INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Emotion-Aware Audio-Driven Face Animation via Contrastive Feature Disentanglement

Xin Ren, Juan Luo*, Xionghu Zhong, Minjie Cai†*

Hunan University, China

Abstract

In this paper, we tackle the problem of audio-driven face animation which aims to synthesize a realistic talking face given a piece of driven speech. Directly modeling the mapping from audio feature to facial expression is challenging, since people tend to have different talking styles with momentary emotion states as well as identity-dependent vocal characteristics. To address this challenge, we propose a contrastive feature disentanglement method for emotion-aware face animation. The key idea is to disentangle the features for speech content, momentary emotion and identity-dependent vocal characteristics from audio features with a contrastive learning strategy. Experiments on public datasets show that our method can generate more realistic facial expression and enables synthesis of diversified face animation with different emotion.

Index Terms: audio-driven face animation, feature disentanglement, contrastive learning, neural network

1. Introduction

Audio-driven face animation is an emerging multimedia technology with wide potential applications, such as teleconference, virtual reality, film dubbing and so on. Recent works on this topic [1, 2, 3, 4] either directly model the relationship between audio input and facial expression, or focus on learning the intermediate representation between different modalities. However, few works have considered the multiple information (e.g., speech content, emotion state) involved in audio features that have different relationships with facial expression. For example, a person in a joyful mood may speak with more widely open mouth than a person in a depressed mood.

One of the main challenges in audio-driven face animation is the coupling of multiple information within the audio signals. In a audio segment, there are the content of speech, the momentary emotion states of the speaker, and the individual vocal characteristics. While the speech content and the emotion states have strong relationship with facial expression, the vocal characteristics are unique for different persons but are nearly irrelevant to the facial expression. Therefore, directly modeling the relationship between the audio signal and facial expression may lead to limited performance due to the coupling of irrelevant information. Furthermore, to enable the synthesis of diversified face animation with different emotion, the decoupling of the content and emotion from the audio is also needed.

In this work, our aim for emotion-aware audio-driven face animation that is consistent with the input audio. Our key idea is to disentangle the information of speech content, emotion states and pronunciation characteristics from audio features with a contrastive learning strategy. Our method formulates the task as a two-stage audio-driven image translation problem. In the first stage, information of speech content, emotion states and pronunciation characteristics are extracted from input audio as speech code, emotion code and identity code respectively with three branches of Multiple Layer Perceptron (MLP). These codes are used for predicting the face landmarks corresponding to speech content and speaking styles separately. In the second stage, target images are generated by an image-to-image translation network using a reference image and the predicted landmarks as input.

To enforce feature disentanglement, we develop a contrastive learning strategy based on datasets which are composed by speech videos of different people, with each person talking in multiple videos with different emotion states. We train and evaluate our method on public VoxCeleb2 dataset[5] and LRW dataset[6]. Quantitative and qualitative experiments show that our proposed method achieves state-of-the-art performance.

In summary, our contributions include the following:

- We propose an emotion-aware audio-driven face animation method which enables synthesis of realistic facial expression with diversified talking styles.
- We develop a contrastive learning strategy to disentangle multiple information from the audio features.

2. Related Work

Audio-driven face animation. Given an input audio stream, audio-driven face animation aims to generate a sequence of talking face images given a piece of speech audio. Previous work can be mainly divided into three directions. In the first direction, the feature extracted from audio signals are studied. Some methods [4, 7, 8, 9, 10, 11] directly used Mel-scale Frequency Cepstral Coefficients (MFCC). NVP [12], Voca [13], and FICIAL [1] used DeepSpeech [14] to extract audio features. MakeitTalk [3] extracted content embedding and speaker identity embedding from input audio by [15]. In the second direction, intermediate representation learning between audio features and facial expression is studied. [3, 16, 17] used landmarks as their intermediate representation. [1, 12, 18, 19] used 3D Morphable Face Models coefficients as their intermediate representation. The third direction focuses on the image synthesis. Some methods [3, 16] used image-to-image translation to generate the final image, while others [1, 20] used traditional rendering pipeline and fine tuning to generate the final image.

