ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446842

SPEECH-DRIVEN EMOTIONAL 3D TALKING FACE ANIMATION USING EMOTIONAL EMBEDDINGS

*Seongmin Lee, Jeonghaeng Lee, Hyewon Song, and Sanghoon Lee**

Yonsei University

ABSTRACT

Existing emotional talking 3D facial animation primarily focus on animating emotional faces using a specific emotion condition. However, in real-world situations, no one consistently speaks with just one emotion. Thus, previous emotion-based approaches have very limited applicability in real-world applications. To address this issue, we propose SDETalk, a novel learning framework that animates the emotional talking faces by leveraging the emotional source from a speech. Unlike previous studies, which use static one-hot emotion conditions, the proposed network regresses complex emotional states from speech. It enables the network to animate natural facial animation from an emotional speech without using a specific emotional condition. Furthermore, we design the proposed method to produce head motions because head motion is an important factor to enhance the naturalness of talking face animation. By doing this, our approach simultaneously achieves accurate lip motion, natural expressions, and rhythmical head motions from emotional speech. Through extensive experiments in both qualitative and quantitative manners, it is demonstrated that our method outperforms other state-of-the-art methods by animating realistic and expressive 3D faces.

Index Terms— emotional talking face, emotional embedding, complex emotions, head motion

1. INTRODUCTION

Human-like 3D avatars that can portray a character and act like real people have been widely used in various entertaining applications such as dance generation [1], face animation [2, 3], and face reconstruction [4]. Especially, as the demand for communication in the virtual space increases [2], speech-driven talking avatars have become an attractive issue. Using speech-driven talking avatars, many people can make their content more efficient and effective. However, most of the previous speech-driven avatar animation [5, 6, 7, 8, 9] only aims to generate accurate facial lip motion neglecting non-verbal social cues, such as facial emotion and head motions. The lack of emotion and head motion can cause significant degradation of the visual quality during communication.

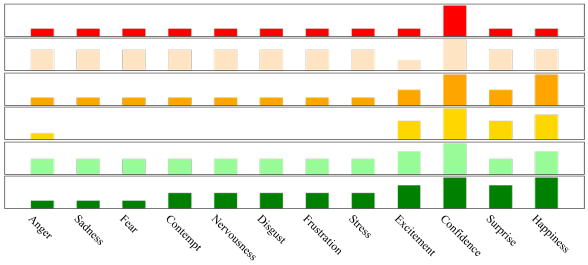

Fig. 1. Subjective emotion recognition results for the same speech. Each row represents an emotion recognition results for each subject.

Emotion and head motion play a vital role in facilitating human communication, allowing individuals to convey their thoughts, feelings, and intentions to others. When facial animations lack the appropriate emotions and motions, they may be perceived as being unresponsive, which can make it difficult to feel connected with them. Therefore, animating emotional talking faces with head motion is important to create realistic content. For the emotional talking face animation, Wang *et al.* proposed the 3DTalkEmo to convert the neutral talking faces into emotional talking faces using emotional condition [10]. This requires a single specific emotional condition for the emotional face generation. However, in real-world scenarios, emotional condition is not provided and no one consistently speaks with a specific emotion. Figure 1 shows 6 subjective test results of participants labeling perceived emotions after listening to speeches [11]. It shows that humans perceive not just one emotion (*i.e.,* confidence) but a combination of emotions (*i.e.,* confidence, excitement, surprise, and happiness). This confirms that humans speak with a complex mixture of emotions rather than a single emotion.

Inspired by this, we propose a **S**peech-**D**riven **E**motional **Talk**ing face generation method (SDETalk) by leveraging the complex emotional source from input speech instead of using a specific emotional condition. To do this, we separately estimate the neutral mouth shape and emotional expressions with head motions. First, the neutral mouth shape is generated according to the context of the speech. Then, we add

*: Corresponding author

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C3011697) and the Yonsei Signature Research Cluster Program of 2023 (2023-22-0008).

