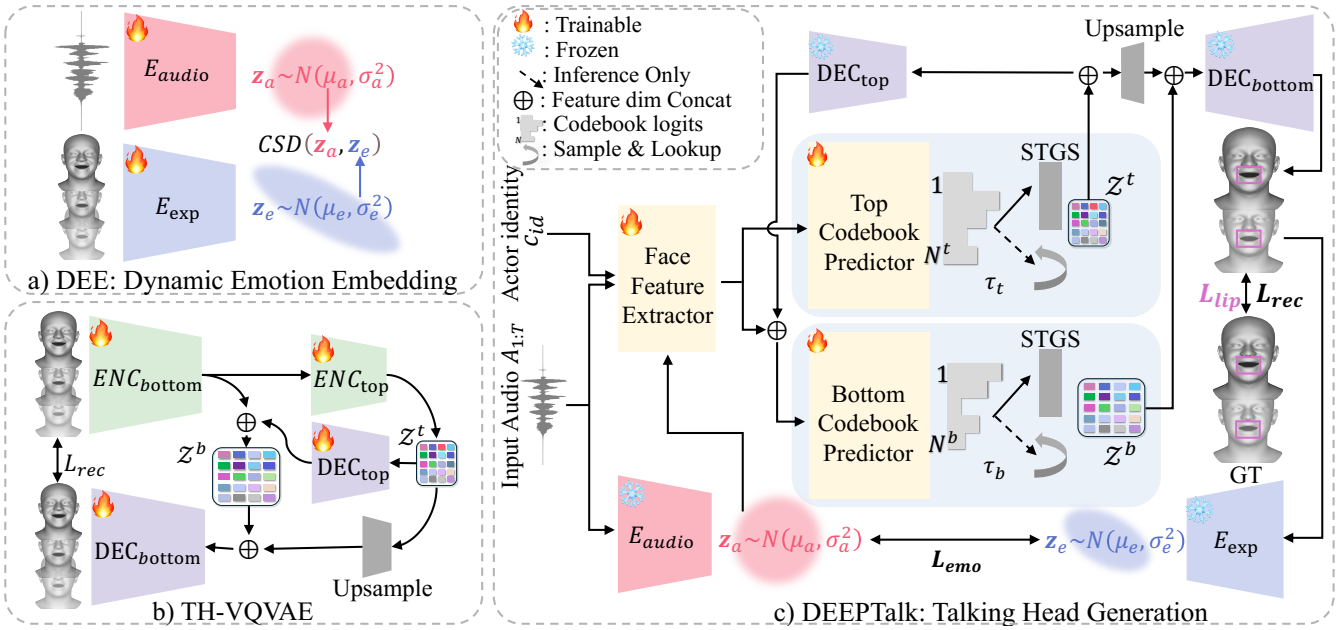

Figure 2: Overall Architecture of Our Method. (a) E_{audio} and E_{exp} are trained to predict mean and variance for a joint audio-facial emotion embedding space, DEE. (b) We train TH-VQVAE with separate codebooks, $\mathcal{Z}^b$ and $\mathcal{Z}^t$, for low and high-frequency motions, respectively. (c) DEEPTalk first extracts face features, predict top and bottom codebook indices, and use frozen TH-VQVAE decoders to decode the quantized motion features. To ensure emotion alignment between input audio and the predicted facial expressions, we introduce an emotional consistency loss L_{emo} by utilizing DEE.

robust emotion prior by leveraging the natural correlation between speech and facial expressions, enabling seamless integration into the talking face training pipeline.

Method

Overview Our goal is to synthesize emotional facial motion solely from speech. However, the complex many-to-many relationship between speech and facial expressions poses challenges for generating expressive emotional faces. To overcome this, we build (1) DEE, a joint probabilistic emotional space to encode both audio and facial motion. DEE sample emotion embeddings from audio, and enable generated facial motions aligned to the corresponding embedding. Also for robust and expressive motion prior, we design (2) TH-VQVAE to learn a discrete motion space capable of capturing both high and low-frequency motions. Building on these priors, we train (3) DEEPTalk to map emotional audio to motion codebooks, resulting in facial animations that are both emotionally expressive and accurately lip-synced. Additional details on each component are provided in the Supplementary.

