Data-Driven Expressive 3D Facial Animation Synthesis for Digital Humans

Kazi Injamamul Haque
k.i.haque@uu.nl
Utrecht University
Utrecht, The Netherlands

ABSTRACT

This doctoral research focuses on generating expressive 3D facial animation for digital humans by studying and employing data-driven techniques. Face is the first point of interest during human interaction, and it is not any different for interacting with digital humans. Even minor inconsistencies in facial animation can disrupt user immersion. Traditional animation workflows prove realistic but time-consuming and labor-intensive that cannot meet the ever-increasing demand for 3D contents in recent years. Moreover, recent data-driven approaches focus on speech-driven lip synchrony, leaving out facial expressiveness that resides throughout the face. To address the emerging demand and reduce production efforts, we explore data-driven deep learning techniques for generating controllable, emotionally expressive facial animation. We evaluate the proposed models against state-of-the-art methods and ground-truth, quantitatively, qualitatively, and perceptually. We also emphasize the need for non-deterministic approaches in addition to deterministic methods in order to ensure natural randomness in the non-verbal cues of facial animation.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Animation**; • **Human-centered computing** → *User studies.*

KEYWORDS

facial animation synthesis, deep learning, digital humans, mesh animation, blendshape animation

ACM Reference Format:
Kazi Injamamul Haque. 2023. Data-Driven Expressive 3D Facial Animation Synthesis for Digital Humans. In *SIGGRAPH Asia 2023 Doctoral Consortium (SA Doctoral Consortium '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3623053.3623369

1 INTRODUCTION AND MOTIVATION

Computer animation is a form of digital art that has become an increasingly popular medium in the entertainment industry, with applications in film, video games, advertising, and more. With the inception of "metaverse", the demand for animated content has now reached an unprecedented level with additional applications beyond those pertaining to the entertainment, such as- healthcare, education, and virtual social interaction. Traditional workflows for creating animated content cannot meet this ever-increasing demand, as they involve several steps, including modeling, rigging, texturing, animation, and rendering, which require specialized software tools and a skilled team of animators, artists, and technicians. Although traditional animation workflows are irreplaceable for scripted scenes in high-budget projects such as animated films and video game cut-scenes, they are not feasible for real-time 3D applications pertaining to XR. In such applications, animations need to be generated on-the-fly, depending on the interaction context. As a result, new workflows for generating animation driven by data need to be proposed, implemented, and properly evaluated. Recent developments in deep learning techniques have proven to be particularly useful for such data-driven workflows for gesture synthesis [Nyatsanga et al. 2023] and they hold immense potential for being applied to facial animation synthesis as well. More specifically, speech-driven facial animation synthesis is now being widely explored in both academic research [Cudeiro et al. 2019; Fan et al. 2021a; Xing et al. 2023] and industry [JALI 2023]. While these works mostly concentrate on lip-sync, the ability to synthesize appropriate expressive non-verbal facial cues, is not yet being addressed as much. In this research, motivated by the aforementioned facts, we will address the gap by employing and evaluating state-of-the-art deep learning techniques for the 3D facial animation generation task ensuring not only lip synchrony but also emotional expressiveness that resides throughout the face.

As humans, we are particularly perceptive to even the most subtle facial cues while interacting with others. To ensure an immersive experience, digital humans must faithfully convey real-life naturalness not only with body gestures but also with appropriate facial expressions coherent with the context of the interaction. Extensive research has been conducted and is ongoing in the domains of 3D face shape generation[Li et al. 2017; Sanyal et al. 2019], detection and analysis of facial action units and expressions[Tian et al. 2001], and performance-based or motion-capture-based facial motion mapping in both offline and online settings[van der Struijk et al. 2018]. In contrast, fewer research works are being carried out, and even fewer are open researches that address audio/speech-driven facial animation. Moreover, current production workflows can benefit from this as these models can generate speech-driven facial animation in games and movies by just using voice acting, which would significantly decrease production time and cost. Some studies investigated only audio-driven 3D facial animation[Cudeiro et al. 2019; Fan et al. 2021a; Xing et al. 2023], while others investigated text-driven animation[Hu et al. 2021]. Another direction investigates both audio and text-driven facial animation[Fan et al.