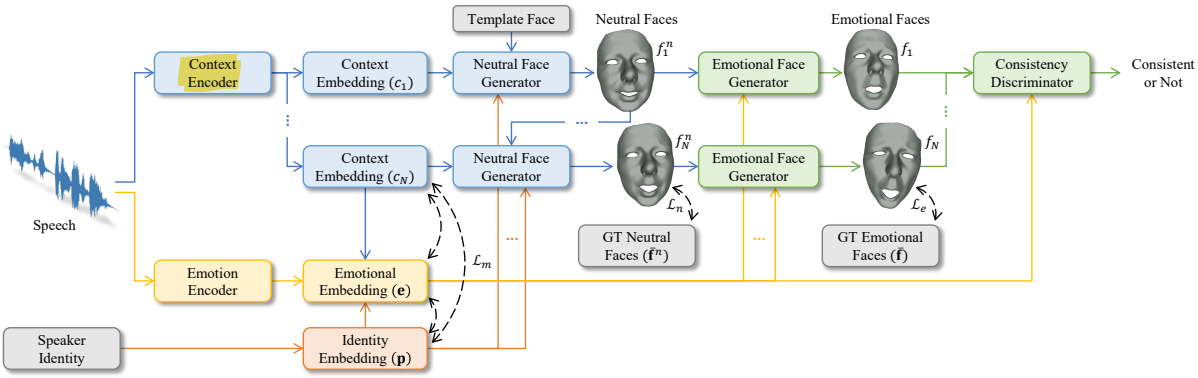

Fig. 2. An overall framework of the SDETalk, which generates emotional talking 3D faces and head motions from speech.

emotional embeddings extracted from the speech to synthesize an emotional face. This allows for accurate lip shapes and natural emotional representations. Furthermore, we generate head motion because it is a decisive factor in maximizing the expressiveness of emotions [12]. For example, nodding the head up and down can indicate positive reactions, while shaking the head from side to side can signify negative reactions. Here, the head motion should be diverse while consistent with the emotions. To ensure diversity, we construct our network using generative adversarial network (GAN) architecture. Based on this, the SDETalk is able to stochastically produce diverse head motions corresponding to the speaker's emotional state. Thus, the SDETalk generates emotional talking faces with natural head motions.

2. EMOTIONAL TALKING FACE ANIMATION

2.1. Network Architecture

Neutral Talking Face Generator. Figure 2 illustrates the entire framework of the *SDETalk*. Given a speech input $\mathbf{x} = \{x_t\}_{t=1}^{kN}$, we first generate sequences of neutral talking faces $\mathbf{f}^n = \{f_t^n\}_{t=1}^{N}$. The neutral talking face generator first produces context representations from the input speech. For the context representation embedding, we construct the context encoder following the state-of-the-art speech representation learning method, wav2vec 2.0 [13]. We initialize the context encoder with the pre-trained wav2vec 2.0 weights and freeze it during the training.

After the context embedding, it should be aligned with face sequences because audio frequency f_a ($f_a = 16$kHz) is conventionally different from a facial animation frequency f_v ($f_v = 30$fps). Thus, the data length of speech (kN) and face (N) are different with ratio of k, where $k = \lceil \frac{f_a}{f_v} \rceil$. To align speech and face sequences, we add a linear interpolation layer after the context encoder to resample the context representations $\mathbf{c} = \{c_t\}_{t=1}^{N}$ according to the frequency of the facial animation as follows:

$$\mathbf{c} = \text{Interp}_{kN \to N}(E_c(\mathbf{x})), \tag{1}$$

where E_c is the context encoder and $\text{Interp}_{kN \to N}$ is the linear interpolation operator from kN samples to N samples. We concatenate personal identity embedding $\mathbf{p}$, which is encoded from a one-hot vector of personal identity using a single linear connected layer for the person-specific generation. Then, it is feds toward the neutral face generator G_n, which is composed of one linear layer. The neutral face is finally defined by adding the template face f_0^n, as follows:

$$f_t^n = G_n(c_t \oplus \mathbf{p}) + f_0^n. \tag{2}$$

Emotional Embedding Network. We extract the emotional embedding from speech for emotional face generation. Figure 3 shows the details of the emotional embedding network. Based on [14, 15], we first convert the audio signal of speech to Mel-spectrogram for emotional embedding. Mel-spectrogram is also resampled by using the Eq. (1). Then, we feed it into temporal and frequency attention blocks sequentially. The temporal attention block is utilized to extract inter-frame temporal features. The frequency attention block is to extract inter-frequency features by looking at global frequency response which is similar to the "classification token" in vision transformer [16]. By fusing temporal and frequency features, we estimate the emotional embedding of speech. In addition, to explicitly guide the emotional embedding, we add an auxiliary emotional classifier. The emotion classifier is composed of a global average pooling layer and two linear layers with GELU activations [17]. It enforces the emotional embedding to accurately establish the emotional information.