Style-aware face generation. Considering that people may

*These two authors contributed equally to the paper.

†Corresponding Author: Minjie Cai (caiminjie@hnu.edu.cn).

This work is supported by the Hunan Provincial Natural Science Foundation of China under Grant 2022JJ20015, and by the National Natural Science Foundation of China under Grant 61971186.

10.21437/Interspeech.2023-358

speak in different styles even for the same content, style-aware face generation aims to identify different talking styles from audio input in addition to the content of the speech [21]. [22] used BiLSTM [23] network to learn the information related to posture in speech. [13] used the one-hot encoding method to encode the data in the training set, and used this one-hot encoding vector as a control variable to distinguish different people's styles. However, the model needs to be retrained if new people join, which is not time-efficient. [12] used a linear mapping to represent the talking styles of different people. If new persons attend, they only need to train the linear mapping corresponding to the new person without retraining the model, thus saving a lot of training time. However, training such a linear mapping requires collecting a video of the person and the three-dimensional facial structure corresponding to the face in the video, which greatly reduces the convenience. [3] separated the content of the conversation and the style of the speaker explicitly, however they didn't take into account that the same person may have momentary emotion states and thus changing talking styles in different situations. In contrast, our method explicitly considers the disentanglement of speech content, momentary emotion and identity-dependent vocal characteristics.

3. Method

Task definition. Given a reference image $\boldsymbol{I}$ containing a face to be animated and a piece of speech audio $\boldsymbol{A}$, the goal of the task is to generate a video $\boldsymbol{V}$ with facial expression that is consistent with the speech with respect to the content and temporal duration. The task can be formulated as a two-stage process. In the first stage, the input audio are transformed to a sequence of displacement of the face landmarks. In the second stage, the target video of face animation is generated with an image-to-image translation model which transfers the reference image and the sequence of face landmark images to a sequence of synthesized face images. The overview of the proposed method is demonstrated in Figure 1.

3.1. Preprocessing

Audio feature extraction. To eliminate the effects of different languages, recording artifacts and noise in the audio, we extract audio feature by AutoVC [15] which is a few-shot voice conversion method to separate the audio into the speech content and the identity information. The obtained audio features are $\boldsymbol{F}_a \in \mathbb{R}^{T\times D}$, where $\boldsymbol{T}$ is the frame number of audio and $\boldsymbol{D}$ is the dimension of AutoVC feature. We divide every 20ms of audio into a frame.

Face alignment. The head poses in a video are usually dynamic and may affect the position of face landmarks even for a static facial expression. In order to learn a stable mapping between audio features and face landmarks, we need to eliminate the effects of head poses by aligning the faces into a consistent head pose. First, we extract the 3D facial landmarks $\boldsymbol{L}_m \in \mathcal{R}^{68\times 3}$ from a video [24]. Then we calculate the average facial landmarks, and adopt Iterative Closest Point[25] algorithm to align facial landmarks of all face images with the average. Finally, the 3D face landmarks are projected onto the image space by orthogonal projection. Our method is trained on aligned 2D facial landmarks $\boldsymbol{L} \in \mathcal{R}^{68\times 2}$.

3.2. Contrastive Feature Disentanglement

We first encode the speech content, emotion and identity from audio features, and then develop a contrastive learning strategy to disentangle the multiple information.

Content encoder. The purpose of this part is to separate out the audio features which are only relevant to the speech content. Similar to [3], we first encode the audio feature with a three-layer MLP, and then use a LSTM [26] to capture temporal information of the speech. After that, the content encoding $\boldsymbol{E}_{content} \in \mathbb{R}^{T\times 256}$ is obtained:

