

Emotional Speech-Driven Animation with Content-Emotion Disentanglement

Radek Daněček¹

rdanecek@tue.mpg.de

Kiran Chhatre²

chhatre@kth.se

Shashank Tripathi¹

stripathi@tue.mpg.de

Yandong Wen¹

ywen@tue.mpg.de

Michael J. Black¹

black@tue.mpg.de

Timo Bolkart¹

tbolkart@tue.mpg.de

¹Max Planck Institute for Intelligent Systems, Tübingen, Germany ²KTH, Stockholm, Sweden

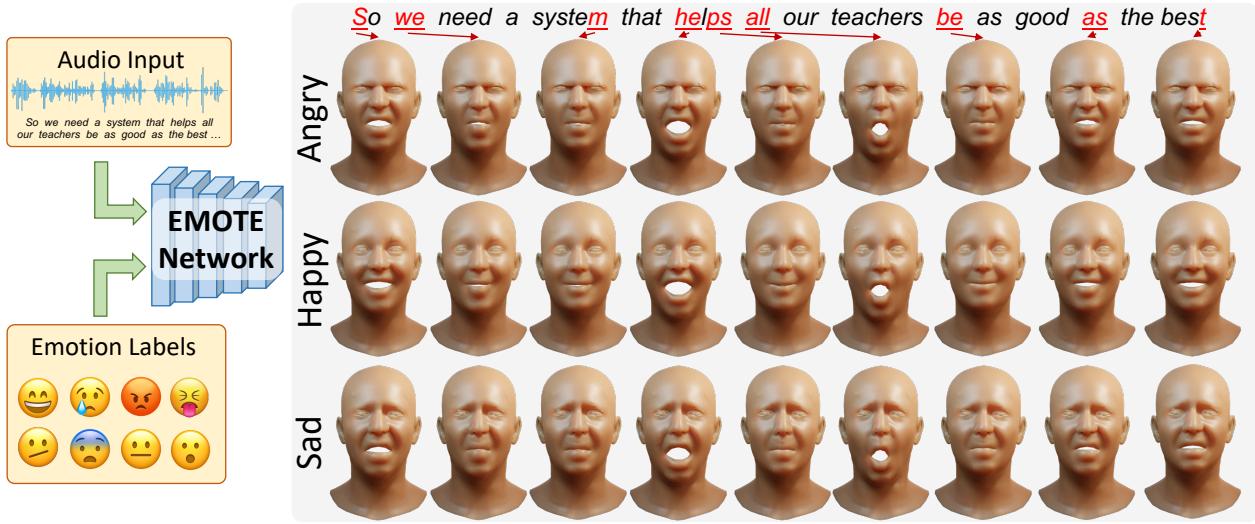


Figure 1. Given audio input and an emotion label, EMOTE generates an animated 3D head that has state-of-the-art lip synchronization while expressing the emotion. The method is trained from 2D video sequences using a novel video emotion loss and a mechanism to disentangle emotion from speech.

Abstract

To be widely adopted, 3D facial avatars must be animated easily, realistically, and directly from speech signals. While the best recent methods generate 3D animations that are synchronized with the input audio, they largely ignore the impact of emotions on facial expressions. Realistic facial animation requires lip-sync together with the natural expression of emotion. To that end, we propose EMOTE (Expressive Model Optimized for Talking with Emotion), which generates 3D talking-head avatars that maintain lip-sync from speech while enabling explicit control over the expression of emotion. To achieve this, we supervise EMOTE with decoupled losses for speech (i.e., lip-sync) and emotion. These losses are based on two key observations: (1) deformations of the face due to speech are spatially localized around the mouth and have high temporal frequency, whereas (2) facial expressions may deform the whole face and occur over longer intervals. Thus we train EMOTE with a per-frame

lip-reading loss to preserve the speech-dependent content, while supervising emotion at the sequence level. Furthermore, we employ a content-emotion exchange mechanism in order to supervise different emotions on the same audio, while maintaining the lip motion synchronized with the speech. To employ deep perceptual losses without getting undesirable artifacts, we devise a motion prior in the form of a temporal VAE. Due to the absence of high-quality aligned emotional 3D face datasets with speech, EMOTE is trained with 3D pseudo-ground-truth extracted from an emotional video dataset (i.e., MEAD). Extensive qualitative and perceptual evaluations demonstrate that EMOTE produces speech-driven facial animations with better lip-sync than state-of-the-art methods trained on the same data, while offering additional, high-quality emotional control.

1. Introduction

Animating 3D head avatars solely from speech has numerous applications, including character animation in films and games, virtual telepresence for AR and VR, and the embodiment of digital personal assistants. For the best user experience, the speech-driven animation methods must be speaker-independent (i.e., generalize to the audio and facial geometry of unseen subjects) and the lip articulation must be synchronized with the speech content. Lip-sync of the generated animation with the audio has drawn significant attention [16, 25, 51, 62, 68]. But to be lifelike, facial animations must also express natural emotions through facial expressions. The modeling of facial expressions in 3D has also been well studied [24]. What is missing, however, is the modeling and animation of *emotion during speech*.

The core issue is that the 3D face shapes that convey emotion are often inconsistent with the lip motions needed to realistically match the audio. That is, there is a conflict between expressing the emotion and synchronizing the lips with the audio. To address this limitation, we present EMOTE (Expressive Model Optimized for Talking with Emotion), a speech-driven 3D facial animation method with semantic animation control over the expressed emotion. EMOTE addresses the core problem, making 3D animations that convey the appropriate emotion without hurting lip-sync possible.

Emotional speech in the context of 3D facial animation has been previously ignored due to the absence of suitable datasets. While existing speech-driven animation methods are trained on 3D scan datasets with paired audio such as VOCASET [16], BIWI [26], or Multiface [67], no large-scale 3D scan dataset with emotional speech sequences exists. For this reason, we train EMOTE on MEAD [66], an emotional video dataset, which does not provide 3D supervision. To compensate for the lack of 3D data, we generate pseudo ground-truth (GT) 3D data using a combination of state-of-the-art (SOTA) monocular reconstruction methods [17, 27, 28, 76] fine-tuned on MEAD.

Directly training a speech-driven animation model following an architecture of a SOTA method such as FaceFormer [25] on such pseudo-GT, however, results in mediocre motions, where speech-dependent and emotion-dependent lip articulations are poorly disentangled. Furthermore, FaceFormer does not enable any control over the output emotions.

We observe that facial animation is the result of two factors that differ temporally and spatially: speech and emotional state. Specifically, speech-induced articulation (or lip-sync) has a high temporal frequency; the lip shape must match the audio at every point in time. In contrast, emotions are a longer-lasting phenomenon that change at a lower temporal frequency compared with speech-driven articulation. Additionally, speech production is localized around the mouth region, whereas facial expressions may occur over

the entire face region. We hypothesize that these temporal and spatial differences make it possible to disentangle these two phenomena.

Specifically, in order to enforce the consistency of the lip-sync, we apply a per-frame lip-reading consistency loss similar to [28] while enforcing the desired emotion at the sequence level through a novel transformer-based dynamic emotion consistency loss. Finally, to separate the effect of emotion from the effect of the spoken words, we propose a novel emotion-content disentanglement mechanism that we use to train our model.

Naively training a SOTA network such as FaceFormer with the aforementioned components leads to temporally unstable and unnatural results. To ensure that the generated motion is natural and temporally consistent, we first train a facial motion prior, specifically a temporal transformer-based VAE that operates over sequences of 3DMM (FLAME [36]) parameters. We then train a regressor to map the speech audio onto the latent space of the prior.

With this, EMOTE generates high-quality 3D facial animations with accurate lip-sync while enabling the editing of the expressed emotion. We demonstrate the ability to edit emotions qualitatively and quantitatively in perceptual studies.

Our contributions are summarized as: (1) The first method for semantic emotion editing of speech-driven 3D facial animation. (2) A novel supervision mechanism with perceptual lip-reading and dynamic emotion losses and a novel content-emotion disentanglement mechanism. (3) A statistical prior for facial motion that is designed to support manipulation of facial motion with perceptual losses while keeping the animation natural. (4) A bidirectional non-autoregressive architecture that is more efficient than autoregressive transformer-based SOTA methods. The pseudo ground-truth 3D (FLAME parameters) for the MEAD dataset, the trained EMOTE, and code to train and generate speech-driven facial animations with emotion control are available for research purposes at <https://emote.is.tue.mpg.de/>

2. Related Work

The field of speech-driven 3D facial animation has a long history [9, 15, 22, 23, 58, 69]. We focus on the most relevant recent work, which leverages deep learning [16, 25, 32, 45, 46, 51, 57, 68, 74].

Semantic Control: Few methods provide the user with any kind of semantic control of the generated 3D avatar. VOCA [16] and FaceFormer [25] allow the speaking style to be controlled by interpolating the style vectors of training individuals; this does not enable simple editing of emotion. While MeshTalk [51] can generate a variety of results for

the same audio input, there is no mechanism that allows for any control of emotion. Karras et al. [32] learn a type of emotional latent space by jointly learning a feature vector for each training sample in an unsupervised way. Changing this feature vector then allows test-time editing. The learned space, however, does not inherently contain a semantic meaning and this must be manually assigned after training. Since there is no disentanglement mechanism, the model lacks the guarantee that mixing different emotion vectors with different audio input will produce the desired result (i.e. correct lip-sync and desired emotion). Concurrent, and most relevant to EMOTE is EmoTalk [44], a method to animate emotional 3D faces from speech input. Unlike EMOTE, EmoTalk requires artist-curated training data, and it only provides control over the intensity of the expression of emotion, but not over the emotion type. In contrast, EMOTE is, to the best of our knowledge, the first to factor the effect of emotion and speech on the resulting 3D animation via a novel emotion-content disentanglement mechanism, allowing semantically meaningful emotion editing at test time.

Works such as [22, 57, 74] automatically animate artist-controllable FACS rigs but also lack explicit speech-driven emotion control and some require additional inputs, i.e., a transcript [22, 23]. JALI for Cyberpunk [23] shows characters with emotional faces, however, the amount of artist work, manually designed rules, and hand-crafted features needed to build the system is unclear.

Supervision: The recent methods are fully supervised [16, 25, 32, 51, 62, 68], requiring a training dataset of 3D scans paired with the synchronized speech. Notably, these methods use a mean squared error loss between the predicted and ground truth mesh vertices (or vertex offsets from a template mesh) at each frame. Richard et al. [51] introduce a cross-modal loss to enforce reconstruction of audio-correlated and uncorrelated information separately in order to learn a categorical motion prior and an explicit eye-closure loss to enforce eye blinks. Thambiraja et al. [62] introduce a mouth-closure loss that is active only when bilabials are spoken, which helps achieve proper mouth closure. EMOTE goes further to use perceptual lip- and emotion-consistency losses with a novel disentanglement framework.

Motion prior: Most methods do not learn or apply any type of motion prior and let the space of valid motions be learned by the architecture itself from the training data [16, 25, 32, 45, 46]. MeshTalk [51], on the other hand, adopts a two-stage approach. In the first stage, a motion prior with a categorical latent space is trained. This pretrained prior is then used in the second stage to autoregressively generate the results. CodeTalker [68] adopts an approach similar to FaceFormer [25] and augments it with a separately trained

VQ-VAE motion prior. Chandran et al. [10] introduce a transformer-based autoencoder for facial motion animation with disentangled identity and shape. The authors demonstrate its effectiveness for tasks like motion compression, retargetting, unconditional generation and others. However, its suitability for regression tasks like speech-driven animation is not investigated.

2D talking head generation: There is a long line of work focused on generating 2D videos of talking heads given speech [3, 11–13, 18, 21, 31, 33, 37, 48, 56, 64, 65, 72] and today there are even commercial systems for this task. These approaches, however, typically ignore emotion and focus on lip-sync. The few methods that address facial-expression animation operate over 2D videos [31, 43, 63, 64]. While some of these methods use 3D parametric models to guide the output expressions (e.g., [43]), their focus is not on outputting the 3D shape and, hence, the underlying 3D shapes are of low quality.

