FaceXHuBERT: Text-less Speech-driven E(X)pressive 3D Facial Animation Synthesis Using Self-Supervised Speech Representation Learning

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

Zerrin Yumak
Utrecht University
Utrecht, The Netherlands
z.yumak@uu.nl

ABSTRACT

This paper presents FaceXHuBERT, a text-less speech-driven 3D facial animation generation method that generates facial cues driven by an emotional expressiveness condition. In addition, it can handle audio recorded in a variety of situations (e.g. background noise, multiple people speaking). Recent approaches employ end-to-end deep learning taking into account both audio and text as input to generate 3D facial animation. However, scarcity of publicly available expressive audio-3D facial animation datasets poses a major bottleneck. The resulting animations still have issues regarding accurate lip-syncing, emotional expressivity, person-specific facial cues and generalizability. In this work, we first achieve better results than state-of-the-art on the speech-driven 3D facial animation generation task by effectively employing the self-supervised pretrained HuBERT speech model that allows to incorporate both lexical and non-lexical information in the audio without using a large lexicon. Second, we incorporate emotional expressiveness modality by guiding the network with a binary emotion condition. We carried out extensive objective and subjective evaluations in comparison to ground-truth and state-of-the-art. A perceptual user study demonstrates that expressively generated facial animations using our approach are indeed perceived more realistic and are preferred over the non-expressive ones. In addition, we show that having a strong audio encoder alone eliminates the need of a complex decoder for the network architecture, reducing the network complexity and training time significantly. We provide the code[1] publicly and recommend watching the video.

CCS CONCEPTS

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

KEYWORDS

facial animation synthesis, deep learning, digital humans

[1]https://github.com/galib360/FaceXHuBERT

ICMI '23, October 09–13, 2023, Paris, France

ACM ISBN 979-8-4007-0055-2/23/10…$15.00
https://doi.org/10.1145/3577190.3614157

ACM Reference Format:
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), October 09–13, 2023, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3577190.3614157

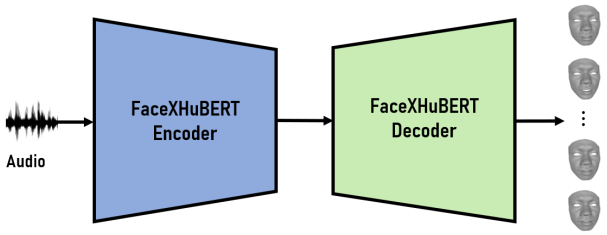


Figure 1: FaceXHuBERT: An end-to-end encoder-decoder architecture that encodes audios using self-supervised pretrained speech model HuBERT and decodes to vertex displacements using GRU followed by a fully connected linear layer that produces 3D facial animation as 3D mesh sequences.

1 INTRODUCTION

Speech-driven 3D facial animation is a growing yet challenging research area with applications to games, VR/AR and film production. Conversational virtual humans with social and emotional interaction capabilities are used in a range of applications such as chatbots for customer service and marketing, simulations for education and healthcare and remote communication [7, 43, 60]. Facial expressions are the first point of attention in conversational communication and humans are very receptive to subtle nuances in facial animation which is explained by the uncanny valley theory [41].

Typically, facial animation workflows rely on professional technical artists using blendshape facial animation [36] or performance capture aiming to mitigate most of the labor intensive work [17, 18, 20]. However, as these characters take place in interactive applications, demand to automatically generate their behavior on-the-fly increases. Research on facial animation focuses on 2D talking faces [32, 38, 50, 66], 3D facial animation constructed from 2D images and videos [15, 24, 39, 70] and 3D speech-driven facial animation [1, 14, 22, 34, 55, 61, 72]. In this paper, we propose a novel approach for emotionally expressive speech-driven 3D facial animation.

3D speech-driven facial animation is either based on phoneme-based approaches using procedural algorithms [11, 31] or data-driven approaches using machine learning [62], motion graphs [9] and deep learning [61, 72]. The former requires explicit definition of co-articulation rules and requires manual work. While the latter aims to eliminate that by learning speech-animation parameters mapping from data, it still relies on intermediary representations of speech units. Recent approaches on 3D speech animation synthesis effectively employ end-to-end deep learning models [14, 22, 34, 55, 72] eliminating the need for intermediary representations. Large speech and language models [6, 26] open the way towards more realistic speech-driven facial animation. However, the lack of 3D facial animation data matching audio and text poses a major bottleneck and current models cannot generalize to arbitrary speech input. More recently, Fan et al. [22] suggested a self-supervised speech representation learning method using transformers to mitigate these issues. However, the method cannot handle expressive animations. Xing et al. [68] later proposed a two-step training approach by learning a codebook using VQ-VAE. Both these methods are computationally very expensive to train.

Our work improves on these works and proposes FaceXHuBERT, a text-less speech-driven expressive 3D facial animation generation method using self-supervised speech representation learning. In our proposed encoder-decoder network (see Fig. 1), we effectively employ self-supervised pretrained HuBERT model to incorporate and encode both lexical and non-lexical information without using a large lexicon and speech-3D data pairs allowing it to generalize to any speech input. The main contributions of our work are:

- **A fast and efficient text-less speech-driven expressive 3D facial animation method using self-supervised speech representation learning**. FaceXHuBERT produces expressive and realistic animations in an efficient way using a HuBERT-based encoder and GRU-based decoder without the use of a large lexicon and using only audio input. Additionally, guiding the training with an emotion condition and speaker identity distinguishes the tiniest subtle facial motions related to emotional expressiveness. The results show that our method produces more realistic results in a more efficient manner (i.e. almost 3 times faster in comparison to state-of-the-art [22]).
- **Demonstration of self-supervised text-less speech model HuBERT [26] incorporated for the downstream task of expressive 3D facial animation synthesis**. Our method produces accurate lip-sync as well as allows to capture personalized and subtle cues in speech (e.g. identity, emotion and hesitation). It can handle audio recorded in a variety of situations (e.g. multiple people speaking, background noise, laughter, lip-smacking).
- **Extensive objective and subjective analyses**. We compared our method to state-of-the-art methods and ground-truth as well as made comparisons between various sequence modelling network architectures using 3D face vertex errors as objective metrics. Subjective analysis includes qualitative generalizability analysis in terms of different languages, text-to-speech, noise, low-quality audio input and single subject training. We also conducted perceptual user studies. Our results demonstrate that our approach produces superior results with respect to the expressiveness of the facial animation 78% of the time in comparison to the state-of-the-art [22].

2 RELATED WORK

Extensive research has been conducted in the domain of automatic facial expression analysis and synthesis in the 2D pixel domain for the purpose of detecting expressions [30, 57, 63], for generating audio-driven talking faces [25, 32, 38, 58, 66, 71] or for video-based facial re-enactment/face swapping [47, 50, 59].