$$\boldsymbol{E}_{content} = LSTM(MLP_c(\boldsymbol{F}_a)) \tag{1}$$

Style encoder. In a piece of speech audio, in addition to the content of the speech, there also involves the momentary emotion and identity-dependent vocal characteristics of the speaker which affect the talking style and thus the facial expression in different way. We use a style encoder to separate the talking style from the audio. Then, we use two MLPs to embed $\boldsymbol{F}_a$ into the emotion embedding $\boldsymbol{E}_{emotion} \in \mathbb{R}^{T\times 128}$ that encodes the momentary emotion and the identity embedding $\boldsymbol{E}_{identity} \in \mathbb{R}^{T\times 128}$ that encodes identity-dependent vocal characteristics. To capture the dependency of talking style on the identity embedding and the emotion embedding, we feed $\boldsymbol{E}_{emotion}$ and $\boldsymbol{E}_{identity}$ into a self-attention encoder [27] to get the style embedding $\boldsymbol{E}_{style} \in \mathbb{R}^{T\times 256}$. The process in the style encoder can be formulated as follows:

$$\boldsymbol{E}_{emotion} = MLP_e(\boldsymbol{F}_a) \tag{2}$$

$$\boldsymbol{E}_{identity} = MLP_i(\boldsymbol{F}_a) \tag{3}$$

$$\boldsymbol{E}_{style} = Attn(\boldsymbol{E}_{emotion}, \boldsymbol{E}_{identity}) \tag{4}$$

Contrastive learning. We develop a contrastive learning strategy to disentangle different feature embeddings of $\boldsymbol{E}_{content}$, $\boldsymbol{E}_{emotion}$ and $\boldsymbol{E}_{identity}$. We carefully construct a batch of four three clips for training. The first and the second are from the same video, and are denoted as $\boldsymbol{V}_r$ and $\boldsymbol{V}_s$ respectively. The third one records the same person speaking but in different videos/scenes and is denoted as $\boldsymbol{V}_p$. It is assumed that the emotion states in different video clips of the same video are more similar than that in different videos. We use $\boldsymbol{V}_r$, $\boldsymbol{V}_s$ and $\boldsymbol{V}_p$ for disentangling the emotion embedding and use $\boldsymbol{V}_r$ and $\boldsymbol{V}_p$ for disentangling the identity embedding.

To disentangle the information of emotion, we use $\boldsymbol{V}_r$, $\boldsymbol{V}_s$ and $\boldsymbol{V}_p$ to construct a contrastive loss. The idea is to enforce similarity of emotion embedding between $\boldsymbol{E}^r_{emotion}$ and $\boldsymbol{E}^s_{emotion}$ for $\boldsymbol{V}_r$ and $\boldsymbol{V}_s$, as well as dissimilarity between $\boldsymbol{E}^r_{emotion}$ and $\boldsymbol{E}^p_{emotion}$ for $\boldsymbol{V}_r$ and $\boldsymbol{V}_p$. Besides, to provide weights for the contrastive loss, we also adopt a pretrained facial expression recognition model [28] to obtain the softmax probabilities of facial expression of the three videos, denoted as $\boldsymbol{e}_r$, $\boldsymbol{e}_s$ and $\boldsymbol{e}_p$. The contrastive loss is formulated as:

$$\mathcal{L}_{emotion} = \lambda_{r,s} \cdot \boldsymbol{d}_{r,s} + (1 - \lambda_{r,p}) \cdot max(0.1 - \boldsymbol{d}_{r,p}, 0) \tag{5}$$

where $d_{r,s}$ represent the mean absolute difference between $\boldsymbol{E}^r_{emotion}$ and $\boldsymbol{E}^s_{emotion}$ and similarly is $d_{r,p}$. $\lambda_{r,s}$ is the weight for $d_{r,s}$ and is computed as cosine similarity of facial expression between $\boldsymbol{e}_r$ and $\boldsymbol{e}_s$. $\lambda_{r,s}$ is computed similarly.