3D Datasets: The ideal dataset for our task would contain ground-truth (GT) 3D face scans synchronized with audio. Such data is limited due to the expense and complexity involved in capturing it. BIWI [26], VOCASET [16], S3DFM [73], and Multiface [67] are publicly available datasets for the audio-driven 3D talking-head task. These datasets are limited in size, richness of emotion, speaking styles, and shapes of the subjects.

2D Datasets: In contrast to the limited richness of 3D datasets, 2D video is plentiful. Specifically, there are many available video-speech datasets [1, 2, 14, 41, 52, 66, 75], and video datasets focused speech emotion recognition (SER) [5, 7, 38, 47, 71]. See the Sup. Mat. for an overview of existing 2D and 3D datasets. Of the existing video datasets, MEAD [66] is most suitable for our task. It is of sufficient size, it is captured in the lab, which makes it easier to perform 3D face reconstruction than in-the-wild video, and, most importantly, it exhibits high emotional variety.

Off-the-shelf 3D face-reconstruction methods can be applied to the video frames, providing pseudo-GT data. This, however, comes with many drawbacks. While the field of image-based 3D face reconstruction has made tremendous progress [17, 19, 27–29, 54, 55, 59–61, 70, 76, 77], SOTA methods are often not robust to occlusion, they produce inaccurate shape or expression, or are not temporally stable. Despite these limitations, the large amount of data available from video outweighs the downsides. Consequently, we generate pseudo-GT data from video by integrating recent SOTA methods [17, 28, 76].

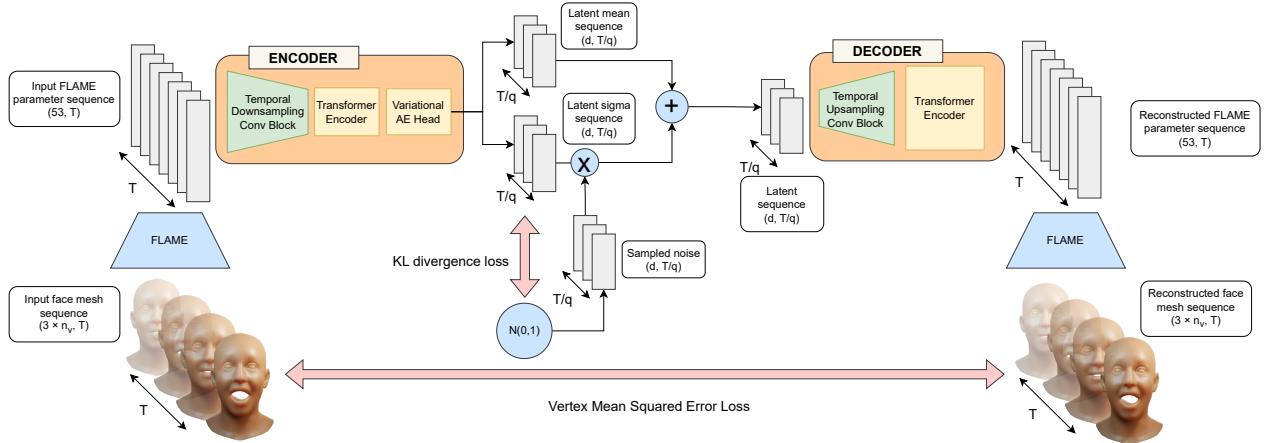


Figure 2. **FLINT motion prior architecture.** Given a sequence of T FLAME parameters, the encoder maps this sequence to a sequence of compact latents. The decoder then reconstructs this latent sequence back into a sequence of T FLAME parameters. The reparametrization trick is employed to sample the latents from predicted sequences of means and sigmas.

3. Background and Notation

Face model: EMOTE predicts the expression and jaw pose parameters of FLAME [36], a parametric 3D head model. FLAME is defined as a function $M(\beta, \theta, \psi) \rightarrow (\mathbf{V}, \mathbf{F})$ that takes parameters for identity shape $\beta \in \mathbb{R}^{|\beta|}$, facial expression $\psi \in \mathbb{R}^{|\psi|}$, and pose $\theta \in \mathbb{R}^{3k+3}$, with $k = 4$ joints, and outputs a 3D mesh with vertices $\mathbf{V} \in \mathbb{R}^{n_v \times 3}$ and triangles $\mathbf{F} \in \mathbb{R}^{n_f \times 3}$.

Emotion feature extraction: We use EMOCA’s [17] publicly available emotion recognition network to predict emotion features from an image. This model consists of a ResNet-50 [30] network trained on AffectNet [40] on the task of in-the-wild emotion recognition to regress valence and arousal and classify the eight basic expressions (neutral, happiness, sadness, surprise, fear, disgust, anger, and contempt). After training the network, the prediction head is discarded, and the output of the final ResNet layer is used as an emotion feature vector, $\epsilon \in \mathbb{R}^{|\epsilon|}$. In the following, we denote the network with $E_{\text{emo}}^{\text{im}}(I) \rightarrow \epsilon$.

Video emotion feature extraction: Emotions are phenomena that generally last at least several seconds or more. Single-frame emotion features are insufficient to describe them, because an expression in a single frame carries the effect of both emotion and speech. This can lead to misinterpreting speech-induced articulation for emotional cues (see Fig. 3 in Sup. Mat.). Hence, emotion features should aggregate information across time. To address this, we train a lightweight transformer-based emotion classifier that takes a sequence of emotion features $\epsilon^{1:T} \in \mathbb{R}^{T \times |\epsilon|}$ and outputs a video emotion classification vector $\mathbf{e} \in \mathbb{R}^8$ and the video emotion feature $\phi \in \mathbb{R}^{|\phi|}$, which is the sequence-aggregated feature pro-

duced by the last transformer layer before the classification head, with $|\phi| = 256$. We refer to the video motion feature extraction as $E_{\text{emo}}^{\text{vid}}(\epsilon^{1:T}) \rightarrow (\mathbf{e}, \phi)$. More details about the video emotion extraction can be found in Sup. Mat.

Speech feature extraction: To encode the audio signal, we employ a pretrained ASR network, Wav2Vec 2.0 [4]. It takes as input the raw waveform sampled at 16kHz. This waveform is first passed through temporal convolutional layers producing a feature sampled at 50Hz. Similar to Fan et al. [25], we use linear interpolation to downsample the feature down to 25Hz to match the frame-rate of our input videos. The resampled feature is then fed into the transformer-based part of Wav2Vec 2.0, producing the output speech feature. Formally, it is defined as $A(\mathbf{w}) \rightarrow \mathbf{s}^{1:T}$, where A is the Wav2Vec 2.0 network, \mathbf{w} is the raw waveform, and $\mathbf{s}^{1:T} \in \mathbb{R}^{T \times d_s}$ is the final speech feature resampled to 25Hz. T denotes the number of frames and each frame is of dimension $d_s = 768$.

4. Method

Motivation: EMOTE follows a two-step pipeline, which first trains a temporal variational autoencoder, and then uses its latent space as a motion prior. Specifically, we train a regressor that maps the speech audio to the latent space of the motion prior conditioned on a given target emotion, its intensity (mild, medium, or high), and a subject-specific speaking style.

4.1. Facial Motion Prior: FLINT

Facial motion is complex and modeling it is challenging. To simplify the problem we represent it in a learned low-dimensional representation. As a foundation, we represent the face in each of the T frames of a sequence using FLAME,

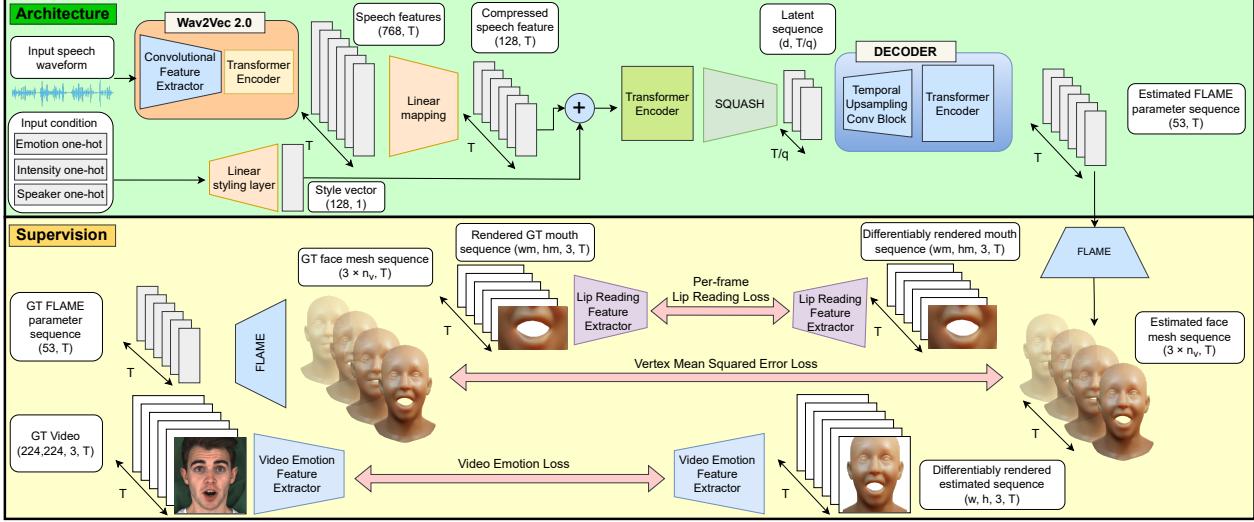


Figure 3. EMOTE architecture. The speech input consists of the raw audio waveform (top left) and conditioning, which includes one-hot vectors of the training speaker ID and, importantly, the emotion class and intensity. The audio is encoded by Wav2Vec 2.0 and the input conditions are mapped with a linear styling layer. These two are then concatenated and passed through additional convolutional layers to map down to the latent space sequence. The pretrained frozen FLINT decoder transforms this sequence back into a FLAME parameter sequence. Finally, the meshes are generated by the FLAME forward pass.

giving $|\psi| + |\theta_{jaw}| = 53$ dimensions (i.e., 50 expression parameters and 3 jaw pose parameters) per frame. Facial motions, however, are not independent between frames and, hence, a sequence can be represented in a lower-dimensional space. To that end, we train a temporal variational autoencoder called FLINT (FLAME IN Time) to represent facial motion sequences. The formulation exploits a transformer encoder to extend the VAE framework to our temporal modeling problem (cf. [42]). We exploit this as a prior in training EMOTE and find that it reduces high-frequency jitter and unnatural jaw rotations.

Architecture: The encoder compresses the sequence of T frames $(\psi^{1:T}, \theta_{jaw}^{1:T})$ into T/q latent frames $\mathbf{z}^{1:T/q}$, where q is the number of consecutive original frames that a single latent frame encapsulates (similar to [42]). The intervals of consecutive latents do not overlap. We empirically set $q = 8$. Specifically,

$$\text{ENC}(\psi^{1:T}, \theta_{jaw}^{1:T}) \rightarrow (\mu^{1:T/q}, \sigma^{1:T/q}). \quad (1)$$

Using the VAE reparametrization gives us the final latent sequence: $\mathbf{z}^t = \sigma^t * \mathbf{z}_s^t + \mu^t$, where \mathbf{z}^t is one latent frame and \mathbf{z}_s^t is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This is done separately for each latent frame $t \in \{1, \dots, T/q\}$, before they are stacked to compose the final sequence of latents, $\mathbf{z}^{1:T/q}$. The sequence $\mathbf{z}^{1:T/q}$ is then decoded back to the original space:

$$\text{DEC}(\mathbf{z}^{1:T/q}) \rightarrow (\hat{\psi}^{1:T}, \hat{\theta}_{jaw}^{1:T}), \quad (2)$$

The architecture of our autoencoder is outlined in Fig. 2, and the hyperparameters are in Sup. Mat.

Losses: We train FLINT with the following loss functions:

$$L_{\text{total}} = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{KL}} L_{\text{KL}}. \quad (3)$$

Reconstruction loss: For each frame t , we compute the mean squared error between the pseudo-GT and predicted meshes:

$$L_{\text{rec}} = \text{MSE}(\hat{\mathbf{V}}^{1:T}, \mathbf{V}^{1:T}), \quad (4)$$

where the vertex coordinates $\mathbf{V}^t, \hat{\mathbf{V}}^t$ are produced by feeding the GT and reconstructed parameters through FLAME: $\mathbf{V}^t = M(\beta, \psi^t, \theta_{jaw}^t)$.