The approaches for 3D facial animation synthesis can be classified into video-driven and audio-driven facial animation. While the former focuses on transferring facial animation from 2D videos to 3D faces, the latter maps speech (audio and text) to 3D facial animation parameters. Earlier works on video-driven facial animation focused on optimization-based 3D facial performance capture [4, 8, 28], while recent works use deep learning [15, 24, 39, 42, 56]. For an extensive survey on 3D face reconstruction, tracking and morphable models, we refer to [19, 40, 73]. Some methods use retargeting algorithms to convert facial expressions from one 3D mesh to the other [10, 54] or from 2D images to 3D faces [69, 70]. Finally, there is a group of research focusing on physics-based animation of faces [3, 29]. In our work, we focus on 3D speech-driven emotionally expressive facial animation using deep learning.

3D Speech-Driven Facial Animation. 3D speech-driven facial animation typically uses phoneme-based procedural approaches [11, 31]. Although these methods come with the advantage of animation control and easy integration to artist-friendly pipelines, they are not fully automatic and require defining explicit rules for co-articulation. Another line of research uses machine learning [62] or graph-based approaches [9] to learn speech-animation mappings from data. These methods rely on blending between speech units and cannot capture the complexity of the dynamics of visual speech [61]. They rather focus on the lower face and are not robust to emotion and style variations. Recent approaches on 3D speech animation synthesis effectively employ deep learning models [1, 14, 22, 34, 55, 61, 72]. Taylor et al.[61] proposes a sliding window approach instead of an RNN focusing on capturing neighborhoods of context and coarticulation effects. VisemeNet [72] builds upon the viseme-based JALI[31] model and combines this with an LSTM-based neural network. However, these two methods [61, 72] still rely on intermediary representations of phonemes and they focus on the mouth movement. Most previous works do not include automatic tongue animation except [1]. Collecting large-scale datasets using professional performance capture workflows is expensive and time consuming but the resulting faces are highly realistic. To elevate this disadvantage, some methods use 3D face reconstruction methods from in-the-wild videos, which are especially useful in situations where professional performance capture systems are not available, e.g. dyadic speech-driven facial animation [33, 46]. However, these methods are prone to 3D reconstruction errors and cannot generate results that are as realistic as the former.

Closest to our work are Karras et al.[34], Cudeiro et al.[14], Richard et al.[55], Fan et al.[21, 22] and Xing et al.[68]. Karras et al.[34] proposes an end-to-end convolutional neural network that learns a mapping from input waveforms to the 3D vertex coordinates of a face model. They aim to resolve the ambiguity in mapping between audio and face by introducing an additional emotion component to the network, which is learned from data. However, the method is not trained on multiple speakers and cannot handle

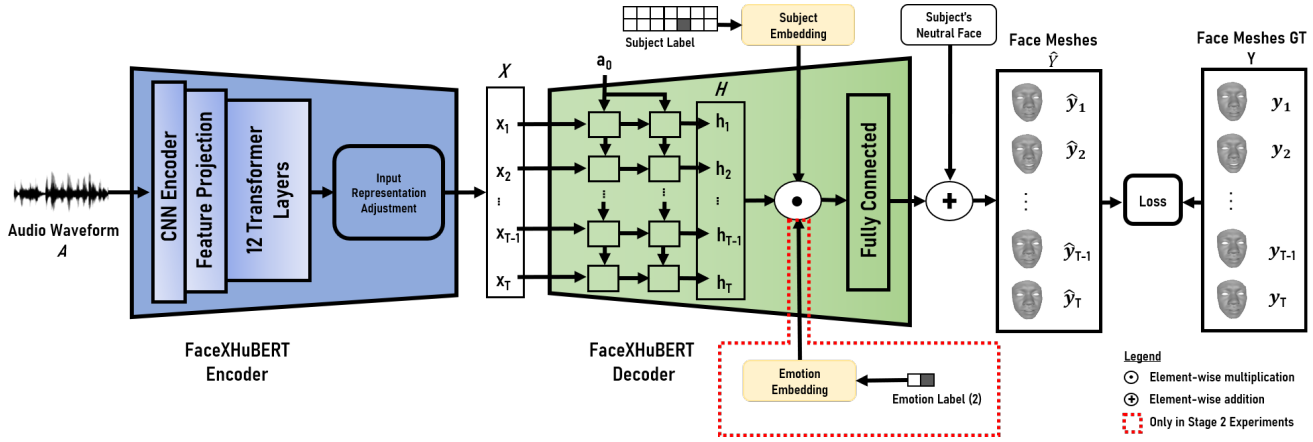


Figure 2: FaceXHuBERT: The encoder encodes the audio waveform A and produces discrete frame level embedding. The Input Representation Adjustment module in the encoder adjusts the encoded information with the output 4D scan data and produces X such that $T_X = T_Y = T$. The Decoder takes in X and with the help of the 2-layered 256 hidden sized GRU, produces the hidden representation H. Additional conditions such as the speaker identity and emotion label are embedded and multiplied with H before the hidden representation is decoded into vertex displacement values and added to the corresponding subject's neutral face to produce the animation output $\hat{Y}$. The loss function is computed based on $\hat{Y}$ and ground-truth (GT), Y.

identity variations and requires a longer-term audio context to infer the emotional state. Instead, Cudeiro et al. [14] presents the audio-driven facial animation method VOCA that generalizes to new speakers using a training dataset with 12 subjects eliminating the need for retargeting. However, VOCA fails to realistically synthesize upper face motion and does not include emotional variations. Similar to VOCA, Richard et al.[55] aims for audio-driven animation that can capture variations in multiple speakers including a much larger dataset of 250 subjects. They address the problem of lack of upper face motions using a categorical latent space that disentangles audio-correlated and audio-uncorrelated information based on a cross-modality loss. Fan et al.[21] proposes an audio and text-driven facial animation method that incorporates the large language model GPT-2[53] to encode the textual information. The authors found that combined audio and text input yielded better results than audio-only or text-only model. However, the results still have problems regarding accurate lip-sync. Most closely related to our work is FaceFormer[22] which uses a self-supervised pretrained speech model that addresses the scarcity of available data in existing audio-visual datasets. The model produces superior results in comparison to Cudeiro et al. [14] and Richard et al.[55] using a modified version of transformers to handle longer sequences of data. However, none of these methods can handle arbitrary variations in speech input while producing accurate lower and upper facial animation for multiple identities and emotions. Most recently, CodeTalker[68] incorporates self-supervised Wav2Vec 2.0 inspired by FaceFormer[22] together with the idea of having a latent codebook using VQ-VAE inspired by Learning2Listen[46] to generated speech-driven facial animation. However, control of emotional expressiveness is not addressed while the training process is subject to computational complexity and done in two steps.