SA Doctoral Consortium '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0392-8/23/12…$15.00
https://doi.org/10.1145/3623053.3623369

2021b]. Since audio is highly correlated with the lower facial region (i.e., lip and jaw), and emotion and contextual information are correlated with the upper face region, extensive research needs to be carried out that generates full facial animation that is expressive, driven not only by speech but also emotion and contextual information. Moreover, facial motion, more specifically non-verbal elements residing in face is non-deterministic in nature [Ng et al. 2022]. Therefore, models that can resemble such non-determinism in generation of diverse facial animations, should be proposed and evaluated.

The main goal of this research is to explore state-of-the-art deep learning techniques in order to propose multimodal data-driven facial animation synthesis approaches for both offline and online use-cases that can process multimodal information and generate facial animations of virtual humans in accordance with input speech, emotion, context, etc. The proposed approaches will be extensively evaluated in terms of both objective and subjective analyses, along with ablations studies.

1.1 Research Question(s)

With the aforementioned motivation, the following main research questions will govern the progress of this doctoral research -

- **Main RQ1.** Can deep learning approaches synthesize data-driven 3D facial animation that is as good as performance capture (i.e.- the ground-truth)?
- **Main RQ2.** Do animations synthesized by the data-driven approaches perform well in terms of perceived realism?

Where RQ1 would be answered by the objective evaluation while RQ2 would be answered by subjective evaluation. Additionally, the following sub research questions will allow us to conduct experiments that will help us answer the two main research questions-

- **Sub RQ1.** Can we disentangle emotional expressiveness from speech input to control the expressiveness of the generated animation?
- **Sub RQ2.** Do non-deterministic generative models synthesize more perceived realism in facial animation compared to deterministic models?
- **Sub RQ3.** Can we employ vision-based reconstruction models to create synthetic datasets with labeled emotion information in order to address the scarcity of large audio-4D datasets?
- **Sub RQ4.** Can vertex-based proposed models be extended to work on blendshape-based data?

2 BACKGROUND AND RELATED WORK

Related work on data-driven facial animation can be divided into two main categories- vision-based and speech-based where our focus is on the latter. There has been extensive work and research done in neural-rendering of talking head animations in 2D pixel space [Guo et al. 2021; Lu et al. 2021; Stypulkowski et al. 2023; Wu et al. 2021]. However, due to the limitations of rendered videos, which are not useful in 3D interactive applications, this research will address speech-driven facial animation in 3D space.

[Karras et al. 2017] proposed an end-to-end convolutional neural network approach that takes advantage of linear predictive coding to learn an encoding of audio or speech that can be used to disambiguate facial expression variations in a latent space. The authors trained their network using their in-house dataset captured with a traditional vision-based industry solution, DI4D [DI4D 2023]. Their approach can generalize to unseen and arbitrarily long audio sequences from any speaker, language, and emotion. However, the animation in the upper face region still suffers from generating realistic variations. [Taylor et al. 2017] proposed a deep learning approach to learn a one-to-one mapping between phonemes and visemes. However, it can be argued that a one-to-one mapping of phonemes and visemes can suffer from generating natural lip motions as a specific phoneme can have many representations in the lip motion. [Cudeiro et al. 2019] presented a generic speech-driven facial animation framework that is trained on 4D scans of 12 subjects uttering 40 sentences each, in which 255 unique sentences are present across the dataset. The authors present the VOCA model that takes audio and a template neutral mesh as input and outputs a facial animation sequence in accordance with the input audio. The authors take advantage of the pretrained DeepSpeech [Hannun et al. 2014] to extract audio features and use the FLAME [Li et al. 2017] head model to learn identity factors in the dataset. The proposed model generalizes to unseen audio, controllable via speaking styles, shape, and pose. However, VOCA fails to realistically synthesize upper face motion and argues that the upper facial region is weakly correlated with speech information. [Richard et al. 2021] presented a framework for speech-driven facial animation by learning a categorical latent space based on a novel cross-modality loss together with reconstruction and landmark losses that disentangles speech-correlated and uncorrelated information. During inference, a U-net-like decoder autoregressively generates facial animation on template face mesh. [Fan et al. 2021a] proposed a transformer-based approach to address speech-driven facial animation called FaceFormer. The authors proposed an autoregressive transformer architecture to solve a sequence modeling problem for speech-driven facial animation. The encoder leverages a self-supervised speech model, Wav2Vec 2.0 [Baevski et al. 2020], which is a pretrained speech model to address the scarcity of available data in existing audio-visual datasets. [Aylagas et al. 2022] proposed an audio-driven approach that incorporates retargeting to a rigged face and includes tongue animation, which has not been addressed by other works. In Learning2Listen [Ng et al. 2022], the authors proposed a network based on VQ-VAE to model the listener animation in a dyadic setting. More recently, [Xing et al. 2023] proposed a VQ-VAE based speech-driven autoregressive transformer animation generation model, CodeTalker, that is inspired by both FaceFormer and Learning2Listen.