KL divergence: For each latent frame in the sequence, we compute the standard VAE KL divergence term [35].

$$L_{\text{KL}} = 0.5 \left[-\sum_i (\log \sigma_i^2 + 1) + \sum_i \sigma_i^2 + \sum_i \mu_i^2 \right]. \quad (5)$$

Please note that the individual latent means and sigmas are not treated here as a sequence but separately.

4.2. Emotional Speech-Driven Animation: EMOTE

Architecture: EMOTE is an encoder-decoder architecture summarized in Fig. 3. The encoder uses Wav2Vec 2.0 [4] to extract the audio feature sequence: $A(\mathbf{w}) = \mathbf{s}^{1:T}$. Each

extracted audio feature \mathbf{s}^t is concatenated with a style vector: $\mathbf{s}_s^{1:T} = [S(\mathbf{c})^{1:T} | \mathbf{s}^{1:T}]$, with $S(\mathbf{c})$ denoting the styling function, which is a linear projection of the input condition \mathbf{c} . At training time, \mathbf{c} is the ground-truth emotion type, emotion intensity, and speaker ID:

$$\mathbf{c} = [\mathbf{c}_{emo} | \mathbf{c}_{int} | \mathbf{c}_{id}], \quad (6)$$

where \mathbf{c}_{emo} , \mathbf{c}_{int} , \mathbf{c}_{id} are one-hot vectors of emotion, intensity and identity indices. At test time, \mathbf{c} can be set manually, which provides animator control over the emotion of the output sequence. After the style is incorporated, the speech feature is mapped to the latent space of the motion prior. Specifically, it is temporally downsampled by concatenating q consecutive frames together and then projecting it with a linear layer down to a single latent frame: $\text{SQUASH}(\mathbf{s}_s^{1:T}) = \mathbf{z}^{1:T/q}$. Finally, the obtained latent sequence $\mathbf{z}^{1:T/q}$ is fed to the pretrained, frozen, motion decoder to produce the output FLAME parameters using Eq. 2, obtaining the estimates of $\hat{\psi}^{1:T}, \hat{\theta}_{jaw}^{1:T}$.

During training, both the GT and predicted geometry are rendered with a differentiable renderer [50], and the images are passed to a lip-reading network E_{lip} and video emotion network E_{emo}^{vid} . We denote the forward pass, including the differentiable rendering, and the extraction of emotion and lip-reading features as:

$$\text{EMOTE}(\mathbf{s}^{1:T}, \mathbf{c}) \rightarrow (\hat{\mathbf{V}}^{1:T}, \hat{\eta}^{1:T}, \hat{\phi}), \quad (7)$$

where $\hat{\mathbf{V}}^{1:T}$ is the generated vertex sequence, $\hat{\eta}^{1:T}$ is the sequence of lip-reading features, and $\hat{\phi}$ is the video emotion feature.

Note that unlike recent transformer-based SOTA methods, EMOTE is not autoregressive. Hence, the decoder is only called once, making the decoding computational complexity $O(1)$. This is more efficient than the $O(T)$ autoregressive decoding loop of FaceFormer and CodeTalker. It also allows us to consider future context by employing bidirectional decoding similar to BERT [20].

Training: During training, we supervise the model using the following loss functions:

$$L_{\text{total}} = L_{\text{rec}} + L_{\text{emo}} + L_{\text{lip}} + L_{\text{emo}}^{\text{dis}} + L_{\text{lip}}^{\text{dis}}. \quad (8)$$

Reconstruction loss: For each frame t in the sequence we compute the mean squared error between the pseudo-GT and predicted meshes:

$$L_{\text{rec}} = \text{MSE}(\hat{\mathbf{V}}^{1:T}, \mathbf{V}^{1:T}). \quad (9)$$

Video emotion loss: We extract the video emotion feature from the original video $E_{emo}^{\text{vid}}(E_{emo}^{\text{im}}(\mathbf{I}^{1:T})) = \phi$ and from the

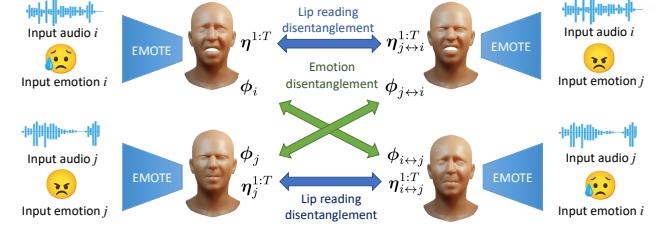


Figure 4. Disentanglement mechanism. During training, we duplicate batches and exchange the emotion condition in the duplicated batch (right), the augmented batch is also passed through the model, and we compute the disentanglement losses such that the result of the augmented batch keeps the original articulation but takes on the desired emotion.

differentiably-rendered predicted sequence and call this $\hat{\phi}$. Their emotional content should be the same, so we penalize their distance:

$$L_{\text{emo}} = d_e(\hat{\phi}, \phi), \quad (10)$$

where d_e is the negative cosine similarity.

Lip-reading loss: For each frame t in the sequence, we also compute a perceptual lip-reading loss. We crop out the mouth region and feed it to the lip-reading network. We extract per-frame lip-reading features using E_{lip} and calculate the distance between the pseudo-GT lip-reading features and the predicted lip-reading features:

$$L_{\text{lip}} = d_l(\hat{\eta}^{1:T}, \eta^{1:T}), \quad (11)$$

where d_l is the negative cosine similarity.

Disentangling emotion and content: Our goal is to disentangle content and emotion such that we can control one while retaining the other. The conditioning described in Eq. 6 is not sufficient to achieve this. Hence, we devise a novel emotion-content disentanglement mechanism, outlined in Fig. 4. During training, we take two sequences with different emotions and switch their emotion conditions. The lip-reading loss for each decoded result should match the original despite the change in emotion, but the emotion in the decoded result should change to match the new condition.

More formally, let $\text{EMOTE}(\mathbf{s}_i^{1:T}, \mathbf{c}_i) = (\hat{\mathbf{V}}_i^{1:T}, \hat{\eta}_i^{1:T}, \hat{\phi}_i)$ be a forward pass of EMOTE for sample i and similarly for a distinct index j , $\text{EMOTE}(\mathbf{s}_j^{1:T}, \mathbf{c}_j) = (\hat{\mathbf{V}}_j^{1:T}, \hat{\eta}_j^{1:T}, \hat{\phi}_j)$. We also generate sequences with swapped emotion conditions, i.e., $\text{EMOTE}(\mathbf{s}_i^{1:T}, \mathbf{c}_{j↔i}) = (\hat{\mathbf{V}}_{j↔i}^{1:T}, \hat{\eta}_{j↔i}^{1:T}, \hat{\phi}_{j↔i})$ and $\text{EMOTE}(\mathbf{s}_j^{1:T}, \mathbf{c}_{i↔j}) = (\hat{\mathbf{V}}_{i↔j}^{1:T}, \hat{\eta}_{i↔j}^{1:T}, \hat{\phi}_{i↔j})$, with $i ↔ j$ denoting generations using audio j with the emotion and intensity condition of audio i .

Disentanglement losses: We apply both emotion and lip-reading perceptual losses to the augmented samples:

$$L_{\text{emo}}^{\text{dis}} = d_e(\hat{\phi}_{i \leftrightarrow j}, \phi_i) \quad L_{\text{lip}}^{\text{dis}} = d_l(\hat{\eta}_{i \leftrightarrow j}^{1:T}, \eta_j^{1:T}). \quad (12)$$

Since we treat emotion as a sequence phenomenon, rather than a per-frame phenomenon, we can bypass the requirement for temporal alignment between emotion features of $\phi_{i \leftrightarrow j}$ and ϕ_i .

5. Implementation Details

Data: FLINT and EMOTE are trained on the MEAD dataset [66]. MEAD is an emotional video dataset of 48 subjects, each uttering around 30 short English sentences. Each subject utters all sentences several times, once for neutral and three times for seven basic emotions (i.e., anger, disgust, fear, happiness, contempt, sadness, and surprise), where each of the basic emotions is articulated with three intensity levels. The subjects are actresses and actors fluent in English. We use 39 subjects and split the dataset such that 32 subjects are included in the training set and 7 in the validation set; i.e., training and validation sets use different subjects.

For evaluation, we use audio sequences from LRS3 [2], a large scale dataset of English TED and TEDx talks by a large variety of speakers. We use the LRS3 test set in our evaluations, which is disjoint from the speakers in our training set.

Data processing: Since MEAD does not come with 3D meshes, we recover the 3D faces synchronized with the audio directly from the videos. However, we found that existing monocular 3D face reconstruction methods (e.g., [17, 27, 28]) are insufficient to process the data. Specifically, EMOCA [17] best recovers emotional 3D faces, but it is often over-expressive and does not well match the lip articulation. SPECTRE [28] improves the lip articulation but lacks expressiveness of the emotion. Both methods use DECA [27] to recover shape, but this is less accurate than MICA [76]. To get the best of these methods, we augment EMOCA with SPECTRE’s lip-reading loss, replace its predicted identity face shape with MICA’s prediction, and use Mediapipe [39] keypoints instead of FAN [6] keypoints as supervision. We then finetune the image encoders of this combined model on MEAD by minimizing EMOCA’s losses with an additional keypoint loss, and the added lip articulation loss. Once finetuned, running inference of the refined model on MEAD gives the reference 3D face shapes. More details can be found in Sup. Mat.

Motion prior: FLINT is trained on the MEAD dataset for 500 epochs with batch size 4 and sequence lengths of 32 frames. Adam [34] is used as optimizer, with a learning rate

of 1e-4. The size of the latent space is set empirically to 128 and $q = 8$, and the KL divergence term is weighted with a factor of 1e-3.

Speech-driven animation model: We train EMOTE in two stages. In the first stage, we only supervise with the vertex loss without the disentanglement mechanism. This is computationally efficient since it does not require differentiable rendering. We train the model for 20 epochs with batch size 4 and sequence lengths of 64 frames with the Adam optimizer and a learning rate of 1e-4. The first stage is only supervised using the MSE of vertex differences with weight: $\lambda_{\text{rec}} = 1$. In the second stage, we freeze the wav2vec weights, add the differentiable rendering, enable the perceptual losses and the disentanglement mechanism, and fine-tune for two more epochs. The perceptual loss weights are: $\lambda_{\text{emo}} = \lambda_{\text{emo}}^{\text{dis}} = 2.5e - 6$ and $\lambda_{\text{lip}} = \lambda_{\text{lip}}^{\text{dis}} = 2.5e - 5$.

6. Experiments

Evaluation must consider two components: the sync between the lip articulation and the input speech and the quality of the emotional content. As both tasks are difficult to evaluate automatically, we conduct two perceptual experiments and provide qualitative evaluations to demonstrate the quality and effectiveness of EMOTE.

6.1. Perceptual Studies

We conduct two perceptual studies on Amazon Mechanical Turk. First, we compare EMOTE’s lip-sync quality with that of publicly available SOTA methods. Second, individual model components are ablated in order to evaluate the influence of each component on the perceived quality of the results.

Lip articulation evaluation: This study compares EMOTE with CodeTalker [68], FaceFormer [25], MeshTalk [51], and VOCA [16]. We randomly selected 15 input audio sequences from the LRS3 test set and used these to synthesize the facial motion. For results generated with EMOTE, the input emotion condition was set to neutral. In the study, we showed the participants two audible output videos of two different methods side by side (the left-right order was randomized for every hit). The participants played both videos separately. After playing both videos at least once, the participant was allowed to select the result with better lip-sync on a 5-point Likert scale (strong/weak preference for one or the other model, or equally good). Each of the two-way comparison studies was completed by 15 participants. Three catch trials with obvious answers (videos with animation generated by a different audio than the one playing) were added, and participants that preferred the catch trials were excluded (see figures for details).