Formulation. Our task can be formulated as follows: Let $\boldsymbol{A}_{1:T} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_T)$ be a sequence of input speech snippets where each $\boldsymbol{a}_t \in \mathbb{R}^D$ has D sampled audio, and let $\boldsymbol{F}_{1:T} = (\boldsymbol{f}_1, \cdots, \boldsymbol{f}_T)$, where $\boldsymbol{f}_t \in \mathbb{R}^{d_{exp}+3}$ is a sequence of FLAME expression parameters $\boldsymbol{\Psi}_t \in \mathbb{R}^{d_{exp}}$ concatenated with jaw parameters $\boldsymbol{\theta}_t^{\text{jaw}} \in \mathbb{R}^3$.

$$\boldsymbol{f}_t = [\boldsymbol{\Psi}_t, \boldsymbol{\theta}_t^{\text{jaw}}], \tag{1}$$

Our goal is to analyze the content and emotion of $\boldsymbol{A}_{1:T}$ and, together with the one-hot speaker identity $\boldsymbol{c}_{id}$, predict FLAME expression parameters $\hat{\boldsymbol{F}}_{1:T} = (\hat{\boldsymbol{f}}_1, \cdots, \hat{\boldsymbol{f}}_T)$ that align with the input speech. As our talking head model is non-auto-regressive, denoting θ as model parameters, our end-to-end procedure can be written as

$$\hat{\boldsymbol{F}}_{1:T} = \text{DEEPTalk}_\theta(\boldsymbol{A}_{1:T}, \boldsymbol{c}_{id}), \tag{2}$$

DEE: Dynamic Emotional Embedding

Emotion feature extraction. We utilized the recently proposed emotion2vec (Ma et al. 2023b) to extract emotion features in our audio encoder. Denoting feature extractor as F_{audio}, this process is defined as :

$$F_{audio}(\boldsymbol{A}_{1:T}) \rightarrow \epsilon_{audio} \tag{3}$$

For the expression feature extractor in the expression encoder, we first trained an emotion classification model using the AffectNet dataset (Mollahosseini, Hasani, and Mahoor 2017). As Affectnet is an image dataset, we applied a 3D flame parameter reconstruction method (Daněček, Black, and Bolkart 2022) to generate 3D pseudo ground truth for each image in AffectNet and trained an emotion recognition model. The trained encoder was then repurposed as the facial expression feature extractor F_{exp}. This process is defined as:

$$F_{exp}(\boldsymbol{F}_{1:T}) \rightarrow \epsilon_{exp} \tag{4}$$

Emotional space construction. With both audio and expression feature extractors in place, DEE trains the audio and expression encoders separately. The overview of DEE

is illustrated in Figure 2(a). Each encoder E comprises two distinct heads, followed by a Generalized Pooling Operator (GPO) (Chen et al. 2021), with one head dedicated to μ and the other to $\log \sigma^2$, similar to the approach in PCME++(Chun 2023). The final probabilistic emotion embeddings for audio and expression can be written as follows:

$$E_{audio}(\epsilon_{audio}) \rightarrow Z_a \sim N(\mu_a, \sigma_a^2) \tag{5}$$

$$E_{exp}(\epsilon_{exp}) \rightarrow Z_e \sim N(\mu_e, \sigma_e^2) \tag{6}$$

Objectives. We train DEE on the Closed-Form Sampled distance (CSD), between audio and expression probabilistic embeddings Z_a, Z_e:

$$CSD(Z_a, Z_e) = \|\mu_a - \mu_e\|_2^2 + \|\sigma_a^2 + \sigma_e^2\|_1 \tag{7}$$

TH-VQVAE: Temporally Hierarchical VQ-VAE

We addressed the challenge of modeling high-frequency lip movements and slower facial motions by extending VQ-VAE2(Razavi, van den Oord, and Vinyals 2019) into the temporal motion domain, creating TH-VQVAE. This model uses distinct codebooks for different motion frequencies, enhancing reconstruction quality and the capabilities of the Talking Head Generator. It enables fine-grained lip movements and dynamic facial expressions while also offering controllability at each hierarchical level.

Model Architecture. TH-VQVAE consists of a bottom encoder ENC_{bottom}, a top encoder ENC_{top}, a bottom decoder DEC_{bottom}, a top decoder DEC_{top}, and two facial motion codebooks $\mathcal{Z}^b$ and $\mathcal{Z}^t$. Each codebook can be formulated as

$$\mathcal{Z}^b = \left\{ \mathbf{z}_k^b \in \mathbb{R}^{C^b} \right\}_{k=1}^{N^b}, \mathcal{Z}^t = \left\{ \mathbf{z}_k^t \in \mathbb{R}^{C^t} \right\}_{k=1}^{N^t} \tag{8}$$

where each represents fine and coarse facial motion. Any sequence of facial motion $\boldsymbol{F}_{1:T}$ can be represented by one item on each codebook $\boldsymbol{z}_i^b, \boldsymbol{z}_j^t$ and decoded through DEC_{bottom} into the corresponding facial motion. As depicted in Figure 2(b), the input motion segment $x = \mathbf{F}_{1:T} \in \mathbb{R}^{T \times (d_{exp}+3)}$ is first encoded to a bottom motion feature and then encoded to top motion feature.