To our knowledge, [Karras et al. 2017; Peng et al. 2023] addressed emotional expressiveness for audio-driven 3D facial animation synthesis task. Our goal is to explore and study the emotional expressiveness in speech-driven 3D facial animation synthesis in more detail and answer the research questions introduced in the previous section by proposing novel approaches for the synthesis task.

3 PROBLEM FORMULATION

The problem of generating output 3D facial animation driven by corresponding input data can be formalized as a sequence modeling problem where the input can be audio together with other modalities (e.g. emotion, text semantics, context, etc.) and the output is a temporally aligned sequence of 3D facial animation. Both deterministic and non-deterministic sequence-to-sequence modeling approaches can be formalized as follows -

Given an input data, X (can be combination of different modalities such as- speech, emotion, subject ID, etc.) and temporally input aligned output sequence $Y_T = (y_1, y_2, y_3, ..., y_T)$, where T is the total number of visual frames (can be both 4D mesh sequence or vector of parameterized blendshape data). The goal is to propose neural network approaches to learn the mapping between X and Y so that in inference time, the model can synthesize $\hat{Y}_T = (\hat{y}_1, \hat{y}_2, \hat{y}_3, ..., \hat{y}_T)$ from arbitrary unseen input data $\hat{X}$.

4 GENERAL APPROACH

The general approach throughout this research would be to address the Sub RQs one-by-one. For each Sub RQ, we will (i) explore and review the current state-of-the-art in data-driven facial animation as well as core deep learning techniques, (ii) propose novel method for data-driven models, and (iii) evaluate the method with respect to state-of-the-art methods and ground-truth.

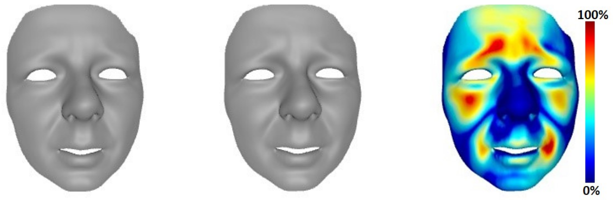
4.1 Datasets

In order to conduct experiments and answer the research questions presented earlier, we will use multiple audio-visual datasets. We will use 3D mesh-based datasets (i.e. BIWI[Fanelli et al. 2010], VOCASET[Cudeiro et al. 2019], and Multiface [Wuu et al. 2022]) and blendshape-based datasets (i.e. BEAT [Liu et al. 2022] and our in-house dataset, UUDaMM- Utrecht University Dyadic Multimodal Motion Dataset).

4.2 Sub RQ1

Sub RQ1 states - **"Can we disentangle emotional expressiveness from speech input to control the expressiveness of the generated animation?"**

To answer this, we need a dataset that comprises utterance sequences in a binary manner in terms of emotional expressiveness. BIWI is such a dataset where actors were asked to speak 40 sentences, firstly with a neutral expression and secondly with emotion. This results in a balanced and appropriately labeled dataset which is perfect for experiments related to this Sub RQ. We employ the state-of-the-art self-supervised HuBERT[Hsu et al. 2021] model to encode speech and semantics in a text-less manner to propose an end-to-end deterministic sequence modelling approach that we call FaceXHuBERT [Haque and Yumak 2023]. In this model, we incorporated an emotion embedding based on the binary emotion label to be able to distinguish the emotion-specific motion that resides mostly in the upper face region. Furthermore, we also found that due to the robustness and generalizable capabilities of HuBERT, using a simple GRU to decode facial animation would suffice, decreasing training time and complexity by a large margin. Our experiments demonstrate better results in terms of quantitative metrics compared to other methodologies. Additionally, our qualitative analysis shows how the network distinguishes the emotion-specific and speech uncorrelated motions residing in the face (see Fig. 1). The user study also demonstrates that emotionally expressive synthesized animations by our model are perceived as more realistic than the ones generated by other models. Furthermore, this work shows that given a strong audio-encoder and a balanced dataset with correctly labeled emotion information, we can generate controllable emotionally expressive facial animation that is perceived as realistic. Visual results of the proposed approach can be seen in the accompanying video. For further details, we refer to the FaceXHuBERT paper [Haque and Yumak 2023].