The Likert plot in Fig. 8 shows preferences averaged across participants. Note that all SOTA methods were trained on high quality audio-4D scan datasets ([16, 67]), while EMOTE is trained on pseudo-GT (i.e., MEAD). EMOTE’s lip-sync is preferred over scan-trained MeshTalk despite being trained on pseudo-GT. Note that the lip-sync of methods trained on VOCASET (VOCA, FaceFormer, CodeTalker) is preferred due to the superior training data quality of VOCASET’s 3D scans. Importantly, in a fair comparison, EMOTE outperforms FaceFormer retrained on MEAD, suggesting the value of our architecture and method.

Ablation experiments: This study evaluates the importance of the individual building blocks of EMOTE. We compare: (1) EMOTE, (2) EMOTE w/o the disentanglement terms, (3) EMOTE w/o disentanglement and emotion loss, (4) EMOTE w/o disentanglement and lip-reading loss, (5) EMOTE w/o FLINT, (6) EMOTE w/o disentanglement and perceptual losses, (7) EMOTE w/ static emotion loss computed per-frame instead of the dynamic emotion loss, (8) FaceFormer-EMO, which is the FaceFormer architecture augmented with a one-hot input for the emotion condition and intensity.

For this study, we randomly select 14 input audios from the LRS3 test set and use these to synthesize the facial motion. We ensure that each of the 7 basic emotions (anger, disgust, fear, happiness, contempt, sadness and surprise) is equally represented. Similar to the study above, we present the participants with two videos. First the videos are muted, and the participant is asked which of the two videos better communicates a particular specified emotion. Then, the same videos are presented but this time with audio and the participant is asked which of the two has better lip-sync (same question as in the study above). Participants answer both questions on a 5-point Likert scale. This process repeats for all 14 video pairs. Each of the two-way comparisons was completed by 15 participants. As above, three catch trials for both emotion and lip-reading were used to automatically filter out uncooperative participants. Figure 7 reports the results of this study for both emotion and lip-sync. The study demonstrates that all design choices are critical. Perhaps the only surprising result is the similar performance of an EMOTE variant that uses a static per-frame emotion loss instead of the dynamic one. For more details, see the visual ablation study.

6.2. Qualitative Results

Comparison with SOTA: In Fig. 5, we qualitatively compare EMOTE with the SOTA methods. While all methods produce good lip articulation in accordance with the spoken words, none of these methods is able to produce emotional animations. FaceFormer trained on our MEAD training data can produce emotional faces, however, speech-content

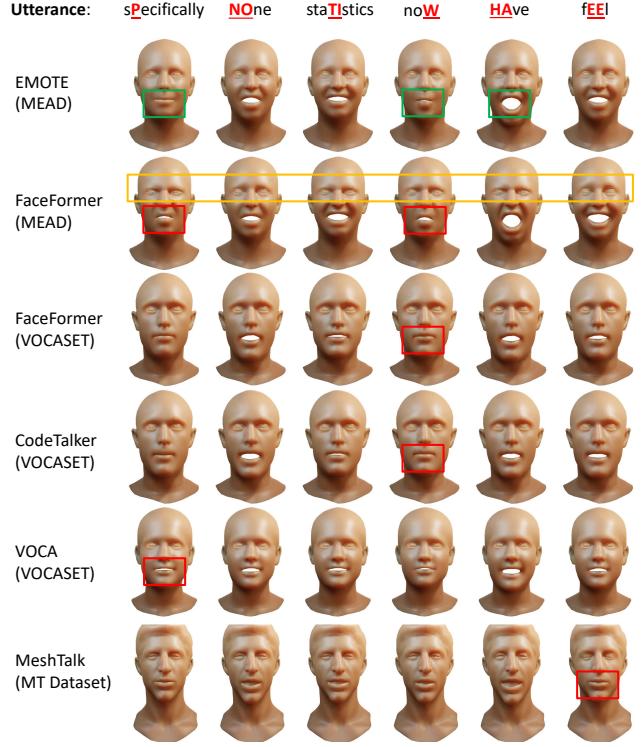


Figure 5. **Visual comparison to SOTA.** The rows show specific frames from a “neutral” sequence generated by EMOTE and baseline models in order: FaceFormer (trained on MEAD pseudo-GT), FaceFormer, CodeTalker, VOCA, and MeshTalk. The training dataset of each method is indicated in the brackets. The input characters of the utterances that were used to generate the animations are highlighted in red (top text description). Our model generates accurate animations of the lips (highlighted in green). FaceFormer trained on the pseudo-GT MEAD data, without emotion condition, struggles to produce a consistent “neutral” emotion given the neutral input speech (highlighted orange). Baseline models, while giving mostly good results, can sometimes also exhibit inferior lip-sync (highlighted in red).

and emotions are not well disentangled, and method lacks emotional control. This highlights the importance of our emotion-content disentanglement mechanism.

Ablation experiment: Figure 6 visualizes the effect of the individual design components. EMOTE (top row) produces accurate mouth shapes for various words and emotions. EMOTE w/o disentanglement terms (second row) starts to lose accurate lip-sync (for instance mouth closures on bilabials), especially during higher intensity emotions. EMOTE w/o the video emotion loss (third row) starts to lose some of the emotional cues as the only supervision signal that remains for emotion is through the geometry loss. EMOTE w/o the lip-reading loss (fourth row) suffers from inaccurate lip-sync. EMOTE w/o FLINT is temporally unstable and pro-

duces undesirable artifacts. EMOTE w/o any perceptual loss terms suffers from lack of emotional expressivity. EMOTE with a static emotion loss instead of the dynamic one has results that are comparable to EMOTE but sometimes suffers from undesirable artifacts such as eye closure. Finally, FaceFormer-EMO, a variant of the FaceFormer architecture augmented with emotion conditioning of Eq. 6, lacks both expressiveness and accurate lip-sync.

6.3. Emotion editing

EMOTE provides animators with emotion control. Figure 9 demonstrates this effect by editing emotions over the course of a sequence.

6.4. Limitations

High frequency speech: We have demonstrated that EMOTE produces emotional performances while maintaining lip-sync superior to the SOTA trained on the same pseudo-GT data. However, the model is not perfect, as it can fail with very fast high-frequency speech. This is due to our data collection process. While our pseudo-ground truth has good shapes that capture emotion, it is only sampled at 25fps. Using 3D scan data, with a higher sampling rate, or elaborate data augmentation, could produce more accurate results.

Eye blinks: EMOTE does not model eye blinks since those are only weakly correlated with speech and emotion and are hence difficult to capture with a deterministic method like EMOTE. Incorporating findings from [53] may help alleviate this limitation.

Paralinguistics: EMOTE does not model paralinguistic motion such as raising eyebrows on words that require emphasis since the lip-reading loss affects only the mouth. Solving this may entail incorporating language semantics (i.e., language models), a richer training set and non-deterministic prediction modelling.

Emotion granularity: EMOTE is capable of generating 8 basic emotions in various speaking styles corresponding to the number of training individuals. However, realistic emotion-induced motion can take on many more emotions and many more styles. Incorporating this would require training on large-scale datasets of sufficient richness and a more granular emotion model.

Mouth cavity: EMOTE and existing SOTA methods focus on the face shape and ignore the teeth and tongue, which can be important in speech perception.

Automatic emotion control: While EMOTE is capable of producing emotional faces, the emotion label must be provided by the user. This process could be automated by using automatic speech emotion recognition to provide the emotion condition.

7. Conclusions

We have presented EMOTE, the first framework to generate 3D talking head avatars with explicit control over the type and intensity of emotional expression. Unlike current SOTA methods that require high-quality scan datasets for training, EMOTE is trained from an emotional video dataset. Despite training on data without high-quality 3D ground truth, EMOTE’s lip-sync is of high quality, and better than that of SOTA methods trained on the same data. This is enabled by (co-)supervising the training with perceptual losses, i.e., a video emotion loss and a lip-reading loss, which give EMOTE an edge compared to the SOTA methods that are supervised solely with pseudo-GT geometry. Without high-quality 3D data, a geometric loss alone is insufficient. EMOTE’s loss terms ensure that the results carry emotional content as well as accurate lip articulation that is in accordance with the speech signal. A novel content-emotion exchange mechanism ensures that the lip articulation is driven by the spoken word and the expression is controlled solely by the specified emotion condition, effectively disentangling the two naturally entangled phenomena. To utilize the power of the perceptual losses without artifacts, we devise a temporal transformer-based VAE coined FLINT that operates on FLAME parameter sequences. We then use its decoder as our motion prior by mapping the speech features and the emotion condition into its latent space. Unlike the SOTA methods, EMOTE regresses FLAME expression and jaw pose, enabling more direct control over face shape by varying FLAME’s identity shape parameters. EMOTE makes use of a computationally efficient feedforward architecture. We believe that EMOTE opens an important and largely overlooked problem in the speech-driven animation field, i.e., that of emotional animation generation, and makes a considerable advance in that direction.

Acknowledgements: We thank Alpar Cseke, Taylor McConnell and Tsvetelina Alexiadis for their help with design and deployment of the perceptual study. We also thank Benjamin Pelkofer, Joan Piles-Contreras, Eugen Fritzler and Jojumon Kavalan for cluster computing and IT support. Finally, we express our gratitude to Anastasios Yiannakidis, Nikos Athanasiou, Peter Kulits and Maria-Paola Forte for proof-reading and valuable feedback. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No.860768 (CLIPe project).

Disclosure: Michael Black has received research gift funds from Adobe, Intel, Nvidia, Meta/Facebook, and Amazon. Michael

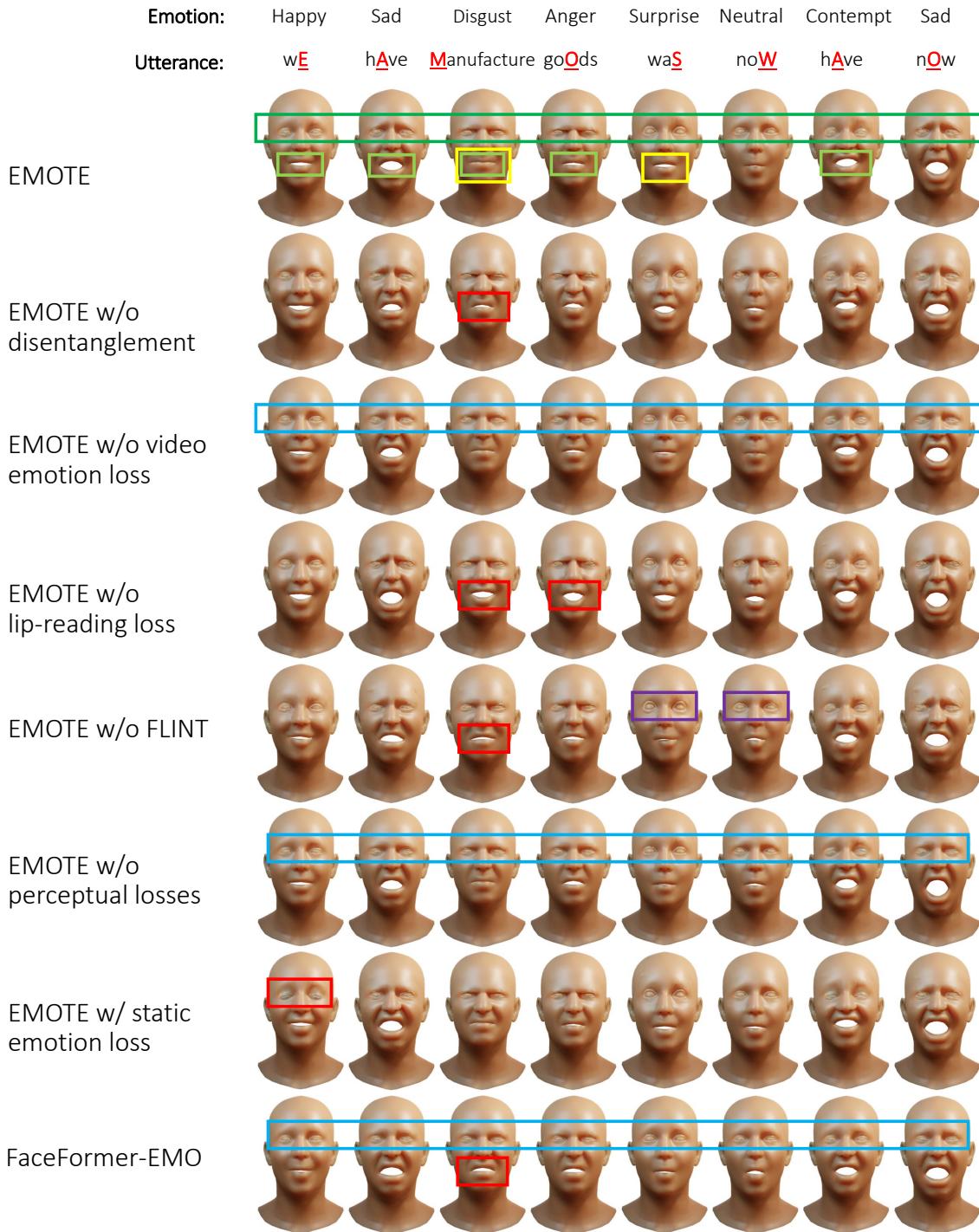


Figure 6. **Visual ablation study.** The rows show generations of the specified model for several spoken words highlighted above. Here we include different visemes and emotions taken out of context (not in sequence). Our model exhibits high emotional fidelity (highlighted green) and accurate lip motions (yellow). The effects of emotion are also visible in the lower part of the face (light green). Models without emotional supervision suffer from poor emotional fidelity (highlighted blue). Models without explicit lip-reading supervision often suffer from inferior lip-sync (such as incomplete mouth closure on bilabials). Finally, the model without FLINT yields uncanny artifacts (purple). Please watch the supplementary video to see the results in motion.