$$\hat{z}^b = \text{ENC}_{bottom}(x) \in \mathbb{R}^{\tau^b \times (C^b - C^t)}, \tag{9}$$

where $\tau^b = \frac{T}{q^b}$ and $\tau^t = \frac{\tau}{q^t}$ is the length of the sequence divided by bottom quant factor q^b and length of the sequence of the bottom motion feature divided by top quant factor q^t. Using the top motion feature, we obtain a top quantized sequence z_q^t, then use this to get decoded top features z_d^t as

$$z_q^t = \arg\min_{z_t^t \in \mathcal{Z}^t} \left\| \hat{z}^t - z_t^t \right\| \in \mathbb{R}^{\tau^t \times C^t}. \tag{10}$$

$$z_d^t = \text{DEC}_{top}(z_q^t) \in \mathbb{R}^{\tau^b \times C^t}, \tag{11}$$

and stack $\hat{z}^b$ and z_d^t along the feature dimension and obtain z_q^b by

$$z_q^b = \arg\min_{z_t^b \in \mathcal{Z}^b} \left\| [\hat{z}^b, z_d^t] - z_t^b \right\| \in \mathbb{R}^{\tau^t \times C^b}. \tag{12}$$

Finally, we upsample z_q^t from τ^t to match the temporal dimension τ^b and decode the stacked upsampled top quantized feature and bottom quantized feature to reconstruct the original facial motion inputs.

$$\hat{x} = \text{DEC}_{bottom}\left([\text{Upsample}(z_q^t), z_q^b] \right) \tag{13}$$

where $\hat{x}$ is a reconstruction of the input motion sequence x. We use the standard VQ-VAE loss to train TH-VQVAE. Details are included in the Supplementary.

DEEPTalk: Talking Head Generator

Shown in Figure 2(c), the Talking head generator is composed of two distinct modules. 1) Face Feature Extractor that extracts facial features from audio and emotion inputs, and 2) Codebook Predictor that predicts codebook indices given facial features.

Face Feature Extractor. In order to generate emotional face features aligned with input speech, we employ Wav2Vec 2.0 (Baevski et al. 2020) to encode content and DEE's audio encoder E_{audio} to encode emotion. During inference, we can sample from the output distribution of E_{audio} and control its uncertainty through scaling the $\log \sigma^2$ of E_{audio} output with the uncertainty control factor α. Both features are concatenated and fed into a transformer model to generate emotional face features.

Codebook Predictor. Instead of predicting each codebook's index at once, we utilize two separate models: the top codebook predictor and the bottom codebook predictor. First, we obtain low-frequency motion features and use them as a condition to predict high-frequency motion features. Specifically, the top codebook predictor first predicts the logit distribution of the top codebook and indexes the top quantized features, which contain low-frequency motion information. It is then decoded by DEC_{top} and concatenated with face features. The bottom codebook predictor then predicts the bottom codebook's logit distribution and indexes the bottom quantized features. Both quantized feature sequences are concatenated and fed to the pretrained and DEC_{bottom} to produce FLAME expression $\hat{\boldsymbol{\Psi}}$ and jaw pose $\hat{\boldsymbol{\theta}}^{\text{jaw}}$ parameter sequence. During training, we utilize the straight-through Gumbel-softmax (STGS) to make indexing of codebooks differentiable, and during inference, we sample from the logit distribution by controlling each codebook's temperature τ_b and τ_t.

Objectives. We employed three distinct loss functions to optimize performance : (i) reconstruction loss, (ii) lip loss, and (iii) emotion consistency loss.

$$L_{total} = \lambda_1 L_{rec} + \lambda_2 L_{emo} + \lambda_3 L_{lip} \tag{14}$$

Reconstruction Loss. We computed the Mean Squared Error between the pseudo-GT and predicted vertices, denoted as L_{rec}.

Emotion Consistency Loss. To ensure that the generated face motion reflects the same emotion as the input audio, we propose an emotion consistency loss. By leveraging the

Method	CREMA-D				RAVDESS			HDTF	MEAD			
	FID↓	FFD↓	Emo-FID↓	LSE-D↓	FID↓	FFD↓	Emo-FID↓	LSE-D↓	FID↓	FFD↓	Emo-FID↓	LSE-D↓
FaceFormer	17.05	62.88	27.69	8.659	27.43	50.40	40.45	11.88	31.40	78.62	35.67	10.70
EmoTalk*	64.75	-	149.1	9.132	-	-	-	-	88.33	-	228.9	10.77
FaceDiffuser	14.78	79.45	23.24	8.686	17.06	81.74	31.23	12.48	30.93	119.2	68.30	10.93
EMOTE	22.24	85.13	33.72	8.612	20.22	62.05	27.40	12.38	27.22	93.17	16.83	10.73
Ours	**11.58**	**50.00**	**11.99**	**8.535**	**11.94**	**32.82**	**11.44**	**11.83**	**26.07**	**64.02**	**15.16**	**10.65**