3 PROBLEM FORMULATION

We formalize the task of audio-driven 3D facial animation as a generic sequence modeling (seq2seq) problem in which the input sequence is a raw audio waveform whereas the output sequence is a 3D face mesh sequence (i.e. 4D scan). Hence, the problem can be formalized as follows:

Given audio A and ground-truth 3D mesh sequence Y= $(y_1, y_2, y_3, ..., y_{T_Y})$, T_Y is the total number of available visual frames or 3D scanned frames in the sequence. Therefore, one sequence of Y is a (T_Y, V) dimensional matrix where V denotes the number of 3D vertices in the mesh topology. On the input side, since audio stream A is a continuous data stream, with the help of an encoder, we encode the continuous audio into a discrete representation $X = (x_1, x_2, x_3, ..., x_{T_X})$ where X is a (T_X, B) dimensional matrix and T_X and B are the discrete time-steps and the encoded representation respectively.

The goal is to train an end-to-end architecture to learn the mapping between A (together with additional modalities) and Y to generate $\hat{Y} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, ..., \hat{y}_{T_y})$ so that the $\hat{Y}$ best approximates Y.

4 PROPOSED APPROACH

We present FaceXHuBERT, an end-to-end encoder-decoder neural network architecture. Our model uses the pretrained HuBERT speech model as the audio encoder while for the decoder, we use Gated Recurrent Unit (GRU) [12]. Fig. 2 shows the overall architecture of our proposed approach. The encoder encodes the continuous audio information into discrete time-step representations and adjusts the representations so that the time-steps match with that of the face scan data. The decoder that incorporates the style labels, consists of a 2-layered GRU with hidden size 256 followed by a fully connected linear layer. We define style in terms of the emotional expressivity present implicitly in the facial motion data of each training subject. The decoder regresses vertex displacements and

adds to the subject's template mesh to generate the predicted mesh sequence. Algorithm 1 in the supplementary material depicts the overall steps of the proposed network. In the next two subsections, we describe the details of the FaceXHuBERT Encoder and Decoder.

4.1 FaceXHuBERT Encoder

Our proposed method effectively adopts the state-of-the-art self-supervised pretrained speech model HuBERT in the encoder for the downstream task of 3D facial animation generation. Since it is able to learn and produce high quality discrete hidden representations of continuous audio streams combining both acoustic and language information, the authors of HuBERT recommend to consider using HuBERT pretrained representations for a variety of downstream tasks [26]. HuBERT architecture introduces a BERT-like [16] masked language modelling encoder for the transformer layers. It introduces a simple cross-entropy loss for predicting masked units in contrast to its predecessor Wave2Vec 2.0's [2] complex contrastive loss. In addition, unlike Wave2Vec 2.0, HuBERT is trained with multiple iterations. During the first iteration, HuBERT uses unsupervised simple k-means clustering for acoustic unit discovery to facilitate the self-supervised masked language modeling learning that takes place in the second iteration. In the second iteration, the training is done on the discovered discrete hidden units with a predictive loss on the masked regions only, forcing the model to learn a combined acoustic and language model using a BERT-like encoder, hence the name H(idden)-u(nit)-BERT. The model is trained on 960 hours of unlabeled speech data [48] which contains English recordings of copyright-free audiobooks by volunteers from the internet. The authors claim that this approach is the first big step towards text-less Natural Language Processing (NLP). For detailed explanation of how different HuBERT models were trained and for comparison to previous work, we refer to the original paper [26].

The FaceXHuBERT Encoder is composed of a CNN encoder that discretizes the continuous audio data into 512 dimensional representations. Feature Projection layer projects the 512 dimensional representation into 768 dimensional representation, a positional convolution embedding layer and 12 transformer layers to capture the contextual information in the sequence. In our approach, we adopt the "base" HuBERT model with 95M parameters which produces 768 dimensional embedding at the last hidden state. We initialize the pretrained weights and freeze the model parameters including the CNN feature encoder layer, feature projection layer and the first two transformer layers. The last ten transformer layers are kept unfrozen and remain trainable. HuBERT generates a feature sequence in 20ms windows (i.e. 50 fps). Therefore, in our architecture, we encode a one second of audio into 50 frames with 768 dimensional embeddings. For example, for a training data with 4 seconds of audio stream, the output from the encoder will be $(4 \times 50, 768) = (200, 768)$ dimensional matrix.

Input Representation Adjustment. This module adjusts the input representation and output representations and do not contain any trainable parameters. This is devised to ensure the one-to-one frame level relationship between decoder input X and output Y such that $T_X = T_Y = T$. This function is generically devised in such a way that it can handle any input-output frequency pair. More details on this can be found in the supplementary material.

4.2 FaceXHuBERT Decoder

For the FaceXHuBERT Decoder, we use a Gated Recurrent Unit (GRU) instead of a complex transformer model. Our extensive analysis shows that, combined with the HuBERT encoder, GRU-based decoder produces realistic results in a more efficient manner. Our GRU-based decoder consists of 2 layers with hidden unit size of 256, followed by one fully connected linear layer that maps the last hidden state to vertex displacement values of the 3D vertices of the face. It represents the faces in terms of their displacement values with respect to the neutral template vertices of a given subject. Between the GRU and the fully connected layer, we add the additional conditions and fuse them with the hidden state representation with element-wise multiplication. The additional conditions are (i) subject identity and (ii) emotion (neutral or expressive in our experiments). We defined the training subjects and the emotion label as one-hot vectors and linearly embed them with two separate 256 dimensional vectors to facilitate the element-wise multiplications.

Eq. (1), Eq. (2), Eq. (3) and Eq. (4) show the core computations of our decoder's forward propagation.

$$\Gamma_r = \sigma(W_r^l[a_{t-1}^l, a_t^l] + b_r^l) \tag{1}$$

$$\tilde{a}_t^l = tanh(W_a^l[\Gamma_r \odot a_{t-1}^l, a_t^l] + b_a^l) \tag{2}$$

$$\Gamma_u = \sigma(W_u^l[a_{t-1}^l, a_t^l] + b_u^l) \tag{3}$$

$$a_t^l = \Gamma_u \odot \tilde{a}_t^l + (1 - \Gamma_u) \odot a_{t-1}^l \tag{4}$$

The subscript t denotes the frame number or time-step in the sequence whereas the superscript l denotes the hidden layer where $l = [1, 2]$ (i.e. two hidden layers). When $l = 1$, the activation values a_t^l (not a_{t-1}^l) in Eq. (1), Eq. (2) and Eq. (3) take the input value x_t, hence, $a_t^1 = x_t$. When $l = 2$, the activation values $a_t^2 = h_t$, each being 256 dimensional hidden unit. The second GRU layer produces the hidden units $H = [h_1, h_2, h_3, ..., h_T]$ for a sequence with T frames. Furthermore, we initialize $a_0 = \vec{0}$ to start the training.