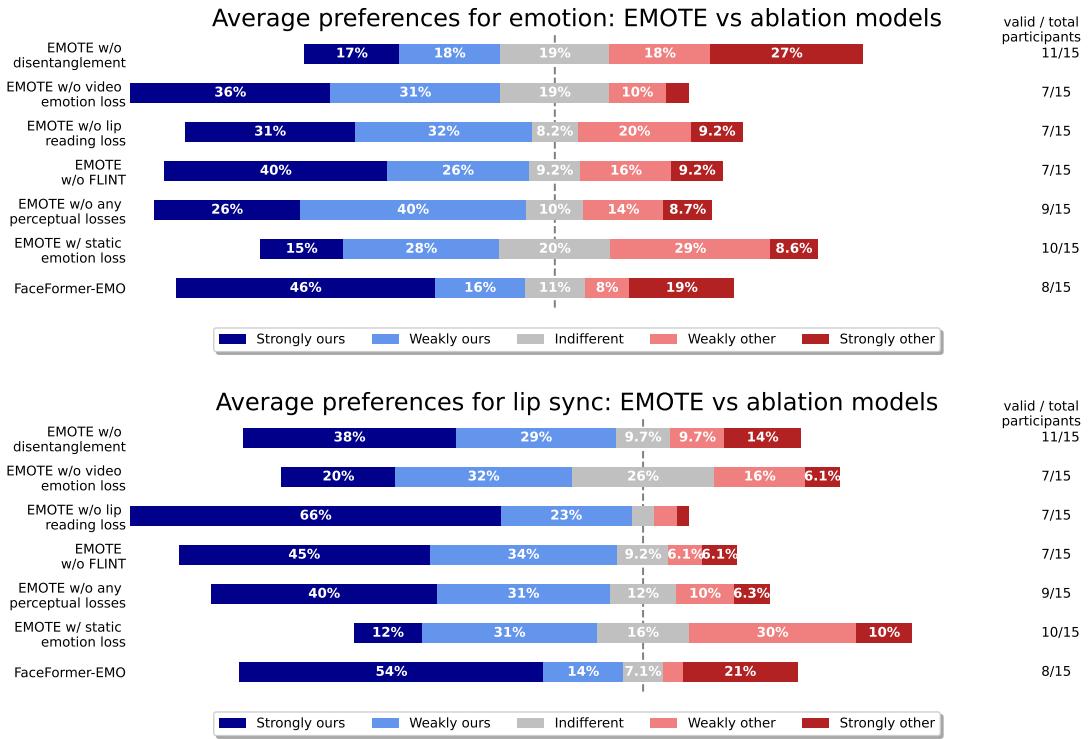


Figure 7. Ablation perceptual study results for emotion quality (top) and lip-sync quality (bottom). While participants prefer EMOTE w/o disentanglement on the emotion task its inferior articulation hurts the lip-sync preferences (see Fig. 6 and Sup. Video). EMOTE /w static emotion loss performs comparably on both metrics but occasionally results in artifacts (see Fig. 6 and Sup. Video). EMOTE is preferred on both tasks to all other ablated models.

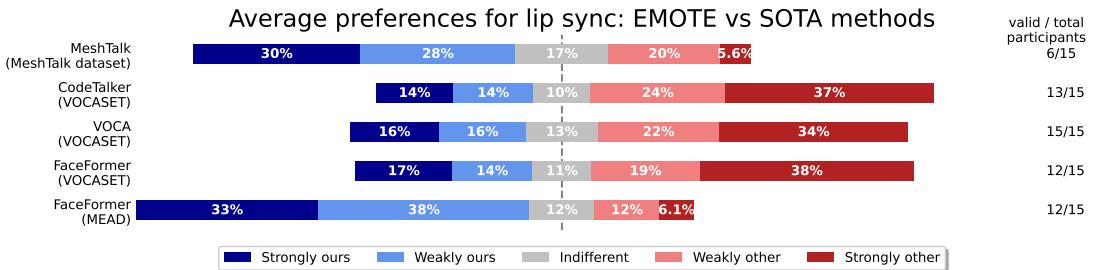


Figure 8. The results of the perceptual study comparing lip-sync of our method with the SOTA. The training dataset of each model is indicated in the brackets. The participants prefer EMOTE over MeshTalk. VOCA, FaceFormer and CodeTalker, which are trained on VOCASET, are preferred over EMOTE thanks to the superiority of the scanned training data. However, when we train FaceFormer on MEAD, its lip-sync preference is considerably lower, highlighting the benefits of our approach over SOTA methods on inferior data.

Black has financial interests in Amazon, Datagen Technologies, and Meshcapade GmbH. While Michael Black is a consultant for Meshcapade and Timo Bolkart a full-time employee of Google, their research was performed solely at, and funded solely by, the Max Planck Society.

References

- [1] Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8717–8727, 2022. 3
- [2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech

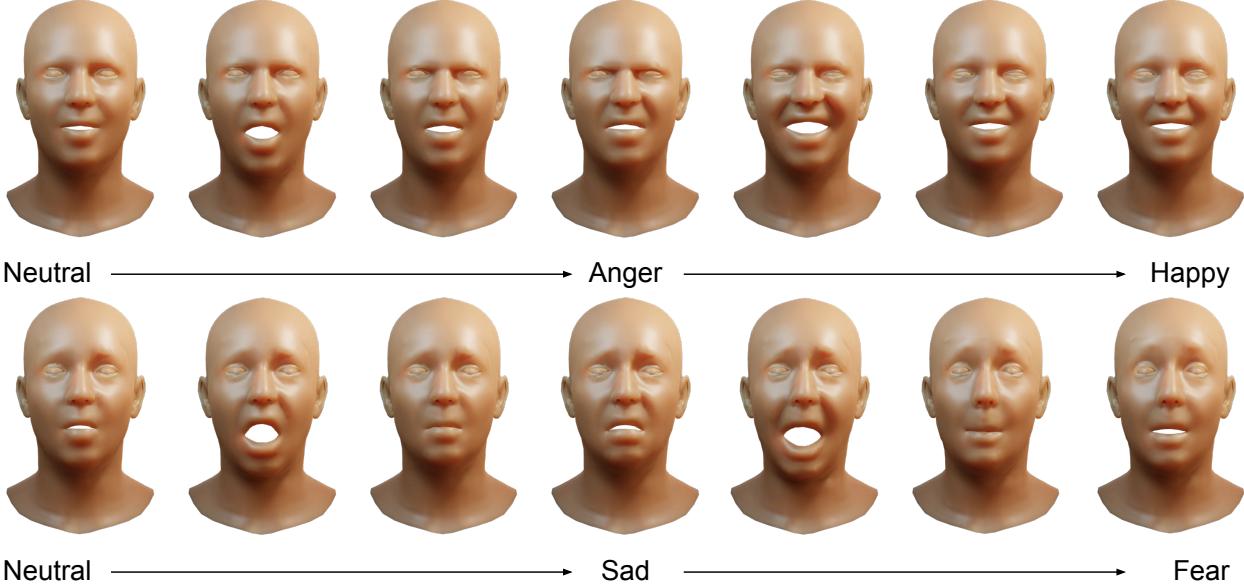


Figure 9. **Emotion editing.** The figure demonstrates that the emotion can be smoothly edited between emotions **during** speech. Despite the fact that EMOTE is trained with constant emotion labels, it can naturally generalize to edit emotion transitions in a single sequence, thus giving an animator the ability to direct the emotion and its intensity over the course of the animation as they see fit.

- recognition. *CoRR*, abs/1809.00496, 2018. 3, 7
- [3] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. Bringing portraits to life. *ACM Trans. Graph.*, 36(6):196:1–196:13, 2017. 3
 - [4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020. 4, 5
 - [5] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. 3
 - [6] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1021–1030. IEEE Computer Society, 2017. 7, 16
 - [7] Huawei Cao, David Cooper, Michael Keutmann, Ruben Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5:377–390, 10 2014. 3
 - [8] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an*,

- China, May 15-19, 2018, pages 67–74. IEEE Computer Society, 2018. 16
- [9] Yong Cao, Wen C. Tien, Petros Faloutsos, and Frédéric Pighin. Expressive speech-driven facial animation. *ACM Trans. Graph.*, 24(4):1283–1302, oct 2005. 2
- [10] Prashanth Chandran, Gaspard Zoss, Markus H. Gross, Paulo F. U. Gotardo, and Derek Bradley. Facial animation with disentangled identity and motion using transformers. volume 41, pages 267–277, 2022. 3
- [11] Lele Chen, Ross K. Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7832–7841. Computer Vision Foundation / IEEE, 2019. 3
- [12] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In Soon Ki Jung, Jehee Lee, and Adam W. Bargteil, editors, *SIGGRAPH Asia 2022 Conference Papers, SA 2022, Daegu, Republic of Korea, December 6-9, 2022*, pages 30:1–30:9. ACM, 2022. 3
- [13] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8185–8194. Computer Vision Foundation / IEEE, 2020. 3
- [14] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the In-*

- ternational Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 1086–1090. ISCA, 2018. 3
- [15] Michael M. Cohen, Rashid Clark, and Dominic W. Massaro. Animated speech: research progress and applications. In Dominic W. Massaro, Joanna Light, and Kristin Geraci, editors, *Auditory-Visual Speech Processing, AVSP 2001, Aalborg, Denmark, September 7-9, 2001*, page 200. ISCA, 2001. 2
- [16] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J. Black. Capture, learning, and synthesis of 3d speaking styles. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10101–10111. Computer Vision Foundation / IEEE, 2019. 2, 3, 7, 8
- [17] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: emotion driven monocular face capture and animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 20279–20290. IEEE, 2022. 2, 3, 4, 7, 16, 18
- [18] Stefano d’Apolito, Danda Pani Paudel, Zhiwu Huang, Andres Romero, and Luc Van Gool. GANmut: Learning interpretable conditional space for gamut of emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, June 2021. 3
- [19] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 285–295. Computer Vision Foundation / IEEE, 2019. 3
- [20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, 2019. 6
- [21] Hui Ding, Kumar Sriraman, and Rama Chellappa. Exprgan: Facial expression editing with controllable expression intensity. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6781–6788. AAAI Press, 2018. 3
- [22] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: An animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph.*, 35(4), jul 2016. 2, 3
- [23] Pif Edwards, Chris Landreth, Mateusz Popławski, Robert Małinowski, Sarah Watling, Eugene Fiume, and Karan Singh. Jali-driven expressive facial animation and multilingual speech in cyberpunk 2077. In *ACM SIGGRAPH 2020 Talks, SIGGRAPH ’20, New York, NY, USA, 2020*. Association for Computing Machinery. 2, 3
- [24] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3d morphable face models - past, present, and future. *ACM Trans. Graph.*, 39(5):157:1–157:38, 2020. 2
- [25] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18749–18758. IEEE, 2022. 2, 3, 4, 7
- [26] G. Fanelli, J. Gall, H. Romdorfer, T. Weise, and L. Van Gool. A 3-d audio-visual corpus of affective communication. *IEEE Transactions on Multimedia*, 12(6):591 – 598, October 2010. 2, 3
- [27] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)*, 40(4):88:1–88:13, 2021. 2, 3, 7, 16
- [28] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniu, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos, 2022. 2, 3, 7, 16
- [29] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8377–8386. Computer Vision Foundation / IEEE Computer Society, 2018. 3
- [30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 4
- [31] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14080–14089. Computer Vision Foundation / IEEE, 2021. 3
- [32] Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 2, 3
- [33] Hyeongwoo Kim, Mohamed Elgharib, Hans-Peter Zollöfer, Michael Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt. Neural style-preserving visual dubbing. *ACM Transactions on Graphics (TOG)*, 38(6):178:1–13, 2019. 3
- [34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7
- [35] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. 2014. 5
- [36] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2, 4