Table 1: Quantitative Evaluation Results. Best performance in bold, and the second best underlined. *EmoTalk does not predict FLAME parameters preventing evaluation of FFD and was trained on RAVDESS and HDTF. LSE-C reported in Supplementary.

trained DEE, we predict the mean from audio input μ_a and generated motion μ_e and enforce a high cosine similarity between these pairs to ensure they represent the same emotion.

$$L_{emo} = \frac{\mu_a \cdot \mu_e}{\|\mu_a\|\|\mu_e\|} \tag{15}$$

Lip Loss. To provide additional lip supervision, we use lip reading perceptual loss L_{lip} (Daněček et al. 2023).

Experiments

Datasets. We utilize MEAD for train and test, and incorporate CREMA-D (Cao et al. 2014), RAVDESS, and HDTF (Zhang et al. 2021) for evaluation. MEAD, CREMA-D, and RAVDESS are lab-recorded emotional talking face videos, while HDTF comprises YouTube-sourced talking face videos. Due to RAVDESS's limited utterances and HDTF's lack of emotion, we evaluate only emotion and lip sync, respectively. Additionally, MEAD's overlapping utterance between train and test datasets make unsuitable for fair comparisons with models untrained on it, and given no clear benchmark for lip generalization on emotional in-the-wild speeches, we constructed an audio test set, Emo-Vox, derived from VoxCeleb2. To create a reliable pseudo-3D ground truth dataset, we employ the SOTA 3D face reconstruction method (Daněček, Black, and Bolkart 2022) exclusively on lab setting datasets for accurate FLAME parameter extraction. Further details are in the Supplementary.

Baseline Impelmentations. To compare DEEPTalk with facial parameter-based models, we train FaceFormer, FaceDiffuser, and EMOTE on MEAD dataset. As EmoTalk employs a unique training approach specifically tailored for RAVDESS and utilizes an in-house blend shape reconstruction method, we utilize its pre-trained weights. Additionally, we conduct experiments with vertex-based models (FaceFormer, FaceDiffuser, MeshTalk, CodeTalker) trained on the ground truth face mesh from the VOCASET dataset.

Quantitative Evaluation

Evaluation Metrics. We adopt **FID** to evaluate the realism of rendered faces. To further assess the realism of facial movements in the FLAME parameter space, we developed **FFD (Frechet Face Distance)**, inspired by FGD (Yoon et al. 2020), and computed it on sequences of FLAME parameters using an encoder trained on MEAD and BEATv2 (Liu et al. 2024). To evaluate emotional expressiveness, we compute **Emo-FID**, an adaptation of FID that uses an emotion feature extractor from AffectNet (Mollahosseini, Hasani, and Mahoor 2017), replacing the inception network to focus on emotion properties. For lip-sync evaluation, we used SyncNet metrics (Chung and Zisserman 2016) **LSE-D** (Lip Sync Error Distance) and **LSE-C** (Lip Sync Error Confidence) following (Aneja et al. 2024). Unlike Lip Vertex Error (LVE), which is unsuitable for DEEPTalk due to its diverse emotional faces, these metrics evaluate lip-sync without requiring ground truth lip movements. Finally, we measure **Diversity** (Ren et al. 2023).

Evaluation on Realism and Emotion. For realism comparison, DEEPTalk outperforms all other methods in FID and FFD across all datasets, shown in Table 1, indicating superior natural expressions and facial movements. DEEPTalk significantly surpasses other methods in Emo-FID, demonstrating superior emotional expressiveness that closely resembles real human expressions. It is noteworthy that, despite utilizing ground truth emotion labels to generate expressions, EMOTE still lags behind our approach due to its lack of diversity and tendency for exaggerated expressions.

Method	Train Dataset	LSE-D↓	LSE-C↑
FaceFormer	VOCASET	11.55	0.763
MeshTalk	VOCASET	11.54	0.555
CodeTalker	VOCASET	11.51	0.782
FaceDiffuser	VOCASET	11.82	0.707
EmoTalk	RAV/HDTF	11.40	0.658
FaceDiffuser	MEAD	11.61	0.569
FaceFormer	MEAD	11.31	**0.893**
EMOTE	MEAD	11.40	0.879
Ours	MEAD	**11.28**	0.889

Table 2: Lip sync results on Emo-Vox.