Eq. (5), Eq. (6), Eq. (7) and Eq. (8) show how the subject identity and emotion conditions are incorporated into the network.

$$S = W_S \cdot [SubjectOneHot] + b_S \tag{5}$$

$$E = W_E \cdot [EmotionOneHot] + b_E \tag{6}$$

$$\tilde{H} = H \odot S \odot E \tag{7}$$

$$\hat{Y} = (W_{\hat{Y}} \cdot \tilde{H} + b_{\hat{Y}}) \oplus [NeutralFace] \tag{8}$$

The subject label and the emotion label are represented as one-hot vectors that we linearly embed to 256 dimensional S and E vectors using Eq. (5) and Eq. (6) respectively. In Eq. (7), the hidden representation H with dimensions $(T, 256)$ is multiplied in an element-wise manner with both the embedding vectors of the given subject, S (i.e. training subjects) and the expressiveness of the given sequence, E (i.e. neutral or expressive). Finally, in Eq. (8), the fully connected linear layer decodes the hidden representation into (T, V) dimensional vertex displacement values which is then added to the respective subject's neutral face vertex values, where $[NeutralFace]$ is a $(1, V)$ dimensional data.

5 EXPERIMENTS

The experiments are setup in two stages. In the first stage, we solve the problem of generating speech-driven 3D facial animation and achieve better results than the state-of-the-art methods. In the second stage, we take the best method from the first stage based on the objective metrics and incorporate the additional emotion modality in the neural network architecture in order to address emotion controllable animation generation. In this section, we describe the datasets along with the data-splits we used in both the stages and explain the two stages of experiments.

5.1 Dataset

For our experiments, we used the BIWI[23] and VOCASET[44] datasets. Both the datasets are suitable for the experiments in the first stage (i.e. speech-driven animation), whereas only BIWI is suitable for the second stage (i.e. speech-driven emotionally expressive animation) as BIWI contains both identity and emotion labels. Furthermore, VOCASET[44] and Multiface[67] datasets allow incorporating identities but not emotions.

BIWI contains synchronized audio-4D scan pairs of 14 human subjects uttering 40 phonetically balanced English sentences twice: first neutrally, second with emotional expressions. Therefore, the dataset contains $(14 \times 40 \times 2) = 1120$ sequences of audio-4D scan pairs. In reality, due to some missing sequences from 4D scans, there are in total 1088 audio-4D pairs. The sequences in the dataset are 4.39 seconds in duration on average including both neutral and emotional sequences. To ensure efficient training, we pre-process the whole dataset by scaling the 3D vertices to have a uniform range of values for all three coordinates across the dataset (e.g. [-0.5,0.5]). More details on how we preprocessed the dataset can be found in the supplementary material document. The data is captured at 25fps with 23370 3D vertices in the mesh topology. VOCASET contains 480 audio-4D pairs of sequences, each lasting about 4 seconds, captured from 12 human actors. The scans are captured at 60fps and registered with FLAME[37] model comprising of 5023 3D vertices.

During the first stage, we use both BIWI and VOCASET. In order to ensure fair comparison with respect to the state-of-the-art methods, we follow the exact splits that VOCA[14] (for VOCASET), FaceFormer [22] and CodeTalker[68] (for both VOCASET and BIWI) used in their respective works. Furthermore, during the first stage, we use the data preprocessing workflow provided in the official code repository of [68] to preprocess the subset (i.e. only the emotional sequences) of the BIWI dataset used by the previous works to ensure fair comparisons in terms of objective metrics. We adopt the same training (VOCA-Train), validation (VOCA-Val), and testing (VOCA-Test) splits as VOCA, FaceFormer and CodeTalker for VOCASET. On the other hand, for BIWI dataset, following the data split in FaceFormer and CodeTalker for fair comparisons, we only take the emotional subset of the whole dataset and train on 6 subjects. More specifically, the training split (BIWI-Train) has 192 sequences, the validation split (BIWI-Val) contains 24 sequences. We then have two testing splits, namely, BIWI-Test-A (includes 24 sequences by six training subjects) and BIWI-Test-B containing 32 sequences spoken by the remaining eight unseen subjects. BIWI-Test-A is used for both quantitative and qualitative evaluation due to the seen subjects during training, while BIWI-Test-B is used for perceptual user study.

Method	Full Face Vertex Error ($\times 10^{-4}$)mm	Lip Vertex Error ($\times 10^{-4}$)mm	FDD ($\times 10^{-5}$) mm
VOCA	6.73	5.86	6.86
FaceFormer	6.24	5.10	4.55
CodeTalker	6.01	4.79	4.11
HuBERT-LSTM	5.95	4.79	4.65
HuBERT-Transformer[1]	14.985	14.15	9.73
HuBERT-Transformer[2]	5.88	4.84	4.62
HuBERT-FaceFormer	5.85	4.77	4.77
HuBERT-RNN	5.84	4.66	**3.51**
HuBERT-GRU (selected)	**5.72**	**4.56**	4.96

[1] Teacher-forcing scheme. Produces static animations.
[2] Autoregressive scheme.

Table 1: Stage 1 Objective Evaluation: This table reports the comparisons of the objective evaluation metrics for stage 1 experiments. Our proposed approach with GRU in the decoder yielded the best results in terms of both full face vertex error and lip vertex error. Hence, we select the HuBERT-GRU variant for stage 2 experiment.

For the second stage, we take the best method yielded from the first stage and incorporate the emotion modality in training the proposed neural network. To achieve this, we use the whole BIWI dataset (including neutral and emotional sequences) and train on all 14 subjects. In this stage, we split the dataset in 90% train, 5% validation and 5% test sets. This split results in 977 training, 56 validation and 55 test sequences across all training subjects. We do not use VOCASET for the stage 2 experiments. VOCASET lacks emotional expressive sequences and upper face motion is absent, hence not an apt dataset to address emotion controllable animation.

5.2 Stage 1: Speech-driven Animation

For the first stage of experiments, we implement our proposed network (without the additional emotion modality depicted in dashed red box in Fig. 2) with varying decoders and compare the generated results with three state-of-the-art speech driven 3D facial animation methods: VOCA[14], FaceFormer[22] and CodeTalker[68]. First, for VOCA, we train on BIWI dataset and generate test animations using the official implementation while for VOCASET, we used the officially released pretrained model trained on VOCASET to generate test sequences. Second, we train FaceFormer on both the dataset and generate test sequences for comparisons. Lastly for CodeTalker, we used their officially released models for BIWI and VOCASET to generate the test sequences. For our proposed approach, we kept HuBERT as a constant for our audio encoder while we experimented with varying decoder structures (e.g. RNN, LSTM, GRU and Transformer) in order to find which variant yields the best result in terms of objective metrics. Furthermore, we also conduct a perceptual user study during this stage to compare our speech-driven synthesis method against the mentioned state-of-the-art methods.