For the diversity of expressions and head motions, we construct the diversity mapping process based on the reparameterization trick [18]. Through a single linear layer, the emotional embedding of each frame is encoded into the mean and variance. Based on this, the diversified emotional embedding is computed by multiplying variance and adding mean from the Gaussian noise $\mathcal{N}(0, 1)$ in each frame.

Emotional Face Generator. As shown in Fig. 3, we construct the emotional face generator G based on the transformer decoder composed of N_d cross-modal attention blocks between faces and emotional embeddings. We use the neutral face $\mathbf{f}^n$ as the query and use the emotional embedding

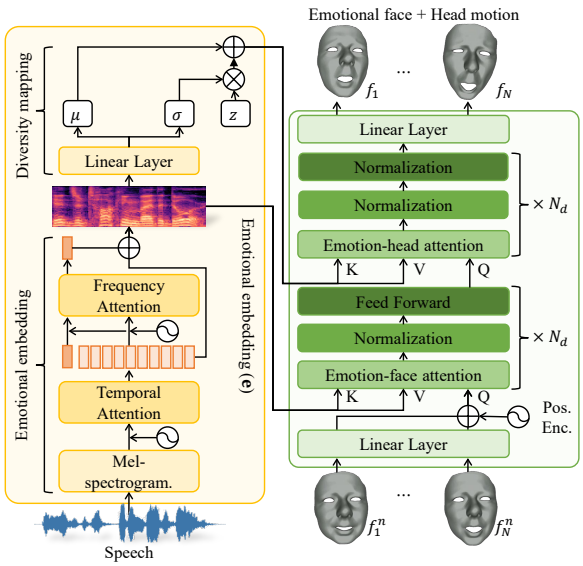

Fig. 3. Details of the emotional embedding network and emotional face generator.

as the key and value of the cross-modal attention block. In addition, for diversity in facial expression and head motion, we feed the diversified emotional embedding $\tilde{\mathbf{e}}$ computed from the diversity mapping network as the key and value into the cross-modal attention block. Emotional faces and head motions (i.e., rotation and translation) are generated through the last linear layer, as follows:

$$[\mathbf{f}^e, \mathbf{R}, \mathbf{T}] = G(\mathbf{f}^n \oplus \mathbf{e}, \tilde{\mathbf{e}}), \tag{3}$$

$$\mathbf{f} = \mathbf{f}^e * \mathbf{R} + \mathbf{T} \tag{4}$$

where $\mathbf{f}^e = \{f_t^e\}_{t=1}^N$, $\mathbf{R} = \{R_t\}_{t=1}^N$, and $\mathbf{T} = \{T_t\}_{t=1}^N$ are the sequence of emotional face, head rotation, and head translation respectively. $\oplus$ is a concatenation operator, and $*$ is a matrix multiplication operator.

Consistency Discriminator. To ensure consistency between speech and generated faces, we design the consistency discriminator based on the transformer encoder architecture. The consistency discriminator D takes emotional embedding $\mathbf{e}$ and real (fake) facial animation with head motion $\mathbf{f}$ ($\bar{\mathbf{f}}$). It regresses concatenated results of emotional embedding and faces to the personal identity and emotion labels.

2.2. Network Training

Our network is trained using adversarial loss including reconstruction, emotional guidance, and diversity losses.

Reconstruction Loss. We set the reconstruction loss by dividing it into neutral face reconstruction loss and expressive face reconstruction loss. The neutral face reconstruction loss $\mathcal{L}_n$ is defined as $\mathcal{L}_n = \left\|\mathbf{f}^n - \bar{\mathbf{f}}^n\right\|_2$, which is the L_2 distance between generated $\mathbf{f}^n$ and ground truth $\bar{\mathbf{f}}^n$ neutral faces. Thus, the loss for expression reconstruction $\mathcal{L}_e$, rotation $\mathcal{L}_r$, and translation $\mathcal{L}_t$ estimation losses are defined as $\mathcal{L}_e = \left\|\mathbf{f}^e - \bar{\mathbf{f}}^e\right\|_2$, $\mathcal{L}_r = \left\|\mathbf{R} - \bar{\mathbf{R}}\right\|_2$, and $\mathcal{L}_t = \left\|\mathbf{T} - \bar{\mathbf{T}}\right\|_2$, where $\bar{\mathbf{f}}^e, \bar{\mathbf{R}}$, and $\bar{\mathbf{T}}$ the are ground truth sequence of expressive faces, rotation, and translation. Therefore, the total reconstruction loss $\mathcal{L}_{rec}$ is defined as:

$$\mathcal{L}_{rec} = \mathcal{L}_n + \mathcal{L}_e + \lambda_r \mathcal{L}_r + \lambda_r \mathcal{L}_t, \tag{5}$$

where λ_r and λ_t are the balance factors between losses. In our implementation, λ_r and λ_t are set as 0.001.

Emotional Guidance Loss. The previous linguistic analysis discovered that the speech includes speaker identity, context, and emotions [14, 15]. Since these factors are independent of each other, we adopt the loss for emotional guidance that minimizes the mutual information between these factors. To compute the upper bound of mutual information, we adopt vCLUB [19]. The mutual information loss $\mathcal{L}_{mi}$ among context $\mathbf{c}$, identity $\mathbf{p}$, and emotion $\mathbf{e}$ is defined as:

$$\mathcal{L}_{mi} = \hat{I}(\mathbf{p}, \mathbf{c}) + \hat{I}(\mathbf{p}, \mathbf{e}) + \hat{I}(\mathbf{c}, \mathbf{e}), \tag{6}$$

where $\hat{I}$ represents the unbiased estimation for vCLUB between identity, context, and emotional factors. In addition, we add the auxiliary emotion classifier to enforce the emotional embedding according to emotion labels correctly. Here, emotion labels represent the mixture of emotional conditions rather than one-hot condition as shown in Fig. 1. We use auxiliary emotion classification loss as $\mathbb{L}_2$ distance between the predicted emotion y_e and the ground truth emotion label $\bar{y}_e$ as $\mathcal{L}_{cls} = \|y_e - \bar{y}_e\|_2$ rather than binary cross entropy loss. Thus, the emotional guidance loss is defined as $\mathcal{L}_e = \mathcal{L}_{mi} + \mathcal{L}_{cls}$.

Diversity Loss. Unlike neutral facial lip motions, expressions and head motions can be different even in the same speech. To encourage pluralistic and diverse expressions and head motion generations, we explicitly regularize G using diversity loss. Following the diverse image generation task [20, 21], we define the diversity loss $\mathcal{L}_d$ as follows:

$$\mathcal{L}_d = \frac{\lambda_f \left\|\mathbf{f}_1^e - \mathbf{f}_2^e\right\|_2 + \left\|\mathbf{R}_1 - \mathbf{R}_2\right\|_2 + \left\|\mathbf{T}_1 - \mathbf{T}_2\right\|_2}{\left\|z_1 - z_2\right\|_2}, \tag{7}$$

where the generated pair of expressions $\mathbf{f}_1^e$, $\mathbf{f}_2^e$, rotations $\mathbf{R}_1$, $\mathbf{R}_2$, and translations $\mathbf{T}_1$, $\mathbf{T}_2$ are produced by emotional flows diversified by two different Gaussian noise z_1, z_2, and λ_f is the balance factor between the diversity of expressions and head motions. Without the diversity loss, the generated results easily to be collapsed into the same mode. In our experiments, λ_f is set as 0.1 because head movement is much more diverse than expression when talking.

Overall Loss. The overall adversarial losses for training the entire network are as follows:

$$\mathcal{L}_{dis} = \mathbb{E}_{\mathbf{e}}[\log D(\mathbf{e}, \bar{\mathbf{f}})] + \mathbb{E}_{\mathbf{e}_f, \mathbf{f}}[1 - \log D(\mathbf{e}_f, \mathbf{f})],$$

$$\mathcal{L}_{gen} = \mathcal{L}_{rec} + \lambda_e \mathcal{L}_e - \lambda_d \mathcal{L}_d + \lambda_g \mathbb{E}_{\mathbf{e}, \mathbf{f}}[\log D(\mathbf{e}, \mathbf{f})],$$

where λ_d, λ_e, and λ_g are the balancing factors for diversity, emotional guidance, and discriminator losses, respectively. In our experiments, we use $\lambda_d = 0.1, \lambda_e = 0.1$, and $\lambda_g = 0.1$.