- [37] Alexandra Lindt, Pablo V. A. Barros, Henrique Siqueira, and Stefan Wermter. Facial expression editing with continuous emotion labels. volume abs/2006.12210, 2020. 3
- [38] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018. 3
- [39] Camillo Lugaressi, Jiuqiang Tang, Hadon Nash, Chris Mc-Clanahan, Esha Ubweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines. *CoRR*, abs/1906.08172, 2019. 7, 16
- [40] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahooz. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 4
- [41] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. pages 2616–2620, 2017. 3
- [42] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. pages 20363–20373, 2022. 5
- [43] Foivos Paraperas Papantoniou, Panagiotis Paraskevas Filintisis, Petros Maragos, and Anastasios Roussos. Neural emotion director: Speech-preserving semantic control of facial expressions in “in-the-wild” videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18759–18768. IEEE, 2022. 3
- [44] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. *CoRR*, abs/2303.11089, 2023. 3
- [45] Hai Xuan Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2328–2336. IEEE Computer Society, 2017. 2, 3
- [46] Hai Xuan Pham, Yuting Wang, and Vladimir Pavlovic. End-to-end learning for 3d facial animation from raw waveforms of speech. *CoRR*, abs/1710.00920, 2017. 2, 3
- [47] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David R. Traum, and Lluís Márquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics, 2019. 3
- [48] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia, MM ’20*, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery. 3
- [49] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 16, 18
- [50] Nikhila Ravi, Jeremy Reizenstein, David Novotný, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *CoRR*, abs/2007.08501, 2020. 6
- [51] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1153–1162. IEEE, 2021. 2, 3, 7
- [52] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *CoRR*, abs/1803.09179, 2018. 3
- [53] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. *Comput. Graph. Forum*, 34(6):299–326, sep 2015. 9
- [54] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7763–7772. Computer Vision Foundation / IEEE, 2019. 3
- [55] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV*, volume 12360 of *Lecture Notes in Computer Science*, pages 53–70. Springer, 2020. 3
- [56] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.*, 36(4):95:1–95:13, 2017. 3
- [57] Sarah L. Taylor, Taewan Kim, Yisong Yue, Moshe Mahler, James Krahe, Anastasio Garcia Rodriguez, Jessica K. Hodgins, and Iain A. Matthews. A deep learning approach for generalized speech animation. *ACM Trans. Graph.*, 36(4):93:1–93:11, 2017. 2, 3
- [58] Sarah L. Taylor, Moshe Mahler, Barry-John Theobald, and Iain A. Matthews. Dynamic units of visual speech. In Je-hee Lee and Paul G. Kry, editors, *Proceedings of the 2012 Eurographics/ACM SIGGRAPH Symposium on Computer Animation, SCA 2012, Lausanne, Switzerland, 2012*, pages 275–284. Eurographics Association, 2012. 2
- [59] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez,

- Michael Zollhöfer, and Christian Theobalt. FML: face model learning from videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10812–10822. Computer Vision Foundation / IEEE, 2019. 3
- [60] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2549–2559. Computer Vision Foundation / IEEE Computer Society, 2018. 3
- [61] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 3735–3744. IEEE Computer Society, 2017. 3
- [62] Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation, 2023. 2, 3
- [63] Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. 12361:716–731, 2020. 3
- [64] Soumya Tripathy, Juho Kannala, and Esa Rahtu. Icface: Interpretable and controllable face reenactment using gans. pages 3374–3383, 2020. 3
- [65] Soumya Tripathy, Juho Kannala, and Esa Rahtu. FACEGAN: facial attribute controllable reenactment GAN. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 1328–1337. IEEE, 2021. 3
- [66] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, volume 12366 of *Lecture Notes in Computer Science*, pages 700–717. Springer, 2020. 2, 3, 7, 17
- [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 M. Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shouo-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. volume abs/2207.11243, 2022. 2, 3, 8
- [68] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. pages 12780–12790, 2023. 2, 3, 7
- [69] Yuyu Xu, Andrew W. Feng, Stacy Marsella, and Ari Shapiro. A practical and configurable lip sync method for games. In Rachel McDonnell, Nathan R. Sturtevant, and Victor B. Zordan, editors, *Motion in Games, MIG '13, Dublin, Ireland, November 6-8, 2013*, pages 131–140. ACM, 2013. 2
- [70] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 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 2020, Seattle, WA, USA, June 13-19, 2020*, pages 598–607. Computer Vision Foundation / IEEE, 2020. 3
- [71] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016. 3
- [72] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor S. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9458–9467. IEEE, 2019. 3
- [73] Jie Zhang and Robert B. Fisher. 3d visual passcode: Speech-driven 3d facial dynamics for biometrics. *Signal Process.*, 160(C):164–177, jul 2019. 3
- [74] Yang Zhou, Zhan Xu, Chris Landreth, Evangelos Kalogerakis, Subhransu Maji, and Karan Singh. Visemenet: Audio-driven animator-centric speech animation. *ACM Trans. Graph.*, 37(4), jul 2018. 2, 3
- [75] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. 13667:650–667, 2022. 3
- [76] Wojciech Zienonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In Shai Avi-dan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIII*, volume 13673 of *Lecture Notes in Computer Science*, pages 250–269. Springer, 2022. 2, 3, 7, 16
- [77] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. *Comput. Graph. Forum*, 37(2):523–550, 2018. 3

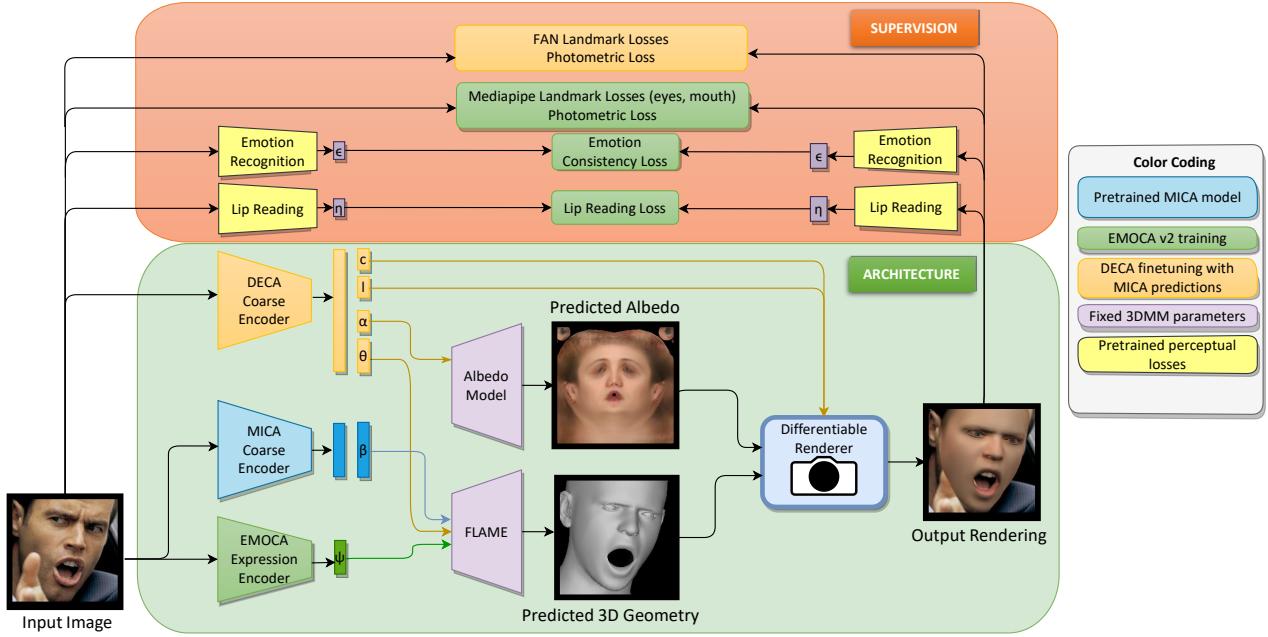


Figure 10. Face Reconstruction Architecture. The bottom part of the figure shows the architecture components. The top part shows the loss computation. The system is trained in stages, the trainable components of each stage and the corresponding losses are color-coded with the same colors (beige for DECA and green for EMOCA).

A. Additional Technical Details

A.1. Motion Prior: FLINT

Here we provide additional details about FLINT. Both transformers (in encoder and decoder) use a modified version of ALiBi [49], that allows the transformer to look into the future steps (as opposed to the past steps only, which is what the original ALiBi does). The ALiBi mechanism is preferred over additive positional encodings since it generalizes to arbitrary sequence lengths better [49]. Both convolutional blocks in encoder/decoder have three convolutional layers that temporally downsample/upsample the sequence. The bottleneck dimension is empirically set to 128. In training, λ_{rec} is set to 1000000 and λ_{KL} to 0.001, which makes the converged KL divergence term less than one order of magnitude lower than the reconstruction terms. The model code and weights will be made publicly available.

A.2. Face Reconstruction Network

In this part we describe our face reconstruction network used to generate pseudo-GT for MEAD. In order to obtain the highest quality possible, we employ a combination of four SOTA in-the-wild face reconstruction methods, each of which tackles a particular aspect of the problem. Specifically, we augment DECA [27] with MICA [76], the SOTA on face shape prediction. Next, we utilize the expression prediction of EMOCA [17] to get the SOTA quality of facial expressions and emotions. Finally, in training we incorporate the lip-reading loss term from SPECTRE [28] in order to produce the SOTA-level of lip articulation.

Architecture: The architecture is depicted in Fig. 10. The input image is passed through all three encoders (MICA, DECA and EMOCA). MICA’s encoder is used to output the facial shape vector β . DECA’s encoder predicts the rest of the parameters: camera c , spherical harmonics coefficients for lighting I , albedo coefficients α , global head pose and jaw pose θ . EMOCA’s encoder predicts the facial expression coefficients ψ . With the regressed predictions, we can now reconstruct the geometry and render an image which can be used for supervision.

Training: We finetune the individual components of the aforementioned architecture in stages.

(1) **Finetuning DECA:** In the first stage, the DECA [27] coarse encoder (beige) is finetuned from the original released DECA model on VGGFace2 [8], the same dataset as the authors. The difference from the original implementation is that we take MICA’s prediction for the facial shape vector β . The encoder of MICA remains frozen. DECA’s encoder predicts the rest of the parameters: c , I , α , θ and also the expression coefficients ψ .