5.2.1 Objective and Perceptual Evaluation. Evaluation of the 3D facial animation generation is challenging and there is no unanimously accepted single objective metric in the literature. Following FaceFormer and Codetalker, we employ the lip vertex error, which is the only proposed objective metric proposed in literature for

Competitors	BIWI-Test-B		VOCA-Test	
	Lip Sync	Realism	Lip Sync	Realism
Ours vs. VOCA	74.50	77.77	63.14	62.55
Ours vs. FaceFormer	56.45	51.79	53.82	47.92
Ours vs. CodeTalker	50.53	50.79	44.90	40.53
Ours vs. GT	31.53	37.01	30.06	32.67

Table 2: Stage 1 User Study Results: We adopt A/B testing and report the percentage of responses where A (i.e. Ours) is preferred over B (i.e. competitor).

speech-driven 3D facial animation in order to measure the lip synchronization of the generated animations. It measures the deviation of all the lip vertices (i.e. vertices belonging to the lip and mouth region) of a synthesized sequence with respect to the ground truth sequence by computing the maximal L2 error for each corresponding frame and by taking the mean across all frames. In addition, we also compute the full face vertex error as we are also interested in accurately generating rest of the face region resembling the ground truth data in stage 2 experiments because the emotional expressiveness resides in upper face and cheek regions of the face. The calculation of full face vertex error is done in the same way as lip vertex error calculation but unlike lip vertex error which computes the error on only lip vertices, the full face vertex error takes into account all the vertices in the face mesh topology. Moreover, we also calculate the FDD metric introduces recently by the authors of CodeTalker[68]. FDD depicts the upper face dynamics deviation, which reports if the observed variation of upper face vertices in generated animations are similar to that of ground truth. It does not necessarily report on the accuracy of the generated animations. Therefore, we select the best model to be incorporated in the next stage of experiments based on the full face and lip vertex errors. Furthermore, since facial animation is perceptual, we conducted user studies to compare our approach with ground truth as well as the state-of-the-art methods.

Tab. 1 reports the comparisons of the objective metrics of stage 1 experiments. For all three objective metrics, lower value is better. HuBERT-RNN and HuBERT-GRU model variants both achieve better results than all the state-of-the-art as well as all other variants in the table. However, since HuBERT-GRU produces the lowest error for both full face and lip vertex errors, we select it for the stage 2 experiments. Furthermore, Tab. 1 also shows having a strong and sophisticated audio encoder (i.e. HuBERT) alone eliminates the need of a complex vertex decoder and can efficiently yield better results.

In addition to objective evaluation, we conduct a perceptual user study to compare our generated animations against VOCA, FaceFormer, CodeTalker and ground truth with A/B testing in terms of lip-sync and realism. We adopt the user study design exactly like in [68] for fair comparisons. In our case, 52 participants took the study. Tab. 2 reports the user study results during stage 1 experiments. Ours is preferred over VOCA on both the datasets. For comparisons against FaceFormer and CodeTalker, the participants preferred ours and the competitors similarly for BIWI-Test-B comparisons while had difficulties choosing one over the other with large standard deviations on VOCA-Test. Therefore, we can conclude that our method generates 3D facial animations that are on-par with the most recent methods. The user study was hosted on Qualtrics [52] and carried out using Prolific [51], ensuring that the participants get compensated appropriately.

Model	Full Face Vertex Error ($\times 10^{-3}$)mm	Lip Vertex Error ($\times 10^{-3}$)mm	Training Time (h)
FaceFormer	3.75	3.28	$\approx$ 16.11
FaceXHuBERT	**1.72**	**1.51**	$\approx$ **5.10**

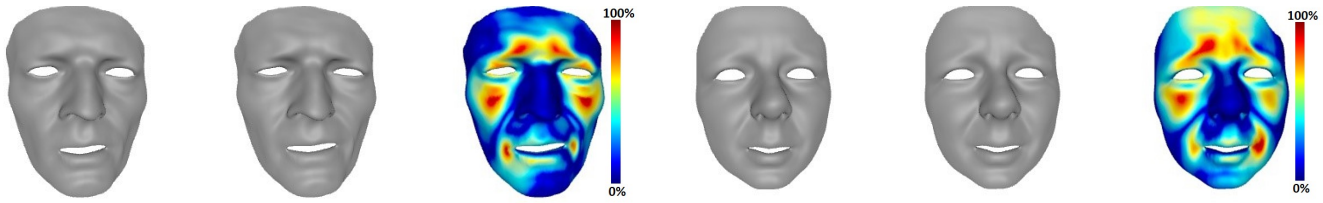
Table 3: Objective evaluation results of the experiments and trained models. Our approach not only produces the minimum Mean Face Vertex Error on test-set sequences but also reduces the training time significantly. FaceXHuBERT is more than 4 times faster than transformer based architecture and almost 3 times faster than the state-of-the-art.

5.3 Stage 2: Emotion Controllable Animation

Based on the results obtained from the first stage of experiments, we select the best performing model (i.e. HuBERT-GRU in Tab. 1) and incorporate the additional emotion modality in the architecture to propose the final model, FaceXHuBERT. In order to do so, we use the whole BIWI dataset including the neutral and the emotional sequences and train on all 14 subjects.

5.3.1 Quantitative and Qualitative Evaluation. We measure and compare our proposed methodology quantitatively based on full face and lip vertex errors. We train FaceFormer and our approach with the same setting and report the comparison in Tab. 3 that reports the error together with corresponding training time. FaceXHuBERT yields the lower full face vertex error while reducing the network complexity and training time significantly. It is to be noted that during stage 2 training, we used our preprocessed dataset described in Sec. 5.1 that has a different data distribution than the preprocessed subset dataset used in the first stage of experiments that was used for fair comparisons with other speech-driven facial animation methods' results.

The quality of animations generated by FaceXHuBERT has been studied carefully to understand the generalizability capabilities of the model in terms of various aspects such as- smoothness of the animation, coherence with respect to speech, different background noises in input audio, multiple speakers, different languages, different subjects, use of TTS (text2speech) instead of real audio and limited training data. The proposed approach is generalizable in terms of all the above-mentioned aspects. Additionally, by training the network without the binary emotion label (FaceXHuBERT- w/o emo) in the decoder, we lose the expressive style control capability during inference. In this case, facial expressiveness of the generated animations rely solely on the audio signal and are qualitatively slightly less expressive than the proposed model's predictions. Yet it still distinguishes between neutral and emotional aspects in speech signals. Furthermore, we qualitatively compared our approach to the FaceFormer (see Fig. 4) and found that FaceXHuBERT generates more expressive animations that are closer to the ground-truth. Unlike our approach, we observed that FaceFormer fails to distinguish variety of noises overlapping with the audio signal and produces visual artefacts. We recommend watching the supplementary video for visual quality judgement.