To disentangle the information of identity, we make an assumption that the identity code $\boldsymbol{E}_{identity}$ is the same for all videos of the same speaker. In concrete, we enforce the similarity between the identity code $\boldsymbol{E}^r_{identity}$ of video $\boldsymbol{V}_r$ and the identity code $\boldsymbol{E}^p_{identity}$ of video $\boldsymbol{V}_p$:

$$\mathcal{L}_{identity} = |\boldsymbol{E}^r_{identity} - \boldsymbol{E}^p_{identity}| \tag{6}$$

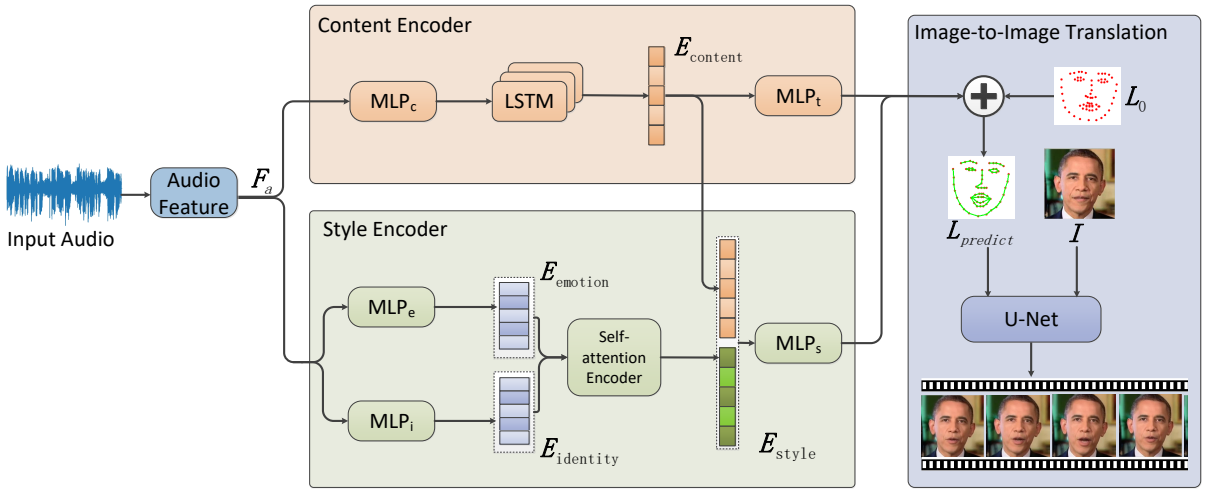

Figure 1: *The overview of the proposed method. Content encoder and style encoder are used to predict displacement of face landmarks from input audio with consideration of the speech content and talking styles respectively. Image-to-image translation is used to synthesize the talking face images from the predicted face landmarks and the reference face image.*

3.3. Landmark Prediction

In this paper, the audio-driven facial expression is represented by the displacement of face landmarks with respect to the reference face image. The displacement of face landmarks related to the speech content $\boldsymbol{D}_{content} \in \mathbb{R}^{T\times 68\times 2}$ is predicted by a MLP given input of the content embedding $\boldsymbol{E}_{content}$. Similarly, the displacement of face landmarks related to the talking style $\boldsymbol{D}_{style} \in \mathbb{R}^{T\times 68\times 2}$ is predicted by a MLP given input of the style embedding $\boldsymbol{E}_{style}$ and the content embedding $\boldsymbol{E}_{content}$. Finally, the predicted face landmarks $\boldsymbol{L}_p \in \mathbb{R}^{T\times 68\times 2}$ are obtained by adding the predicted displacement with the face landmarks of the reference image $\boldsymbol{L}_0 \in \mathbb{R}^{68\times 2}$ which are 2D orthogonal projection of $\boldsymbol{L}_{\boldsymbol{m}}$. The landmark prediction process is formulated as follows:

$$\boldsymbol{D}_{content} = MLP_t(\boldsymbol{E}_{content}) \tag{7}$$

$$\boldsymbol{D}_{style} = MLP_s(Concat(\boldsymbol{E}_{content}, \boldsymbol{E}_{style})) \tag{8}$$

$$\boldsymbol{L}_p = \boldsymbol{L}_0 + \boldsymbol{D}_{content} + \boldsymbol{D}_{style} \tag{9}$$