Evaluation on Lip sync. In Table 1, DEEPTalk achieves the highest performance on LSE-D and ranks first or second on LSE-C across the MEAD test set and other datasets, demonstrating strong generalization and precise lip sync. We also conducted evaluations on Emo-Vox to assess lip sync, while ensuring a fair comparison with both parameter-based and vertex-based models. As shown in Table 2, our model excels in LSE-D and ranks second in LSE-C, demonstrating effective generalization of lip movements to emotional, in-the-wild audio. This performance further indicates our bottom codebook's capability to capture high-frequency details, enhancing dynamic lip movements. Moreover, due to the limited scale of the ground truth scan dataset, vertex-

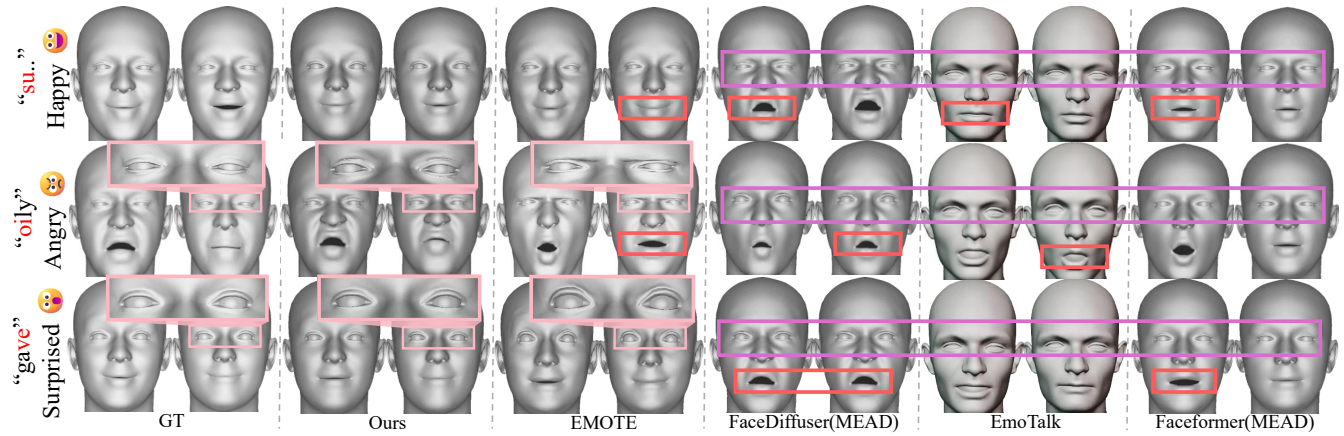

Figure 3: Qualitative results on MEAD test set. Each row displays the predicted facial motions for each utterance and corresponding emotion (left) generated by baseline models. Lip motion deviations from the ground truth are highlighted in red, while incorrect or neutral emotional expressions are indicated in purple. EMOTE, being conditioned on emotion labels, exhibits a high degree of emotional expressiveness. However, this conditioning sometimes results in exaggerated expressions, highlighted in pink in the enlarged images. In contrast, DEEPTalk generates natural emotional faces while maintaining accurate lip sync.

Method	α	τ	Diversity↑	LSE-D↓
FaceDiffuser	-	-	21.56	10.93
Ours	1	argmax	19.23	10.65
Ours	0.1	argmax	25.48	10.60
Ours	-2	argmax	25.50	10.60
Ours	-4	argmax	**25.55**	10.60
Ours	mean	$\tau_t = 1, \tau_b = 0.1$	9.91	10.65
Ours	mean	$\tau_t = 0.1, \tau_b = 1$	13.03	10.65
Ours	mean	$\tau_t = 4.5, \tau_b = 1$	**23.22**	10.68

Table 3: Diversity results on the MEAD test set. DEEPTalk generates diverse faces while maintaining accurate lip sync. LSE-C reported in the Supplementary.

based models fail to achieve accurate lip sync compared to those using pseudo ground truth. Notably, FaceFormer, trained on MEAD, produces nearly neutral expressions (see Figure 3), whereas ours are more expressive, benefiting FaceFormer's lip-sync performance due to the emotion and lip-sync trade off, as demonstrated in ablation studies.

Evaluation on Diversity. DEEPTalk provides control over the diversity of generated facial motions in two ways: 1) **diverse emotional facial motions from the same speech input** by adjusting the control factor α, and 2) **diverse facial motions from the same speech and emotion** by modifying the temperature of the bottom (τ_b) and top (τ_t) codebook predictors. We evaluated each controllability factor by varying them independently while keeping the other factor deterministic. We continued this process until our method surpassed FaceDiffuser in terms of diversity, after which we assessed lip synchronization. As shown in Table 3, DEEPTalk demonstrates superior lip synchronization, even with increased diversity, compared to FaceDiffuser on both control factors, illustrating the model's effectiveness in generating diverse yet accurate facial motion. Notably, these two methods of control are orthogonal and can be used independently, potentially leading to even greater diversity.