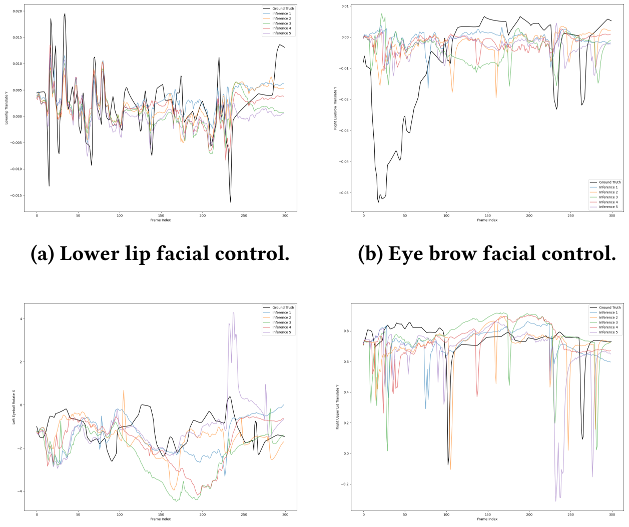
(a) Neutrally generated animation **(b) Expressively generated animation** **(c) Difference between (a) & (b)**

Figure 1: Emotional expressivity disentanglement using our approach, FaceXHuBERT[Haque and Yumak 2023] that can generate facial animations that are style controllable by a binary emotion label. Fig. 1a is generated with neutral expression whereas Fig. 1b is generated with emotional expressivity. Fig. 1c shows the colorized difference based on per-vertex distances between neutrally and expressively generated animations where extreme red depicts 100% of the computed distance and extreme blue depicts 0% of the computed distance. It is evident that the emotional expressivity signal effects the facial regions that are uncorrelated or loosely correlated with speech but correlated with emotional expressiveness.

4.3 Sub RQ2

Sub RQ2 states - **"Do non-deterministic generative models synthesize more perceived realism in facial animation compared to deterministic models?"**

Facial motion is a human action that naturally has certain non-deterministic aspects. This means that even during performance capture, no two takes for a single sequence would be exactly the same at the frame level and will have randomness based on a probabilistic distribution. With this in mind, Sub RQ2 experiments with non-deterministic techniques for generative models to synthesize facial animation. Due to the non-deterministic nature of such models, evaluation is mostly be based on subjective analysis and user studies, along with certain quantitative evaluation metrics. We employed diffusion technique similar to [Tevet et al. 2022] but for face and proposed FaceDiffuser [Stan et al. 2023]. The proposed approach generates non-deterministic facial animation sequences that are coherent with the input audio. Fig. 2 shows the non-deterministic capability of our approach. Furthermore, the perceptual user study also demonstrated that majority of the users preferred the animations generated by the non-deterministic approach to the deterministic ones. For more details, we refer to the FaceDiffuser paper[Stan et al. 2023].



(a) Lower lip facial control.

(b) Eye brow facial control.

(c) Eye ball rotation facial control.

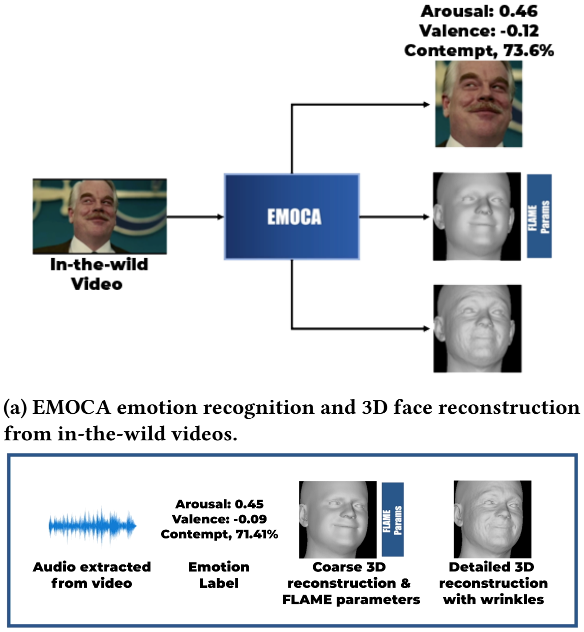
(d) Upper lid facial control.

Figure 2: Animation graphs of some facial controls (i.e.- lowerlip, eyebrow, gaze, upperlid) of the UUDaMM dataset synthesized using our approach, FaceDiffuser[Stan et al. 2023]. We synthesize animation data multiple times using the same audio and plot the graphs together with the ground truth (GT). The black plots depict GT whereas different colored plots depict different generations. It is evident that our approach produces lip control values similar to the ground truth as seen in Fig. 2a while encourages diversity for speech uncorrelated facial controls as seen in Fig. 2b, Fig. 2c and Fig. 2d.