(2) **Training EMOCA:** In the second stage, we train the EMOCA expression encoder similarly to Danecek et al. [17], discarding DECA’s expression predictions. However, there are a few differences. Instead of employing FAN [6] landmarks, we make use of Mediapipe [39] landmarks. We only use the landmarks and not the face contour since the face contour does not affect the expression. Compared to FAN landmarks, the eye and mouth landmarks of Mediapipe are more accurate. In addition to the photometric and emotion consistency loss from the EMOCA authors, we also employ the lip-reading loss from SPECTRE. This stage is akin to the EMOCA v2 the authors released but upgraded with MICA for

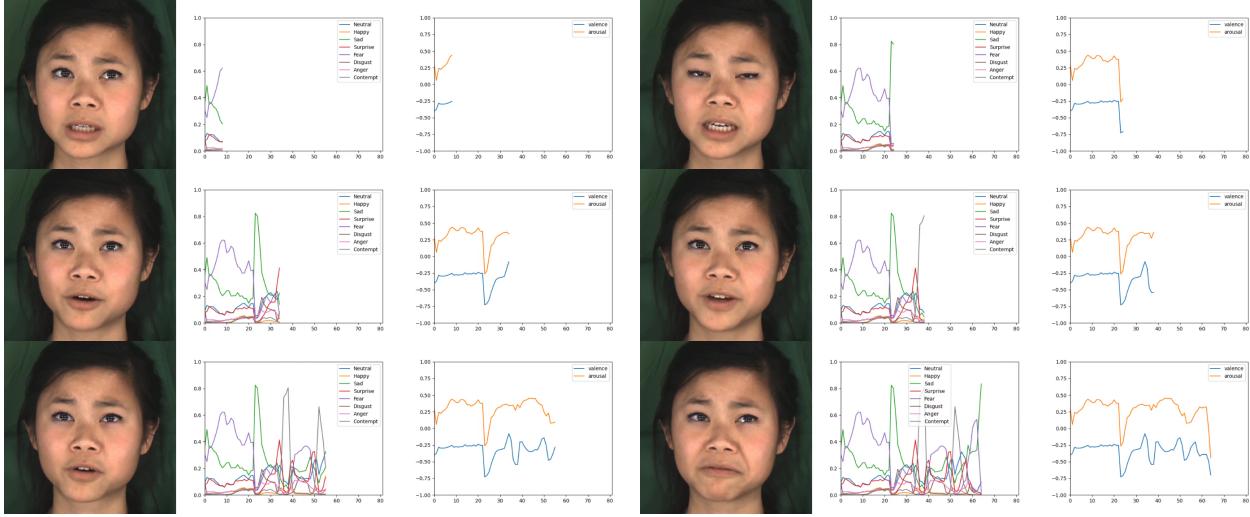


Figure 11. Static Emotion Recognition. In the figure you can see the input images taken out of one of the videos of a sad person from the MEAD dataset (left-hand side) and the outputs of a static emotion recognition system - the basic expression classification (middle) and valence and arousal (right-hand side) over the course of a video, with x-axis representing the temporal dimensions and y-axis the probability of the emotion class (middle) and valence, arousal (right). The person has a constant level of sadness throughout the video. Despite that, the static emotion recognition yields very different classification for individual frames (fear, sadness, surprise, contempt and neutral).

shape prediction.

(3) Finetuning on MEAD: In the final stage, we finetune EMOCA with the same losses as in the previous stage on MEAD in order to get the most accurate MEAD pseudo-GT possible.

A.3. Limitations of Static Emotion Recognition

Static (single-image) emotion recognition is not stable. It can output many different classification results over the course of a video of a person talking under a single emotion as shown in Fig. 11. Even to a human, single frames can be misleading because temporal context is missing and speech-induced facial expressions can lead to misinterpretation. Static emotion recognition suffers from the same limitation. This limitation can be lifted, when considering emotions as a temporal phenomenon. For this reason, we opt to train a video emotion recognition network that is able to leverage contextual information. And then utilize the sequence aggregated features for our video emotion perceptual loss.

A.4. Video Emotion Recognition

Our implementation of the video emotion networks is a lightweight single-layer transformer network with two classification prediction heads - one to classify the emotion and one to classify the intensity. It takes a sequence of static emotion recognition features extracted from a video on the input $\epsilon^{1:T}$, passes it through a transformer encoder to get the video emotion feature ϕ . The video features is then used to classify the emotion class and intesity with linear classification prediction heads. We train this network on the MEAD dataset with the standard cross-entropy classification losses for both classification tasks (emotion and intensity) until convergence. Ground truth labels for both emotion and intensity are provided with the MEAD dataset. The architecture is depicted in Fig. 12. The model code and weights will be made publicly

available.

A.5. Dynamic vs Static Emotion Loss

Fig. 13 compares the performance of the static and dynamic emotion losses. The video emotion classification is significantly superior to the static emotion classifier.

B. Datasets

Existing datasets: Table 1 provides an overview of existing 3D and 2D face datasets with synchronized audio. While there is seemingly many datasets available, each of them come with a particular set of challenges (such as low video quality, too difficult for face reconstruction, not enough variety of emotions and speaking styles, small size of the dataset etc.). Taking all of the above into account, we opt for using MEAD [66] in our experiments, since it

Table 1. Datasets

Datasets	Modality	Number of subjects	Expressions	Duration
BIWI	3D	14 (8F, 6M)	11	1.43 h
VOCASET	3D	12 (6F, 6M)	–	0.48 h
S3DFM	3D	77 (27F, 50M)	–	0.28 h
Multiface	3D	13	65 (v1), 118 (v2)	–
VoxCeleb1	2D	1251 (561F, 690M)	–	352 h
VoxCeleb2	2D	6112 (2351F, 3761M)	–	2442 h
Faceforensics++	2D	–	–	–
CelebV-HQ	2D	15653	83	68 h
LRS2-BBC	2D	–	–	224.5 h
LRS3-TED	2D	5594	–	438 h
RAVDESS	2D	24 (12F, 12M)	8	–
CREMA-D	2D	91 (43F, 48M)	6	–
MELD	2D	407	7 (+ 3 sentiments)	12.96 h
CMU-MOSI	2D	98	sentiment intensity [-3,3]	2.6 h
CMU-MOSEI	2D	1000	6 (+ 5 sentiments)	65.88 h
TalkingHead-1KH	2D	–	–	1000 h
MEAD	2D	60 (30F, 30M)	8	38.95 h

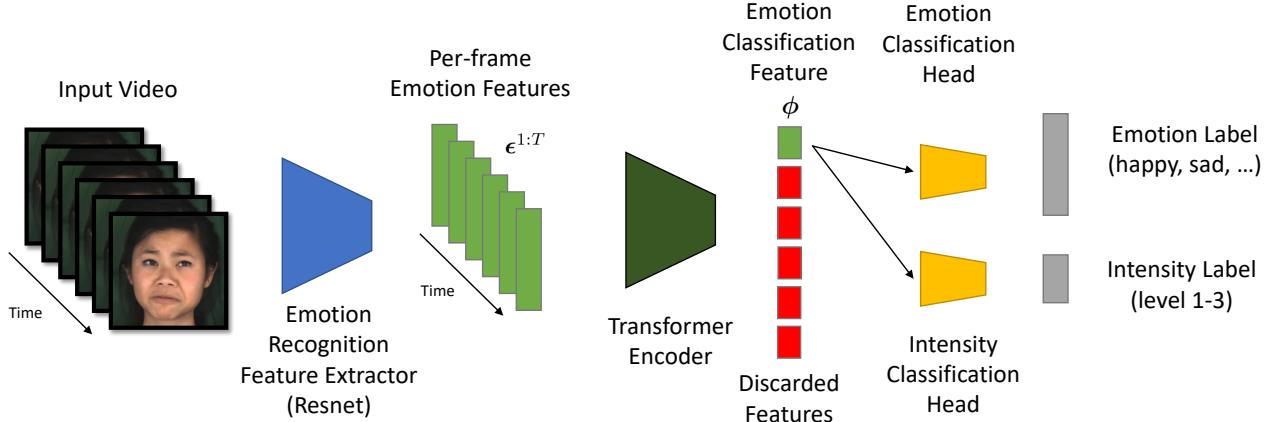


Figure 12. Video Emotion Recognition Architecture. The input video is passed frame by frame into the pretrained and frozen emotion recognition net of Danecek et al. [17] obtaining a $2048d$ emotion feature for each frame $\epsilon^{1:T}$. These are passed into a single-layer transformer encoder to extract a single $256d$ video emotion feature ϕ . Instead of additive positional encodings we opt to use the ALiBi [49] mechanism and we modify it, such that it also considers future frames. This feature is then fed into linear classification heads to classify the emotion and the intensity.

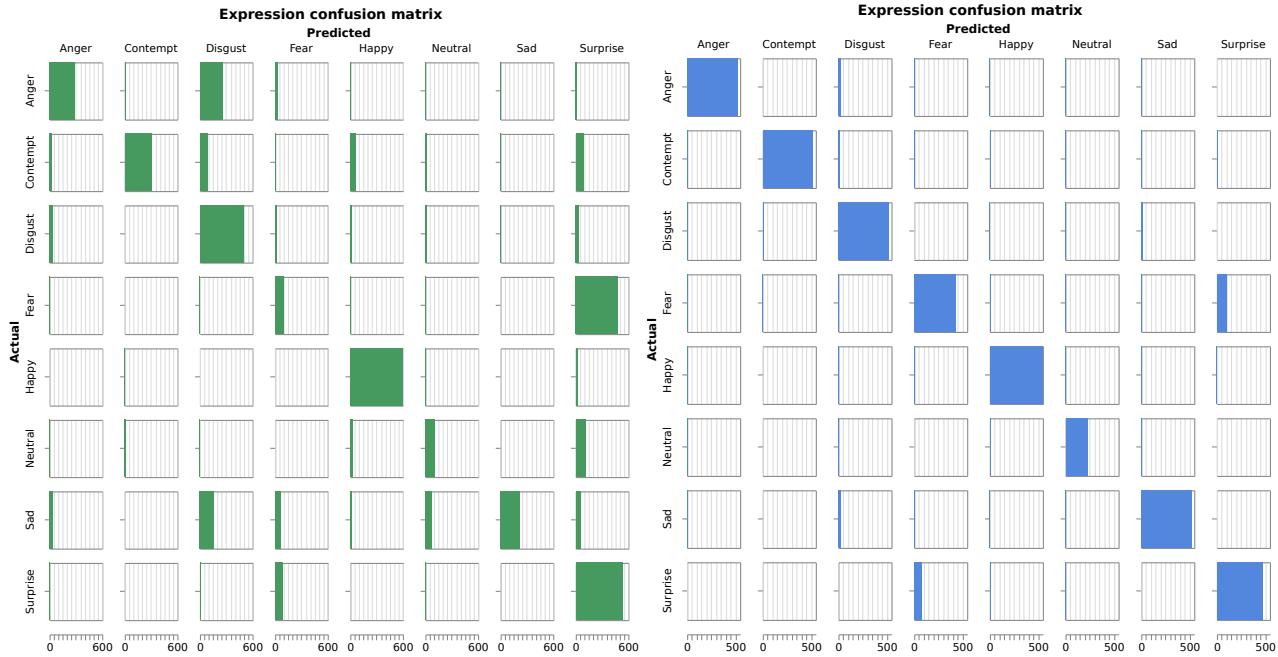


Figure 13. Classification performance of static vs dynamic emotion loss. Here we show two confusion matrices on the MEAD validation set, the one of static emotion recognition network (left) and a dynamic emotion recognition network (right). For static emotion classification results we took the most occurring classification in each video. The accuracy of static emotion recognition network is 57.9 % and the dynamic is 90.8%.

is of sufficient scale, has all emotions in different intensities and thanks to its constrained environment and high quality video, it is relatively easy to reconstruct. It is also considerably easier to reconstruct compared to the other datasets since it is a captured in a lab and hence does not come with the problems of associated with reconstruction in-the-wild (such as occlusions, difficult illumination etc.).

C. Perceptual studies

Here we provide additional detail about our perceptual studies.

Lip articulation evaluation: Fig. 14 show the exact layout used for this perceptual study.