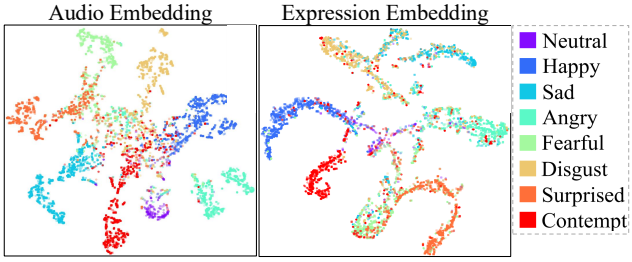

Figure 4: Embeddings are clustered by emotion categories.

Qualitative Evaluation

Visual Comparison. Figure 3 compares our method with SOTA methods on the MEAD test set. While most methods generate natural lip movements, they often misalign with the ground truth, such as opening or closing lips incorrectly. Additionally, except for EMOTE, other methods produce incorrect or unexpressive expressions. This highlights the effect of DEE and L_{emo} on our framework. While EMOTE generates emotional expressions using emotional labels, it often produces exaggerated faces, like unnaturally wide eyes or flat eyebrows, deviating from the ground truth. In contrast, DEEPTalk generates natural emotional expressions that closely resemble the ground truth without emotion labels by leveraging emotions inherent within the speech.

Effect of Emotion Embedding. Figure 4 shows that our emotion embedding clusters by emotions using T-SNE, capturing a meaningful emotion space. Furthermore, to evaluate its efficacy in generating emotional expressions, we randomly selected two audio samples with distinct emotions and interchanged their emotion embeddings from DEE's au-

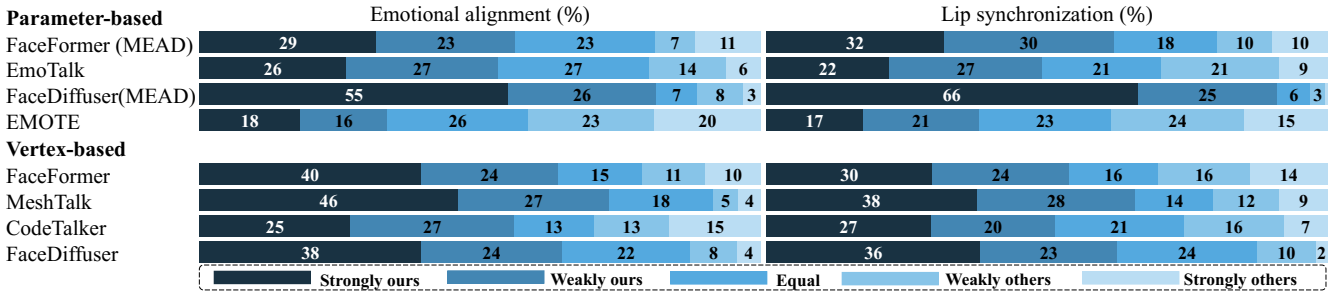

Figure 5: User Study Results. Our method is preferred over most methods on emotional alignment and lip synchronization.

dio encoder, before inputting them into the Face Feature Extractor. As shown in Figure 6, this swap effectively modifies the emotional expression to match the reference speech, while maintaining precise lip synchronization with the original speech. Further analysis is provided in Supplementary.

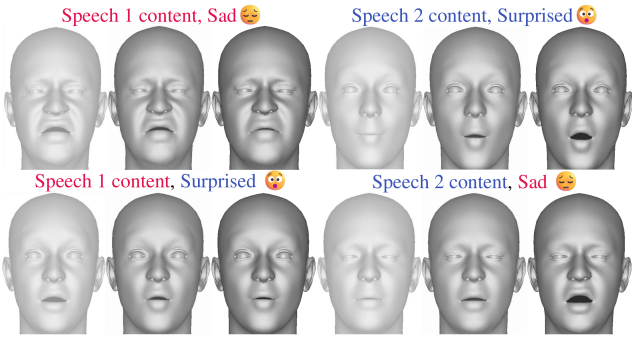

Figure 6: Swapping emotion embedding alters the emotional expression while maintaining precise lip synchronization