Loss function. For the landmark prediction, the goal is to predict realistic face landmarks from audio. We define the loss function $\mathcal{L}_{landmark}$ as the absolute value of the coordinate difference between the landmarks $\boldsymbol{L}_p$ predicted from the audio and the landmarks $\boldsymbol{L}_v$ detected from the video which is considered as the ground-truth. The loss function is formulated as follows:

$$\mathcal{L}_{landmark} = |\boldsymbol{L}_v - \boldsymbol{L}_p| \tag{10}$$

3.4. Image-to-Image Translation

To synthesize the target face image with facial expression that is consistent to the input audio, we adopt an image-to-image translation method. We use a U-Net model which takes as input the stack of the reference face image $\boldsymbol{I} \in \mathbb{R}^{3\times W\times H}$ with the images of predicted face landmarks $\{\boldsymbol{M}_t\} \in \mathbb{R}^{T\times 3\times W\times H}$ that are constructed following previous work [16]. The output of the U-Net model is a sequence of synthesized face images $\{\boldsymbol{I}_t\} \in \mathbb{R}^{T\times 3\times W\times H}$.

Loss function. For image-to-image translation, our goal is to generate the image close to the ground truth image. Thus, the loss function is formulated as the difference of pixel values between the synthesized image and the ground truth image.

$$\mathcal{L}_{image} = |\boldsymbol{I}_s - \boldsymbol{I}_t| \tag{11}$$

4. Experiment

4.1. Dataset

We train and evaluate our method on two public datasets:

VoxCeleb2 Dataset [5]. The dataset contains over 1 million utterances for 6112 celebrities, extracted from videos uploaded to YouTube. In this dataset, each speaker is recorded in multiple videos of different scenes.

LRW Dataset [6]. The dataset consists of up to 1000 utterances of 500 different words, spoken by hundreds of speakers. People in each video only speak one word.

4.2. Implementation Details

Network architecture. The network architectures of MLPs and LSTM are shown in Table 1. We use the LeakyReLU activation function and batch-normal after the hidden layers of each MLP. For the LSTM, we also use a dropout of 0.2.

Table 1: *Dimensions of each layer in MLP and LSTM.*

Network	Layer dimension
MLP_c	(80, 256, 161)
MLP_i	(80, 256, 128, 128)
MLP_e	(80, 256, 128, 128)
MLP_t	(256, 512, 256, 136)
MLP_s	(256, 512, 256, 136)
$LSTM$	(161, 256, 256, 256)

Training procedure. The proposed model is trained in three steps:

Step-1: We first train the content encoder part (MLP_c, $LSTM$ and MLP_t) with $\mathcal{L}_{landmark}$.

Step-2: We then train the style encoder part (MLP_e, MLP_i, self-attention encoder and MLP_s) with $\mathcal{L}_{emotion} + \mathcal{L}_{identity} + \mathcal{L}_{landmark}$.

Step-3: We finally train the image-to-image translation part with $\mathcal{L}_{image}$.

We use PyTorch [29] for implementation. We use Adam optimizer [30] during training and set the learning rate to 1e-5 and with weight decay of 1e-5. The full model is trained on a Nvidia 3090 GPU for nearly 60 hours.

4.3. Quantitative Evaluation

Compared methods. To demonstrate the effectiveness of our method, we compare with state-of-the-art methods including MakeitTalk [3] and PC-AVS [2]. To verify the role of style encoder and the contrastive loss function, we also conduct an ablation study by comparing with two baselines of our method. One baseline is the removal of the style encoder part, which is named as "Ours (w/o D_{style})". Another baseline named as "Ours (w/o $\mathcal{L}_{emotion}$)" is the removal of the contrastive loss function, meaning that the style encoder is trained with the same loss function as the content encoder.

Evaluation metric. We use landmark distance (LMD) and emotion similarity (ES) as evaluation metrics. LMD is computed as the average Euclidean distance between the predicted landmarks and the ground-truth landmarks obtained by [31]. Through the metric of LMD, the accuracy of audio-driven landmark prediction can be measured. ES is computed as the cosine similarity of facial expressions between the synthesized image and the real image. The facial expression is represented as the softmax probability estimated by a pretrained facial expression recognition model [28]. The metric of ES can be used to measure the quality of audio-driven emotion estimation.