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

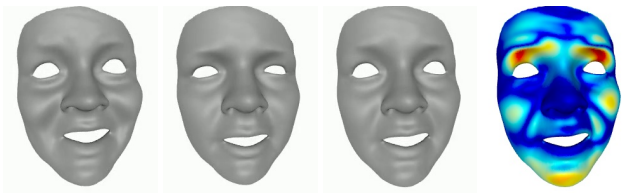
Figure 3: Effect of the emotion label during inference: our approach can generate facial animations that are style controllable by a binary emotion label. Given two in-the-wild audio signal examples, Figs. 3a to 3c correspond to the male example whereas Figs. 3e to 3f correspond to the female example. Figs. 3a and 3d are generated neutrally whereas Figs. 3b and 3e are generated expressively. Figs. 3c and 3f show the colorized differences based on per-vertex distances between neutrally and expressively generated face meshes where extreme red depicts 100% of the computed distance and extreme blue depicts 0% of the computed distance. It is evident that the emotion signal effects the facial regions that are uncorrelated with speech.

5.3.2 Perceptual Evaluation.

> "A work of art doesn't exist outside the perception of the audience."
>
> *Abbas Kiarostami*

In order to demonstrate the realism of facial animation produced by our proposed approach and to compare to ground-truth and FaceFormer, we conducted multiple perceptual user studies. Similar to the user study done during first stage of experiments, these user studies were hosted on Qualtrics [52] and carried out using Prolific [51] in A/B testing settings. We conducted three separate user studies where the users selected their preferences based on the expressiveness of the rendered 3D facial animation videos. In Experiment I, the users indicated their preference between Ground-truth data vs Ours. In Experiment II, the users were shown side-by-side comparisons of FaceFormer and FaceXHuBERT generated with the same audio input. Lastly, in Experiment III the users responded with their preference between neutrally and expressively generated animations both using our approach. Fig. 3 quantitatively and visually shows the difference between neutrally and expressively generated animations. The facial regions (i.e. upper face, eye region, cheeks) that are expected to be effected by the emotion condition deform differently than in the corresponding neutral animation. To solidify our argument, the third experiment was conducted to prove that the expressiveness in facial animation is actually perceived by users.

Tab. 4 depicts the results of the three user studies we conducted during stage 2 of experiments. For all three experiments, the participants where shown a series of A vs. B video pairs and were asked to choose one between the two animations that looks more expressive. In order to ensure good quality data, we included random attention check questions in the study. The attention check questions consist of a video pair in which one of the videos is totally out-of-sync with the accompanied audio. In total, 147 participants were recruited to the user studies in stage 2 in which 51 participated in Experiment I, 31 in Experiment II and 65 in Experiment III. In Experiment I, on average, 25.45% of the participants preferred animation generated by our model over the ground-truth animation. This is expected as the generated animations do not model some of the facial motion such as the eye blinks that are present in the ground-truth. Furthermore, the ground-truth has more variations across the whole face. In Experiment II, on an average, 77.73% of the participants preferred animations generated by our model over the animations generated by FaceFormer. In Experiment III, on average, 66.35% of the participants preferred animations generated by our model with expressive signal over the animations generated without the expressive signal. More details on these user studies can be found in the supplementary material document.



(a) GT **(b) FaceFormer** **(c) Ours** **(d) Difference**

Figure 4: Qualitative comparison of expressiveness between FaceFormer and FaceXHuBERT (Ours). Given an audio sequence from the test-set, Fig. 4a is the ground-truth whereas Figs. 4b and 4c are the corresponding frames from animations generated using FaceFormer and FaceXHuBERT respectively. Fig. 4d shows the colorized per vertex distance computed between Figs. 4b and 4c. Animation generated by our approach is more expressive because the upper face region is more responsive and expressive to the emotionally expressive audio sequences than the one of FaceFormer.

User Preference	Expressiveness
I. Ours *vs.* Ground-Truth	25.45 ± 13.20
II. Ours *vs.* FaceFormer	77.73 ± 17.59
III. Ours-Emotional *vs.* Ours-Neutral	66.35 ± 5.66

Table 4: User study results: The Expressiveness score depicts the average percentage of participants that preferred the left item to the right item in the corresponding row in terms of expressiveness.

Training Details and Tools: All the models in our experiments were trained for 100 epochs on an HP ZBook Fury G7 laptop with an Intel Core i7-10850H 2.7 GHz (12 cores) CPU, 32GB RAM and Nvidia Quadro RTX 3000 6GB VRAM. We optimized on Huber Loss function [27] and used Adam optimizer [35] during the training. The dropout value of the recurrent units was set to 0.3. We used PyTorch [49] for implementing the models. Meshlab [13] and PyMeshLab [45] were used qualitative visualization. Trimesh [65] together with OpenCV [5] and ffmpeg [64] were used for rendering and visualizing the animations. The entire codebase for training, evaluating and visualizing can be found in the supplementary material.

6 ABLATION STUDY

Ablation on FaceXHuBERT Encoder. We conducted extensive experiments to understand and optimize the effect of HuBERT in our encoder. This ablation study is devised by freezing the model's pretrained weights at various layers starting from no-freezing (i.e. all the layer parameters are trainable) to all weights frozen (i.e. all the layers are frozen to their pretrained weights). The results of the ablation study is reported in Tab. 5a. In light of the encoder structure as described in Sec. 4.1, the network configurations for this ablation study are as follows- **(i)** no layers are frozen, **(ii)** CNN encoder is frozen, **(iii)** CNN encoder and feature projection layers are frozen. For models **(iv)** to **(viii)**, in addition to freezing the CNN encoder and feature projection layer, we incrementally freeze two transformer layers. For model **(ix)**, only the last transformer layer is kept trainable whereas for **(x)**, the entire pretrained model is frozen during training. Although all the above mentioned configurations yield qualitatively coherent animations, we found that **(iv)** produces the least mean face vertex error. Furthermore, as we move from model **(iv)** to **(x)**, the animations become more stiffer and less expressive.

Ablation on FaceXHuBERT Decoder. We experimented with different configurations of the GRU structure in terms of number of hidden layers and hidden unit size to optimize for the mean vertex error. Tab. 5b shows the results of the FaceXHuBERT decoder ablation study. The first column represents the number of hidden layers (e.g. 2L) and hidden size (e.g. 128) in the GRU structure. All the configurations mentioned in the table produce qualitatively coherent animations but training the proposed configuration results in producing the least mean face vertex error, ensuring that the predictions are more closer to the ground-truth, hence more realistic and expressive.