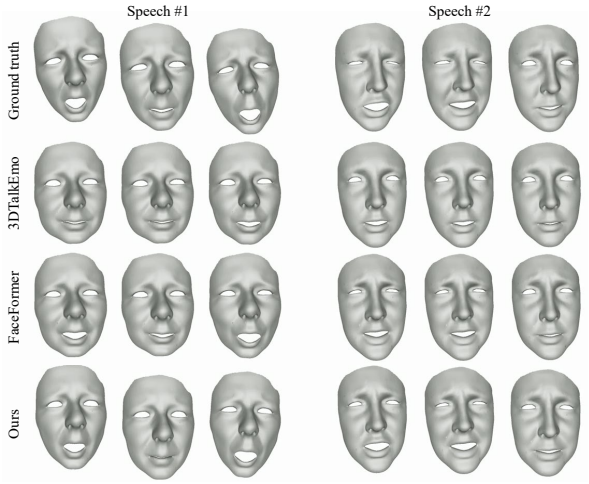

Fig. 4. Qualitative comparison of 3D facial animation results from three different emotional speeches.

Table 1. Comparison of $\mathbb{L}_2$ error on BIWI emotional dataset.

3DTalkEmo	FaceFormer	Ours (w/o $\mathcal{L}_e$)	Ours
7.97	6.64	5.10	**4.92**

3. EXPERIMENTAL RESULTS

Dataset and Implementation Details. We use the BIWI dataset [11] for the training and testing. It contains emotional speech with a complex emotional label. Emotional labels are measured by 12 categories through the subjective test. These labels are not a form of a one-hot label, as shown in Fig. 1. We extract 128-dimensional Mel-spectrograms for emotional embedding. In all attention blocks in our method, we used $N_d = 2$ blocks with 4 attention heads, 128 hidden nodes, and layer normalization [22]. We used the Adam optimizer [23] with a fixed learning rate of $1e^{-4}$ and training converges after 500 epochs using a GPU of NVIDIA RTX 3080 (10GB).

Emotional Talking Facial Animation. For the qualitative comparison, we visualize our animation results and state-of-the-art method, 3DTalkEmo [10] and FaceFormer [5], in Fig. 4. 3DTalkEmo requires specific emotional conditions for the emotional talking face generation. However, since the emotional conditions are not provided in natural speech, we evaluate the 3DTalkEmo without emotional conditions. The result shows that our method produces more emotional face animation than the comparison method. In addition, since our method generates facial expressions and head movements simultaneously, it produces visually more effective results than other methods. For quantitative comparison, we measure the $\mathbb{L}_2$ error between generated and ground truth meshes. Table 1 reports the average $\mathbb{L}_2$ error errors of FaceFormer, and ours in BIWI [11] and our emotional datasets. It implies that emotional guidance loss enables the network effectively extracts emotional information from a speech by reducing the dependency between context and emotion. The result shows that the proposed method outperforms state-of-the-art methods in emotional facial animation.

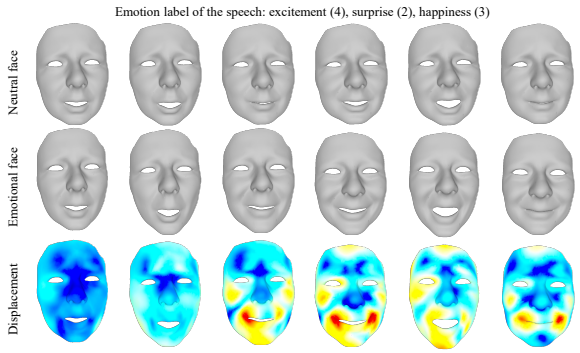

Fig. 5. Visualization of neutral faces, emotional faces, and displacement maps.

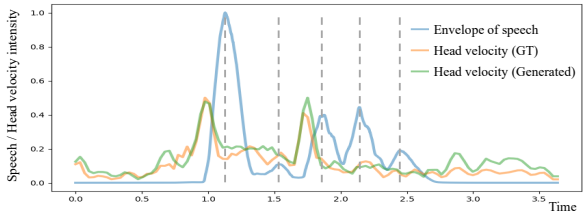

Fig. 6. Consistency evaluation between speech and head motions. Gray dashed lines are the peak of the speech.