User Studies. We conducted A/B test with 5-point Likert scale for 98 users to evaluate model preference across two subtasks: 1) **emotional alignment** between speech and expression 2) **lip synchronization**. We sampled twenty audio clips from MEAD and ten from Emo-Vox. Details are provided in Supplementary. As shown in Figure 5, DEEPTalk outperformed all parameter-based models across all tasks except for EMOTE, which leverages ground truth emotion labels to generate faces, serving as the upper bound for emotional face generation. Note that EMOTE is deterministic and lacks expression diversity. For vertex-based models, DEEPTalk excelled in both tasks for all competitors. Due to the scarcity and limited size of emotional ground truth face scan datasets, vertex-based methods struggle with emotional expression and accurate lip synchronization. This highlights that using pseudo-3D data, as done in DEEPTalk, is a promising approach for achieving accurate emotional 3D talking face animation with precise lip synchronization.

Ablation Studies

We conducted ablation studies on the MEAD test set and Emo-Vox to analyze the contributions of individual components of DEEPTalk. Specifically, we compared (1) DEEPTalk, (2) DEEPTalk with VAE, (3) DEEPTalk with VQ-VAE, (4) DEEPTalk without L_{lip} and (5) DEEPTalk without L_{emo}. As shown in Table 4, utilizing a discrete motion prior is crucial, as its absence results in significant quality degradation. This is due to perceptual losses like L_{emo} and L_{lip} where enforcing such constraints leads to unnatural expressions. Furthermore, employing TH-VQVAE enhances performance across all metrics by effectively capturing both low and high-frequency motion patterns. The incorporation of our proposed L_{emo} significantly improves emotional expressiveness and realism, as evidenced by Emo-FID, FID, and FFD scores. Interestingly, this enhancement results in a minor decrease in lip synchronization, indicating a tradeoff between emotional expressiveness and lip sync precision.

	Emo-Vox	MEAD			
Methods	LSE-D↓	LSE-D↓	Emo-FID↓	FID↓	FFD↓
w/ VAE	10.92	10.78	102.2	43.40	325.1
w/ VQVAE	10.87	10.70	<u>16.58</u>	30.10	66.98
w/o L_{lip}	10.98	10.91	19.94	<u>28.23</u>	**63.51**
w/o L_{emo}	**10.67**	**10.66**	28.78	32.47	91.35
Full (Ours)	<u>10.74</u>	<u>10.65</u>	**15.15**	**26.07**	<u>64.02</u>

Table 4: Ablation results on Emo-Vox and MEAD test set. LSE-C reported in the Supplementary.

Conclusion

This paper presents DEEPTalk, a novel speech-driven talking head framework designed to generate diverse and emotionally expressive facial animations. Unlike previous methods, DEEPTalk achieves precise lip synchronization while ensuring facial expressions accurately reflecting the emotional tone of the input speech. This is accomplished through our dynamic emotion embedding (DEE), which serves as a strong emotion prior, and a temporally hierarchical VQ-VAE (TH-VQVAE), which functions as a robust and dynamic motion prior. Additionally, the probabilistic design of DEEPTalk facilitates non-deterministic generation that can be controlled on various levels. Extensive experiments on various datasets demonstrate the superiority of our model over existing methods across six metrics, with notable improvements in emotional realism.

Acknowledgments

This work was supported by an IITP grant funded by the Korean Government (MSIT) (No. RS-2020-II201361 , Artificial Intelligence Graduate School Program (Yonsei University)) and Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name:Development of multimodal UX evaluation platform technology for XR spatial responsive content optimization, Project Number: RS-2024-00361757) and GIANTSTEP Inc.

References

Albanie, S.; Nagrani, A.; Vedaldi, A.; and Zisserman, A. 2018. Emotion Recognition in Speech using Cross-Modal Transfer in the Wild. *CoRR*, abs/1808.05561.

Aneja, S.; Thies, J.; Dai, A.; and Nießner, M. 2024. FaceTalk: Audio-Driven Motion Diffusion for Neural Parametric Head Models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Cao, H.; Cooper, D. G.; Keutmann, M. K.; Gur, R. C.; Nenkova, A.; and Verma, R. 2014. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4): 377–390.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the Best Pooling Strategy for Visual Semantic Embedding. arXiv:2011.04305.

Chen, Y.; Zhao, J.; and Zhang, W.-Q. 2023. Expressive Speech-driven Facial Animation with controllable emotions. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 387–392. IEEE.

Chun, S. 2023. Improved probabilistic image-text representations. *arXiv preprint arXiv:2305.18171*.

Chun, S.; Oh, S. J.; de Rezende, R. S.; Kalantidis, Y.; and Larlus, D. 2021. Probabilistic Embeddings for Cross-Modal Retrieval. arXiv:2101.05068.

Chung, J. S.; and Zisserman, A. 2016. Out of Time: Automated Lip Sync in the Wild. In *ACCV Workshops*.