Model variant	Full Face Vertex Error ($\times 10^{-3}$)mm	Trainable parameters (in millions)	Training Time (hours)
(i)	2.04	114.1	≈ 8.19
(ii)	1.79	109.9	≈ 5.75
(iii)	1.84	109.5	≈ 6.38
(iv)	**1.72**	**95.3**	≈ **5.10**
(v)	1.81	81.2	≈ 5.83
(vi)	1.73	67.0	≈ 5.50
(vii)	1.86	52.8	≈ 5.27
(viii)	1.95	38.6	≈ 5.13
(ix)	1.82	31.6	≈ 4.86
(x)	2.05	19.7	≈ 3.88

(a) FaceXHuBERT Encoder Ablation Study Results.

Model variant	Full Face Vertex Error ($\times 10^{-3}$)mm	Trainable parameters (in millions)	Training Time (hours)
2L-256	**1.72**	**95.3**	≈ **5.10**
1L-256	1.87	95.0	≈ 5.55
2L-128	1.86	85.3	≈ 5.33
2L-64	1.88	80.4	≈ 5.27
2L-32	2.07	78.0	≈ 4.30

(b) FaceXHuBERT Decoder Ablation Study Results.

Table 5: FaceXHuBERT Ablation Study Results.

7 DISCUSSION AND LIMITATIONS

Using a self-supervised pretrained speech model such as HuBERT produces significant improvements for the 3D speech-driven facial animation task. It clearly shows the importance of the encoder model, while using a simple decoder component. It does not only have the ability to disambiguate speech uncorrelated factors for facial animation, but also addresses the scarcity of synchronized audio-visual datasets by incorporating pretrained speech representations based on a large speech model. We assume that it can be adopted to solve similar downstream tasks such as audio-driven gesture synthesis. Additionally, we showed that guiding the training with an emotion label generates the facial deformations uncorrelated with speech and correlated with emotion context.

Due to the limitation of the BIWI dataset, we could only guide the learning in a binary manner (i.e. neutral and expressive). However, we assume that with a balanced dataset containing specific emotion categories in the data, we will be able to learn and generate audio-driven facial animations for respective emotion categories. Furthermore, our approach is limited to offline animation generation. In the future, we plan to extend our work to be real-time friendly. Additionally, since the face scans of BIWI do not contain eyes and tongue, our method could not take into account animations of some face parts such as eye gaze and tongue.

8 CONCLUSION

In this paper, we presented FaceXHuBERT, an efficient text-less speech-driven expressive 3D facial animation method using self-supervised speech representation learning. First, we showed that our approach performs better than state-of-the-art on the 3D speech-driven facial animation task. Second, we incorporated emotion modality in the architecture to achieve emotion controllable animation synthesis. At the core of our model is the pretrained HuBERT-based encoder combined with an efficient GRU-based decoder instead of a complex model based on e.g. transformers. Our method can produce accurate lip-sync and expressive facial animations for arbitrary audio input without the need of long training times and large dataset. Our method does not only produce accurate lip-sync but also generates personalized and subtle facial cues (e.g. identity and emotional expressiveness). It is generalizable in terms of language and can handle audio recorded in a variety of situations (e.g. multiple people speaking, background sound, laughter, lip-smacking). Our extensive objective and subjective analysis shows that FaceXHuBERT outperforms the state-of-the-art. We hope that our approach will be a stepping stone towards text-less speech-driven expressive 3D facial animation.

Ethical Consideration Models trained on face scans can easily be used for generating synthetic content that can jeopardize humans and their privacy. We must act responsibly by considering the aspects pertaining to privacy and ethics.

Acknowledgement We would like to thank the authors of VOCA, FaceFormer and CodeTalker for making their codes available and to ETH Zurich CVL for providing us access to the *Biwi 3D Audiovisual Corpus*. We express our gratitude to our colleagues for providing different language audios and for the feedback.

REFERENCES

[1] 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)*.

[2] 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

[3] V. Barrielle and N. Stoiber. 2018. Realtime Performance-Driven Physical Simulation for Facial Animation. *Computer Graphics Forum* 38, 1 (June 2018), 151–166. https://doi.org/10.1111/cgf.13450

[4] Sofien Bouaziz, Yangang Wang, and Mark Pauly. 2013. Online Modeling for Realtime Facial Animation. *ACM Trans. Graph.* 32, 4, Article 40 (jul 2013), 10 pages. https://doi.org/10.1145/2461912.2461976

[5] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[7] Shanna L. Burke, Tammy Bresnahan, Tan Li, Katrina Epnere, Albert Rizzo, Mary Partin, Robert M. Ahlness, and Matthew Trimmer. 2017. Using Virtual Interactive Training Agents (ViTA) with Adults with Autism and Other Developmental Disabilities. *Journal of Autism and Developmental Disorders* 48, 3 (Nov. 2017), 905–912. https://doi.org/10.1007/s10803-017-3374-z

[8] Chen Cao, Qiming Hou, and Kun Zhou. 2014. Displaced Dynamic Expression Regression for Real-Time Facial Tracking and Animation. *ACM Trans. Graph.* 33, 4, Article 43 (jul 2014), 10 pages. https://doi.org/10.1145/2601097.2601204

[9] Yong Cao, Wen C. Tien, Petros Faloutsos, and Frédéric Pighin. 2005. Expressive Speech-Driven Facial Animation. *ACM Trans. Graph.* 24, 4 (oct 2005), 1283–1302. https://doi.org/10.1145/1095878.1095881

[10] Prashanth Chandran, Loïc Ciccone, Markus Gross, and Derek Bradley. 2022. Local Anatomically-Constrained Facial Performance Retargeting. *ACM Trans. Graph.* 41, 4, Article 168 (jul 2022), 14 pages. https://doi.org/10.1145/3528223.3530114

[11] Constantinos Charalambous, Zerrin Yumak, and A.F. van der Stappen. 2019. Audio-driven emotional speech animation for interactive virtual characters. *Computer Animation and Virtual Worlds* 30 (2019).

[12] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. https://doi.org/10.48550/ARXIV.1409.1259

[13] Paolo Cignoni, Marco Callieri, Massimiliano Corsini, Matteo Dellepiane, Fabio Ganovelli, and Guido Ranzuglia. 2008. MeshLab: an Open-Source Mesh Processing Tool. In *Eurographics Italian Chapter Conference*, Vittorio Scarano, Rosario De Chiara, and Ugo Erra (Eds.). The Eurographics Association. https://doi.org/10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2008/129-136

[14] 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/

[15] 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.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[17] DI4D 2023. DI4D. https://di4d.com/. Accessed: 2022-11-05.

[18] Dynamixyz 2023. Dynamixyz. https://www.dynamixyz.com. Accessed: 2022-11-05.

[19] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 2020. 3D Morphable Face Models—Past, Present, and Future. *ACM Trans. Graph.* 39, 5, Article 157 (jun 2020), 38 pages. https://doi.org/10.1145/3395208

[20] Faceware 2023. Faceware. https://facewaretech.com/. Accessed: 2022-11-05.

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

[22] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. 2022. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] 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.