4.4 Sub RQ3

Sub RQ3 states - **"Can we employ vision based reconstruction models to create synthetic datasets with labeled emotion information in order to address the scarcity of large audio-4D datasets?"**

Based on the findings from FaceXHuBERT [Haque and Yumak 2023], which addressed sub RQ1, we believe that we can propose a method to incorporate specific emotion information to generate emotionally expressive facial animation while having the freedom to control the emotional aspect during generation. The bottleneck for this experiment is the lack of appropriate datasets. However, vision-based 4D reconstruction models such as DECA [Feng et al. 2021] and EMOCA [Danecek et al. 2022] have gained traction in recent years for producing emotionally expressive 3D mesh sequences from videos. We have seen in [Ng et al. 2022; Peng et al. 2023], such vision-based models are used to create synthetic datasets using 2D videos. With EMOCA, we plan to employ a similar strategy to create our own synthetic dataset that will have labeled categories of emotion together with continuous valence and arousal information as depicted in Fig. 3.



(a) EMOCA emotion recognition and 3D face reconstruction from in-the-wild videos.

(b) Synthetic dataset creation process by taking the audio from the source video together with the reconstructed output of EMOCA.

Figure 3: Synthetic dataset creation process using EMOCA.

4.5 Sub RQ4

Sub RQ4 states - **"Can vertex-based proposed models be extended to work on blendshape-based data?"**

Finally, for Sub RQ4, we will extend the vertex-based approaches to work with blendshape-based data and evaluate the networks' performance quantitatively and in terms of perceived realism and synchrony. As manipulating blendshapes is real-time friendly and can be integrated with existing interactive 3D graphics applications, such a data-driven model can benefit online interactive applications involving digital humans.

5 EVALUATION

The proposed approaches will be evaluated extensively through quantitative, qualitative, and perceptual methods. While quantitative evaluation metrics will represent how well a network performs in producing facial animation that resembles the ground-truth data, qualitative and perceptual evaluation methods will provide insights on the visual realism and coherence of the synthesized animations. Due to the many-to-many mappings between speech and facial motion in both lower and upper regions of the face, it is recommended to conduct qualitative and perceptual evaluations to gain a more appropriate understanding of the performance of the proposed models, rather than relying solely on quantitative metrics [Cudeiro et al. 2019; Fan et al. 2021a; Karras et al. 2017]. Furthermore, ablation studies will also be conducted to evaluate the performance of the proposed models by leaving out key specific modules from the original approaches. The ablation study will provide key insights into the importance of specific modules in the final proposed architectures.

6 DISCUSSION AND CONCLUSION

Speech-driven 3D facial animation can be used in various ways including movie/game production, XR applications involving digital humans, dubbing 3D content to a different languages etc. NPCs in video games can be animated with ease by voice acting only. Additionally, manually editing animations for different languages (i.e. dubbed movies/game cut-scenes in languages other languages than the original version) is tedious, while such speech-driven animation synthesis models can be used to train on original speech-animation paired dataset and later be synthesized in multiple languages, reducing production complexity and increasing quality. However, we are still far away from having a production-ready and accessible data-driven model that can enhance the animation workflows. With this doctoral research, we plan to contribute towards realizing ideal data-driven 3D facial animation synthesis approaches that not only generates accurate lip-sync but also conveys natural non-verbal facial cues with coherent expressiveness.

ACKNOWLEDGMENTS

I would like to thank Dr. Zerrin Yumak and Prof. Dr. Judith Masthoff for unending support and trust to carry out this research. Many thanks to Stefan Stan for all the enticing discussions and heartfelt gratitude towards Dr. Rachel McDonnell for agreeing to assess my work as an external and for providing valuable insights.

REFERENCES

Mónica Villanueva Aylagas, Héctor Anadon Leon, Mattias Teye, and Konrad Tollmar. 2022. Voice2Face: Audio-driven Facial and Tongue Rig Animations with cVAEs. In *EUROGRAPHICS SYMPOSIUM ON COMPUTER ANIMATION (SCA 2022)*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *CoRR* abs/2006.11477 (2020). arXiv:2006.11477 https://arxiv.org/abs/2006.11477

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. 2019. Capture, Learning, and Synthesis of 3D Speaking Styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10101–10111. http://voca.is.tue.mpg.de/

Radek Danecek, Michael J. Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 20311–20322.