Table 2: *Performance comparison of different methods. LMD and ES are used as the evaluation metric.*

Method	VoxCeleb2 [5]		LRW [6]	
	LMD↓	ES↑	LMD↓	ES↑
MakeitTalk [3]	9.16	0.877	7.13	0.884
PC-AVS [2]	6.88	0.879	**3.93**	-
Ours (w/o D_{style})	4.78	0.891	4.42	0.895
Ours (w/o $\mathcal{L}_{emotion}$)	4.59	0.903	4.06	**0.900**
Ours (full)	**4.15**	**0.919**	-	-

Results. The results are shown in Table 2. It can been seen that our method achieves best performance for both metrics of LMD and ES on the VoxCeleb2 dataset. Since the LRW dataset doesn't provide the speech videos of the same person in different scenes, our model cannot be fully trained with the proposed contrastive loss. Still our method can achieve nearly state-of-the-art performance with LMD of 4.06. The superior performance with the metric of ES on both datasets supports our proposal of the proposed style encoder.

As for the ablation study, Our full model significantly outperforms our baseline without $\mathcal{L}_{emotion}$, demonstrating the effectiveness of the contrastive learning for style encoder. Comparing the two baselines, it can be seen that although the use of style encoder improves the performance, its advantage is not fully exploited without the contrastive learning.

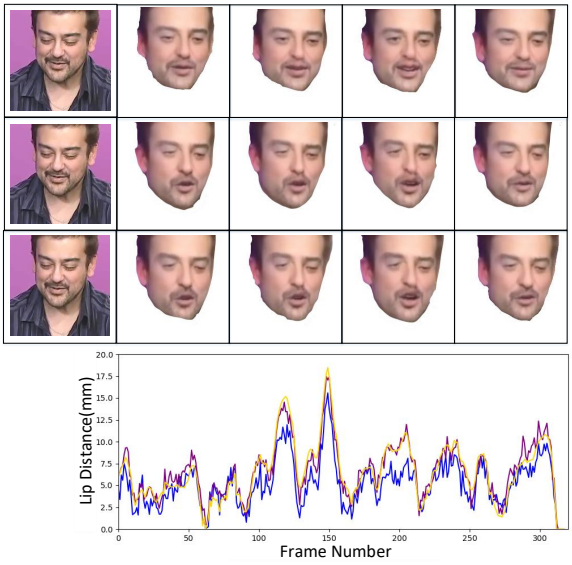

Figure 2: *Qualitative results of face animation for the same driven audio conditioned on different emotion states ($E_{emotion}$). Below is the distance between the lower lip and the upper lip for the face animation.*

4.4. Qualitative Evaluation

To investigate the emotion space learned through our method, we conduct two qualitative experiments. We firstly use the trained model to obtain the emotion embeddings of all training videos. Then we carry out principal component analysis on these emotion embeddings and use the first principal component to sample different emotion embeddings. Three videos are synthesized with the same audio and are shown in the upper part of Figure 2, of which the second and the third are synthesized by adding the extracted emotion embeddings with increasing value along the first principal component. We observe that the facial expression of the three videos changes smoothly from solemn to excited emotion, with gradually exaggerated lip motion.

We also visualize the curve of the distance between the upper and lower lips for the three sampled videos. It can be seen from the bottom of Figure 2 that while different curves show similar trend of variation, the lip distance varies significantly among different videos. The results indicate that our method can learn from audio diversified talking styles while maintaining the consistency between audio and facial expression.

5. Conclusion

In this paper, we propose a method that enables emotion-aware audio-driven face animation. We propose a model to explicitly disentangle talking style from speech content with a contrastive learning strategy. Experiments on two public datasets show that our method not only can synthesize realistic facial expression, but also enables synthesis of diversified facial expression with learned emotion space. Remaining issues of our method are that our method can only synthesize 2D facial expression, and the quality of synthesized images is limited by the quality of the reference image and the change of head positions. In future work, we would try to use 3D parametric face model to model the mapping between audio and 3D facial expression, and to further improve the quality of the synthesized image.