[24] 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

[25] 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

[26] 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. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

[27] Peter J. Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35, 1 (1964), 73 – 101. https://doi.org/10.1214/aoms/1177703732

[28] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D Avatar Creation from Hand-Held Video Input. *ACM Trans. Graph.* 34, 4, Article 45 (jul 2015), 14 pages. https://doi.org/10.1145/2766974

[29] Alexandru-Eugen Ichim, Petr Kadleček, Ladislav Kavan, and Mark Pauly. 2017. Phace: Physics-Based Face Modeling and Animation. *ACM Trans. Graph.* 36, 4, Article 153 (jul 2017), 14 pages. https://doi.org/10.1145/3072959.3073664

[30] Geethu Miriam Jacob and Bjorn Stenger. 2021. Facial Action Unit Detection With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7680–7689.

[31] JALI 2023. JALI Research. https://jaliresearch.com/. Accessed: 2022-11-05.

[32] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. 2022. EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model. In *ACM SIGGRAPH 2022 Conference Proceedings* (Vancouver, BC, Canada) *(SIGGRAPH '22)*. Association for Computing Machinery, New York, NY, USA, Article 61, 10 pages. https://doi.org/10.1145/3528233.3530745

[33] Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. 2020. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *International Conference on Intelligent Virtual Agents (IVA '20)*. ACM.

[34] 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

[35] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980 cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

[36] J. P. Lewis, Ken Anjyo, Taehyun Rhee, Mengjie Zhang, Fred Pighin, and Zhigang Deng. 2014. Practice and Theory of Blendshape Facial Models. In *Eurographics 2014 - State of the Art Reports*, Sylvain Lefebvre and Michela Spagnuolo (Eds.). The Eurographics Association. https://doi.org/10.2312/egst.20141042

[37] 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, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

[38] Yuanxun Lu, Jinxiang Chai, and Xun Cao. 2021. Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation. *ACM Transactions on Graphics* 40, 6 (12 2021), 17 pages. https://doi.org/10.1145/3478513.3480484

[39] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De la Torre, and Yaser Sheikh. 2021. Pixel Codec Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 64–73.

[40] Araceli Morales, Gemma Piella, and Federico M. Sukno. 2021. Survey on 3D face reconstruction from uncalibrated images. *Computer Science Review* 40 (2021), 100400. https://doi.org/10.1016/j.cosrev.2021.100400

[41] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100.

https://doi.org/10.1109/MRA.2012.2192811
[42] Lucio Moser, Chinyu Chien, Mark Williams, Jose Serra, Darren Hendler, and Doug Roble. 2021. Semi-Supervised Video-Driven Facial Animation Transfer for Production. *ACM Trans. Graph.* 40, 6, Article 222 (dec 2021), 18 pages. https://doi.org/10.1145/3478513.3480515
[43] Maher Ben Moussa, Zerrin Kasap, Nadia Magnenat-Thalmann, Krishna Chandramouli, Seyed Navid Haji Mirza, Qianni Zhang, Ebroul Izquierdo, Iordanis Biperis, and Petros Daras. 2010. Towards an Expressive Virtual Tutor: An Implementation of a Virtual Tutor Based on an Empirical Study of Non-Verbal Behaviour. In *Proceedings of the 2010 ACM Workshop on Surreal Media and Virtual Cloning* (Firenze, Italy) *(SMVC '10)*. Association for Computing Machinery, New York, NY, USA, 39–44. https://doi.org/10.1145/1878083.1878096
[44] MPG. 2019. *VOCASET*. https://voca.is.tue.mpg.de/
[45] Alessandro Muntoni and Paolo Cignoni. 2021. PyMeshLab. https://doi.org/10.5281/zenodo.4438750
[46] 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. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).
[47] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject Agnostic Face Swapping and Reenactment. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 7183–7192.
[48] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964
[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf
[50] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr. Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. https://doi.org/10.48550/ARXIV.2005.05535
[51] Prolific 2023. Prolific. https://www.prolific.co. Accessed: 2022-11-05.
[52] Qualtrics 2023. Qualtrics. https://www.qualtrics.com. Accessed: 2022-11-05.
[53] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
[54] Roger Blanco i Ribera, Eduard Zell, J. P. Lewis, Junyong Noh, and Mario Botsch. 2017. Facial Retargeting with Automatic Range of Motion Alignment. *ACM Trans. Graph.* 36, 4, Article 154 (jul 2017), 12 pages. https://doi.org/10.1145/3072959.3073674
[55] 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
[56] 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)*. 7763–7772.
[57] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. 2018. Facial Action Unit Detection Using Attention and Relation Learning. *CoRR* abs/1808.03457 (2018). arXiv:1808.03457 http://arxiv.org/abs/1808.03457
[58] 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
[59] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing Obama: Learning Lip Sync from Audio. *ACM Trans. Graph.* 36, 4, Article 95 (jul 2017), 13 pages. https://doi.org/10.1145/3072959.3073640
[60] Man To Tang, Victor Long Zhu, and Voicu Popescu. 2021. AlterEcho: Loose Avatar-Streamer Coupling for Expressive VTubing. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 128–137. https://doi.org/10.1109/ISMAR52148.2021.00027
[61] 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
[62] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain Matthews. 2012. Dynamic Units of Visual Speech. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation* (Lausanne, Switzerland) *(SCA '12)*. Eurographics Association, Goslar, DEU, 275–284.
[63] 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
[64] Suramya Tomar. 2006. Converting video formats with FFmpeg. *Linux Journal* 2006, 146 (2006), 10.
[65] Trimesh 2023. Trimesh. https://trimsh.org. Accessed: 2022-11-05.
[66] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[67] 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
[68] 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)*.
[69] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. FaceScape: A Large-Scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 598–607. https://doi.org/10.1109/CVPR42600.2020.00068
[70] Juyong Zhang, Keyu Chen, and Jianmin Zheng. 2022. Facial Expression Retargeting From Human to Avatar Made Easy. *IEEE Transactions on Visualization and Computer Graphics* 28, 2 (2022), 1274–1287. https://doi.org/10.1109/TVCG.2020.3013876
[71] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. 2021. Pose-Controllable Talking Face Generation by Implicitly Modularized Audio-Visual Representation. *CoRR* abs/2104.11116 (2021). arXiv:2104.11116 https://arxiv.org/abs/2104.11116
[72] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. 2018. Visemenet: Audio-Driven Animator-Centric Speech Animation. *ACM Trans. Graph.* 37, 4, Article 161 (jul 2018), 10 pages. https://doi.org/10.1145/3197517.3201292
[73] M. Zollhöfer, J. Thies, P. Garrido, D. Bradley, T. Beeler, P. Pérez, M. Stamminger, M. Nießner, and C. Theobalt. 2018. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. *Computer Graphics Forum* 37, 2 (2018), 523–550. https://doi.org/10.1111/cgf.13382 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13382