Cudeiro, D.; Bolkart, T.; Laidlaw, C.; Ranjan, A.; and Black, M. J. 2019. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10101–10111.

Cvejic, E.; Kim, J.; and Davis, C. 2010. Prosody off the top of the head: Prosodic contrasts can be discriminated by head motion. *Speech Commun.*, 52(6): 555–564.

Daněček, R.; Black, M. J.; and Bolkart, T. 2022. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20311–20322.

Daněček, R.; Chhatre, K.; Tripathi, S.; Wen, Y.; Black, M.; and Bolkart, T. 2023. Emotional speech-driven animation with content-emotion disentanglement. In *SIGGRAPH Asia 2023 Conference Papers*, 1–13.

Fan, Y.; Lin, Z.; Saito, J.; Wang, W.; and Komura, T. 2022. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18780.

Fanelli, G.; Gall, J.; Romsdorfer, H.; Weise, T.; and Van Gool, L. 2010. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6): 591–598.

Gan, Y.; Yang, Z.; Yue, X.; Sun, L.; and Yang, Y. 2023. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22634–22645.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Ji, X.; Zhou, H.; Wang, K.; Wu, Q.; Wu, W.; Xu, F.; and Cao, X. 2022. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. arXiv:2205.15278.

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

Li, J.; Kang, D.; Pei, W.; Zhe, X.; Zhang, Y.; He, Z.; and Bao, L. 2021. Audio2Gestures: Generating Diverse Gestures from Speech Audio with Conditional Variational Autoencoders. arXiv:2108.06720.

Liang, B.; Pan, Y.; Guo, Z.; Zhou, H.; Hong, Z.; Han, X.; Han, J.; Liu, J.; Ding, E.; and Wang, J. 2022. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3387–3396.

Liu, H.; Zhu, Z.; Becherini, G.; Peng, Y.; Su, M.; Zhou, Y.; Zhe, X.; Iwamoto, N.; Zheng, B.; and Black, M. J. 2024. EMAGE: Towards Unified Holistic Co-Speech Gesture Generation via Expressive Masked Audio Gesture Modeling. arXiv:2401.00374.

Livingstone, S. R.; and Russo, F. A. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Zenodo.

Ma, Y.; Wang, S.; Hu, Z.; Fan, C.; Lv, T.; Ding, Y.; Deng, Z.; and Yu, X. 2023a. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1896–1904.

Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; and Chen, X. 2023b. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. *arXiv preprint arXiv:2312.15185*.

Mollahosseini, A.; Hasani, B.; and Mahoor, M. H. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1): 18–31.

Ng, E.; Joo, H.; Hu, L.; Li, H.; Darrell, T.; Kanazawa, A.; and Ginosar, S. 2022. Learning to Listen: Modeling Non-Deterministic Dyadic Facial Motion. arXiv:2204.08451.

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

Pell, M. 2005. Prosody–face Interactions in Emotional Processing as Revealed by the Facial Affect Decision Task. *Journal of Nonverbal Behavior*, 29: 193–215.

Peng, Z.; Wu, H.; Song, Z.; Xu, H.; Zhu, X.; He, J.; Liu, H.; and Fan, Z. 2023. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20687–20697.

Razavi, A.; van den Oord, A.; and Vinyals, O. 2019. Generating Diverse High-Fidelity Images with VQ-VAE-2. arXiv:1906.00446.

Ren, Z.; Pan, Z.; Zhou, X.; and Kang, L. 2023. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

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

Stan, S.; Haque, K. I.; and Yumak, Z. 2023. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion.

Tan, S.; Ji, B.; Ding, Y.; and Pan, Y. 2024. Say anything with any style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5088–5096.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, K.; Wu, Q.; Song, L.; Yang, Z.; Wu, W.; Qian, C.; He, R.; Qiao, Y.; and Loy, C. C. 2020. MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation. In *ECCV*.

Wang, S.; Lilienfeld, S. O.; and Rochat, P. 2015. The Uncanny Valley: Existence and Explanations. *Review of General Psychology*, 19(4): 393–407.

Xing, J.; Xia, M.; Zhang, Y.; Cun, X.; Wang, J.; and Wong, T.-T. 2023. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12780–12790.

Yi, H.; Liang, H.; Liu, Y.; Cao, Q.; Wen, Y.; Bolkart, T.; Tao, D.; and Black, M. J. 2023. Generating Holistic 3D Human Motion from Speech. arXiv:2212.04420.

Yoon, Y.; Cha, B.; Lee, J.-H.; Jang, M.; Lee, J.; Kim, J.; and Lee, G. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6): 1–16.

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