6. References

[1] C. Zhang, Y. Zhao, Y. Huang, M. Zeng, S. Ni, M. Budagavi, and X. Guo, "Facial: Synthesizing dynamic talking face with implicit attribute learning," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3847–3856.

[2] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, "Pose-controllable talking face generation by implicitly modularized audio-visual representation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[3] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, "MakeItTalk: Speaker-Aware Talking-Head Animation," *ACM Transactions on Graphics*, vol. 39, no. 6, pp. 1–15, dec 2020. [Online]. Available: https://doi.org/10.1145%2F3414685.3417774

[4] A. Richard, M. Zollhx00F6;fer, Y. Wen, F. de la Torre, and Y. Sheikh, "Meshtalk: 3d face animation from speech using cross-modality disentanglement," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1153–1162.

[5] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[6] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision*, 2016.

[7] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7824–7833.

[8] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 484–492. [Online]. Available: https://doi.org/10.1145/3394171.3413532

[9] M. C. Doukas, S. Zafeiriou, and V. Sharmanska, "Headgan: One-shot neural head synthesis and editing," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[10] L. Chen, Z. Li, R. K. Maddox, Z. Duan, and C. Xu, "Lip movements generation at a glance." Berlin, Heidelberg: Springer-Verlag, 2018, p. 538–553. [Online]. Available: https://doi.org/10.1007/978-3-030-01234-2_32

[11] K. Cheng, X. Liu, Y.-m. Cheung, R. Wang, X. Xu, and B. Zhong, "Hearing like seeing: Improving voice-face interactions and associations via adversarial deep semantic matching network," ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 448–455. [Online]. Available: https://doi.org/10.1145/3394171.3413710

[12] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner, "Neural voice puppetry: Audio-driven facial reenactment," *ECCV 2020*, 2020.

[13] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3d speaking styles," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 093–10 103.

[14] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, and Z. Zhu, "Deep speech 2: End-to-end speech recognition in english and mandarin," *Computer Science*, 2015.

[15] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "AutoVC: Zero-shot voice style transfer with only autoencoder loss," ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 5210–5219. [Online]. Available: http://proceedings.mlr.press/v97/qian19c.html

[16] Y. Lu, J. Chai, and X. Cao, "Live speech portraits: Real-time photorealistic talking-head animation," *CoRR*, vol. abs/2109.10595, 2021. [Online]. Available: https://arxiv.org/abs/2109.10595

[17] O. Wiles, A. S. Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIII*. Berlin, Heidelberg: Springer-Verlag, 2018, p. 690–706. [Online]. Available: https://doi.org/10.1007/978-3-030-01261-8_41

[18] L. Chen, G. Cui, C. Liu, Z. Li, Z. Kou, Y. Xu, and C. Xu, "Talking-head generation with rhythmic head motion," in *European Conference on Computer Vision*, 2020.

[19] A. Lahiri, V. Kwatra, C. Frueh, J. Lewis, and C. Bregler, "Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2754–2763.

[20] Y. Wang, Y. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, and T. Li, "Contrastive regression for domain adaptation on gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 19 376–19 385.

[21] T. Li, X. Wang, Q. Xie, Z. Wang, M. Jiang, and L. Xie, "Cross-speaker emotion transfer based on prosody compensation for end-to-end speech synthesis," in *Interspeech*, 2022.

[22] D. Greenwood, I. Matthews, and S. Laycock, "Joint learning of facial expression and head pose from speech," 09 2018, pp. 2484–2488.

[23] S. Zhang, D. Zheng, X. Hu, and M. Yang, "Bidirectional long short-term memory networks for relation classification," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China, Oct. 2015, pp. 73–78. [Online]. Available: https://aclanthology.org/Y15-1009

[24] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.

[25] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[26] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32, no. 2. Bejing, China: PMLR, 22–24 Jun 2014, pp. 1764–1772. [Online]. Available: https://proceedings.mlr.press/v32/graves14.html

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv*, 2017.

[28] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6896–6905.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[31] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.