DI4D 2023. DI4D. https://di4d.com/.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2021a. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. https://doi.org/10.48550/ARXIV.2112.05329

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2021b. Joint Audio-Text Model for Expressive Speech-Driven 3D Facial Animation. https://doi.org/10.48550/ARXIV.2112.02214

G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. 2010. A 3-D Audio-Visual Corpus of Affective Communication. *IEEE Transactions on Multimedia* 12, 6 (October 2010), 591 – 598.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. 2021. Learning an Animatable Detailed 3D Face Model from In-The-Wild Images. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 40, 8. https://doi.org/10.1145/3450626.3459936

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis. *CoRR* abs/2103.11078 (2021). arXiv:2103.11078 https://arxiv.org/abs/2103.11078

Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep Speech: Scaling up end-to-end speech recognition. *CoRR* abs/1412.5567 (2014). arXiv:1412.5567 http://arxiv.org/abs/1412.5567

Kazi Injamamul Haque and Zerrin Yumak. 2023. FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23)* (Paris, France). ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614157

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. arXiv:2106.07447 [cs.CL]

Li Hu, Jinwei Qi, Bang Zhang, Pan Pan, and Yinghui Xu. 2021. *Text-Driven 3D Avatar Animation with Emotional and Expressive Behaviors.* Association for Computing Machinery, New York, NY, USA, 2816–2818. https://doi.org/10.1145/3474085.3478569

JALI 2023. JALI Research. https://jaliresearch.com/.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion. *ACM Trans. Graph.* 36, 4, Article 94 (jul 2017), 12 pages. https://doi.org/10.1145/3072959.3073658

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics* 36, 6 (Nov. 2017), 1–17. https://doi.org/10.1145/3130800.3130813

Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. 2022. BEAT: A Large-Scale Semantic and Emotional Multi-Modal Dataset for Conversational Gestures Synthesis. (03 2022).

Yuanxun Lu, Jinxiang Chai, and Xun Cao. 2021. Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation. *ACM Transactions on Graphics* 40, 6 (2021), 17 pages. https://doi.org/10.1145/3478513.3480484

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. 2022. Learning To Listen: Modeling Non-Deterministic Dyadic Facial Motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20395–20405.

S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff. 2023. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. *Computer Graphics Forum* 42, 2 (May 2023), 569–596. https://doi.org/10.1111/cgf.14776

Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. 2023. EmoTalk: Speech-driven emotional disentanglement for 3D face animation. arXiv:2303.11089 [cs.CV]

Alexander Richard, Michael Zollhoefer, Yandong Wen, Fernando de la Torre, and Yaser Sheikh. 2021. MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement. https://doi.org/10.48550/ARXIV.2104.08223

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. 2019. Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Stefan Stan, Kazi Injamamul Haque, and Zerrin Yumak. 2023. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion. In *ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23), November 15–17, 2023, Rennes, France* (Rennes, France). ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3623264.3624447

Michal Stypulkowski, Konstantinos Vougioukas, Sen He, Maciej Zieba, Stavros Petridis, and Maja Pantic. 2023. Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation. *CoRR* abs/2301.03396 (2023). https://doi.org/10.48550/arXiv.2301.03396 arXiv:2301.03396

Sarah Taylor, Taehwan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica Hodgins, and Iain Matthews. 2017. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics* 36, 4 (July 2017), 1–11. https://doi.org/10.1145/3072959.3073699

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2022. Human motion diffusion model. *arXiv preprint arXiv:2209.14916* (2022).

Y.-I. Tian, T. Kanade, and J.F. Cohn. 2001. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 2 (2001), 97–115. https://doi.org/10.1109/34.908962

Stef van der Struijk, Hung-Hsuan Huang, Maryam Sadat Mirzaei, and Toyoaki Nishida. 2018. FACSvatar: An Open Source Modular Framework for Real-Time FACS Based Facial Animation *(IVA '18)*. Association for Computing Machinery, New York, NY, USA, 159–164. https://doi.org/10.1145/3267851.3267918

Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. 2021. Imitating Arbitrary Talking Style for Realistic Audio-DrivenTalking Face Synthesis. *CoRR* abs/2111.00203 (2021). arXiv:2111.00203 https://arxiv.org/abs/2111.00203

Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. 2022. Multiface: A Dataset for Neural Face Rendering. In *arXiv*. https://doi.org/10.48550/ARXIV.2207.11243

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. 2023. CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.