In addition, we visualize the neutral and emotional faces estimated from the SDETalk in Fig. 5. This is the result of a speech that is a fusion of emotions of excitement, surprise, and happiness. It shows that the proposed method satisfactorily represents the emotional faces from emotional speech.

Speech and Head Motion Consistency. Figure 6 visualizes an example among speech envelope, computed by using Hilbert transform, and velocity of the ground truth and generated head motions. The results show that the ground truth head motions have a correlation to the audio envelope. Especially, ground truth head motion tends to move a lot right before or after speaking hard. The generated head motions have a similar tendency to the ground truth head motions. It means that the generated head moves naturally by having consistency with the speech signal.

4. CONCLUSION

We present a novel learning framework to animate emotional talking 3D faces with rhythmical head motions from speech. By leveraging emotional sources from speech, our proposed method produces rich facial expressions and rhythmical head motions from speech without any emotional constraints. It effectively matches facial animation and speech, increasing the emotional consistency and naturalness of facial animations. We believe that our method can be applied in various 3D applications, such as telepresence systems, virtual reality, and computer games, by automatically animating emotional talking faces from speech.

5. REFERENCES

[1] Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee, "A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres," in *CVPR*, 2022, pp. 3490–3500.

[2] Kyungjune Lee, Jeonghaeng Lee, Hyucksang Lee, Mingyu Jang, Seongmin Lee, and Sanghoon Lee, "FaceClone: Interactive facial shape and motion cloning system using multi-view images," in *IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 2023, pp. 512–513.

[3] Seongmin Lee, Hyunse Yoon, Jiwoo Kang, Jungsu Kim, Jiwan Son, Jungwoo Huh, and Sanghoon Lee, "Video-based stabilized 3D face alignment using temporal multi-discrimination," in *International Workshop on Multimedia Signal Processing*. IEEE, 2023, pp. 1–6.

[4] Jiwoo Kang, Seongmin Lee, and Sanghoon Lee, "Competitive learning of facial fitting and synthesis using UV energy," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 5, pp. 2858–2873, 2021.

[5] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura, "Faceformer: Speech-driven 3D facial animation with transformers," in *CVPR*, 2022, pp. 18770–18780.

[6] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh, "Meshtalk: 3D face animation from speech using cross-modality disentanglement," in *ICCV*, 2021, pp. 1173–1182.

[7] Jingying Liu, Binyuan Hui, Kun Li, Yunke Liu, Yu-Kun Lai, Yuxiang Zhang, Yebin Liu, and Jingyu Yang, "Geometry-guided dense perspective network for speech-driven facial animation," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[8] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black, "Capture, learning, and synthesis of 3D speaking styles," in *CVPR*, 2019, pp. 10101–10111.

[9] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–12, 2017.

[10] Qianyun Wang, Zhenfeng Fan, and Shihong Xia, "3DTalkEmo: Learning to synthesize 3d emotional talking head," *arXiv preprint arXiv:2104.12051*, 2021.

[11] Gabriele Fanelli, Juergen Gall, Harald Romsdorfer, Thibaut Weise, and Luc Van Gool, "A 3D audio-visual corpus of affective communication," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 591–598, 2010.

[12] Steven R Livingstone and Caroline Palmer, "Head movements encode emotions during speech and song.," *Emotion*, vol. 16, no. 3, pp. 365, 2016.

[13] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *NeurIPS*, vol. 33, pp. 12449–12460, 2020.

[14] Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li, "Disentanglement of emotional style and speaker identity for expressive voice conversion," *arXiv preprint arXiv:2110.10326*, 2021.

[15] Disong Wang, Liqun Deng, Yu Ting Yeung, Xiao Chen, Xunying Liu, and Helen Meng, "VQMIVC: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," *arXiv preprint arXiv:2106.10132*, 2021.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[17] Dan Hendrycks and Kevin Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[18] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[19] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *ICML*. PMLR, 2020, pp. 1779–1788.

[20] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *CVPR*, 2020, pp. 8188–8197.

[21] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang, "Mode seeking generative adversarial networks for diverse image synthesis," in *CVPR*, 2019, pp. 1429–1437.

[22] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *NeurIPS*, 2016.

[23] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *ICLR*, 2014.