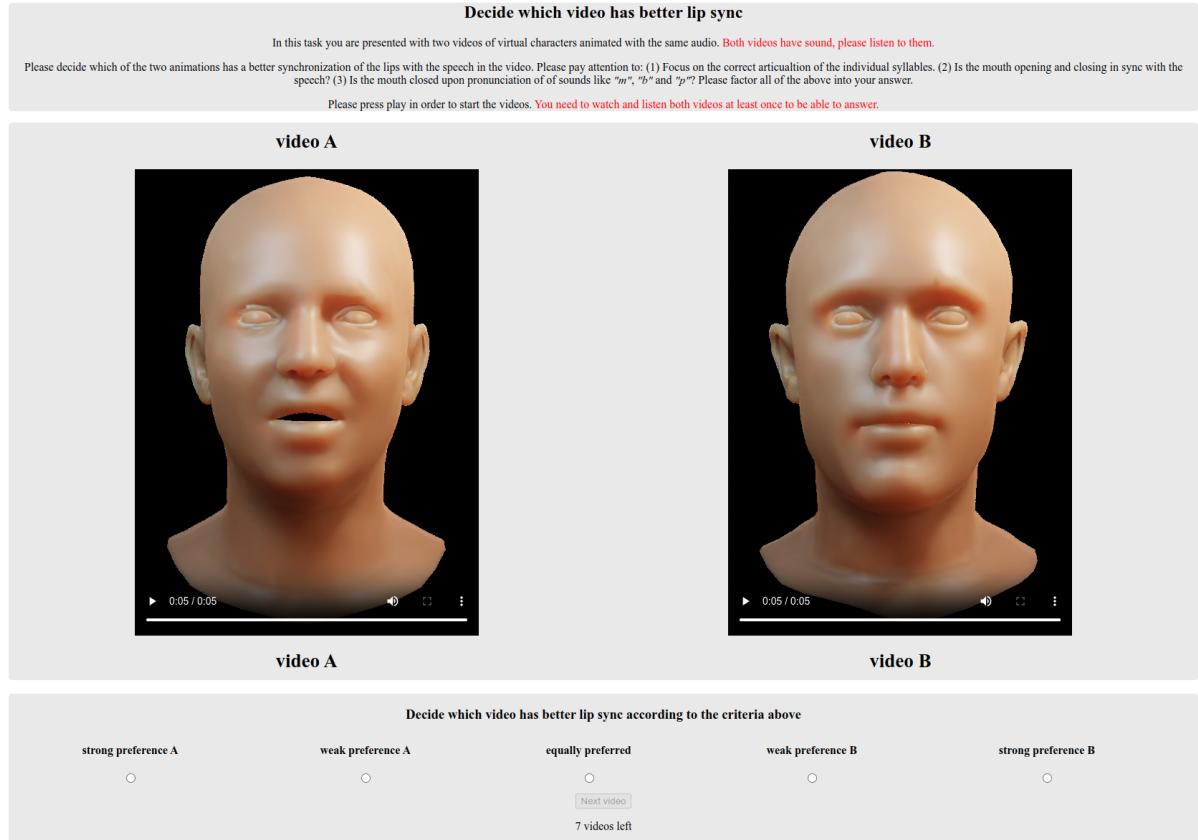


Figure 14. **Lip-sync Comparison with SOTA.** The participants are presented with two videos *with audio*, one generated using EMOTE and others using SOTA methods. The participant is asked to judge the quality of synchronization of the lips with the speech in the video.

Ablation perceptual study: Figures 16 and 15 show the web layout of the emotion quality and lip-sync quality ablation studies respectively. The two layouts alternate after each response.

D. Additional Results

Fig. 17 and Fig. 18 demonstrates the ability of EMOTE to generate emotional animation. A different speaking style was used for each of the two figures.

Lip Sync and Emotion analysis

In this task two questions will alternate:

#1 subtask: you are presented two videos of a virtual character **without sound**. Please decide which of the two characters communicates better the emotion specified below the videos.

#2 subtask: you are presented the same two videos but now **with sound**. Please decide which of the two animations has a better synchronization of the lips with the speech in the video. Please pay attention to: (1) Focus on the correct articulation of the individual syllables. (2) Is the mouth opening and closing in sync with the speech? (3) Is the mouth closed upon pronunciation of sounds like "m", "b" and "p"? Please factor all of the above into your answer.

Please press play in order to start the videos. You need to watch (and listen) both videos at least once to be able to answer.

video A



▶ 0:00 / 0:05

video B



▶ 0:00 / 0:05

video A

video B

#2 subtask:

The emotion the characters are supposed to show is: **disgust**. Which of the two characters **has better lip sync and articulates more in correspondence with the audio?**

strong preference A

weak preference A

equally preferred

weak preference B

strong preference B

Figure 15. **User study: Lip-sync Analysis.** The participants are presented with same two videos as in Fig. 16. This time the videos are *audible* and the user is asked to judge the quality of the articulation, taking the audio into account. Again, the participant must watch both videos in full length before being able to proceed to the next question. Upon answering, the user is redirected to the first question (emotion quality assessment) with a new pair of videos.

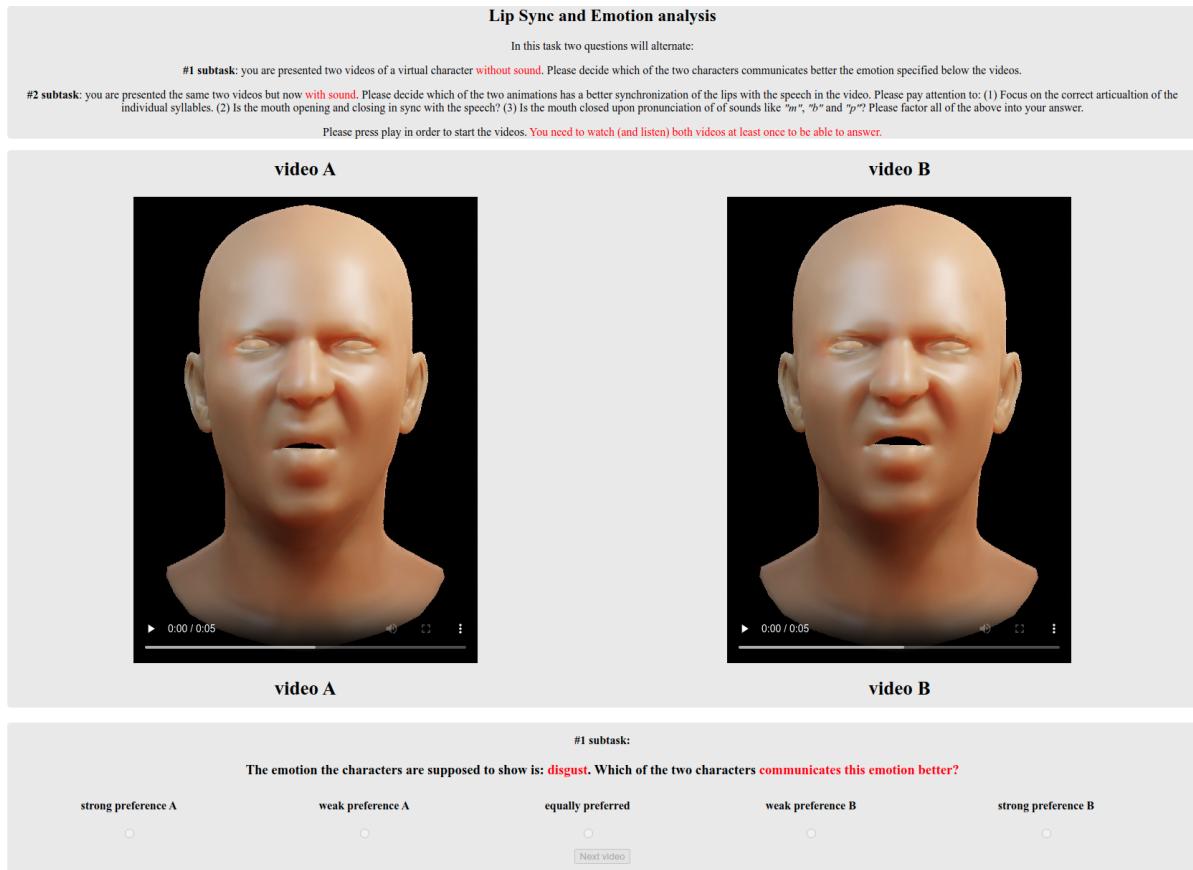


Figure 16. **Emotion quality assessment.** The participants are presented with two *muted* videos and are asked to select which of the two videos better communicates the emotion specified in the text under the video. The participant must watch both videos in full length at least once before submitting an answer becomes available and then they proceed to the next question about lip-sync assessment for the same two videos (see Fig. 15).

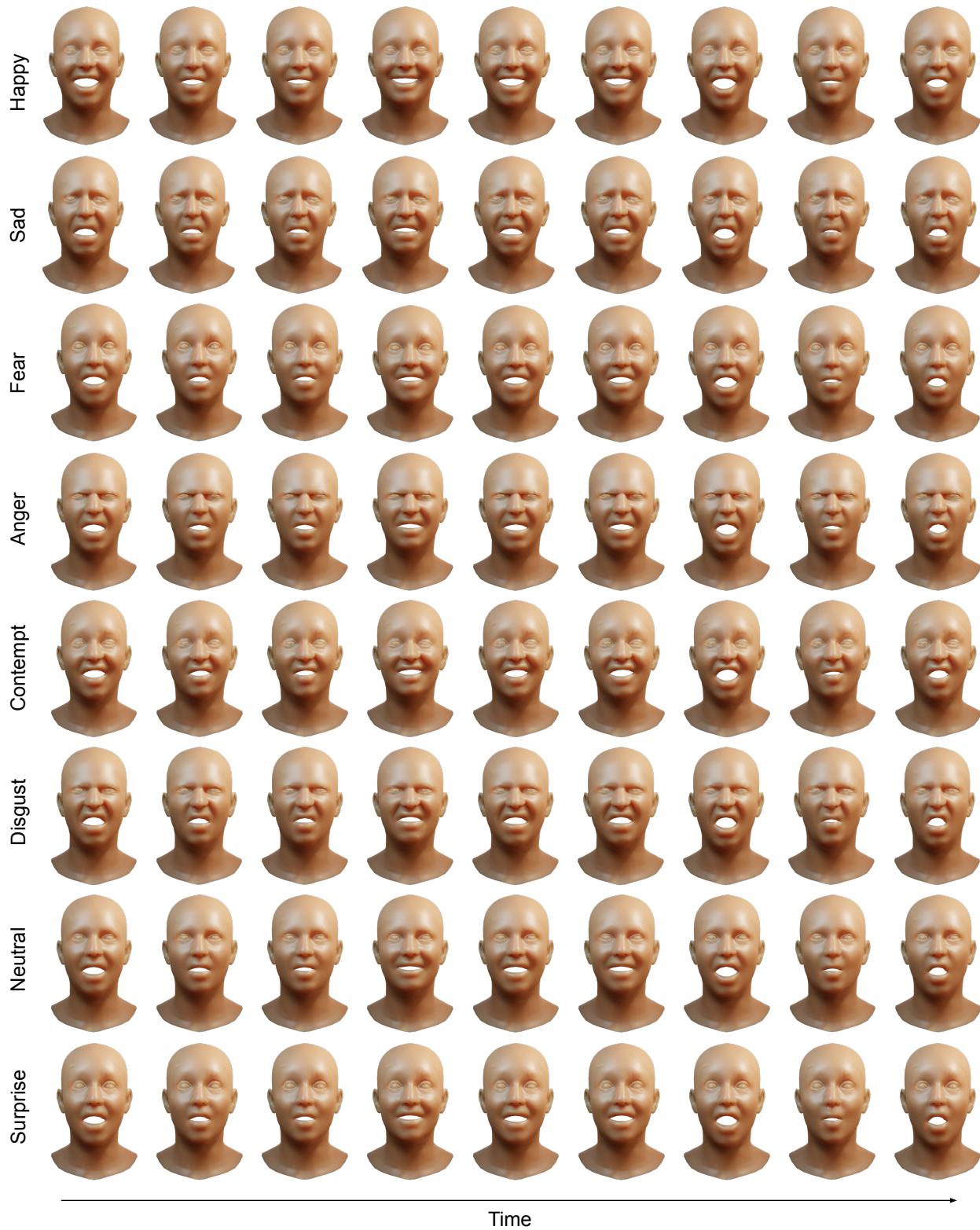


Figure 17. Additional results with the same input audio but different input emotion.

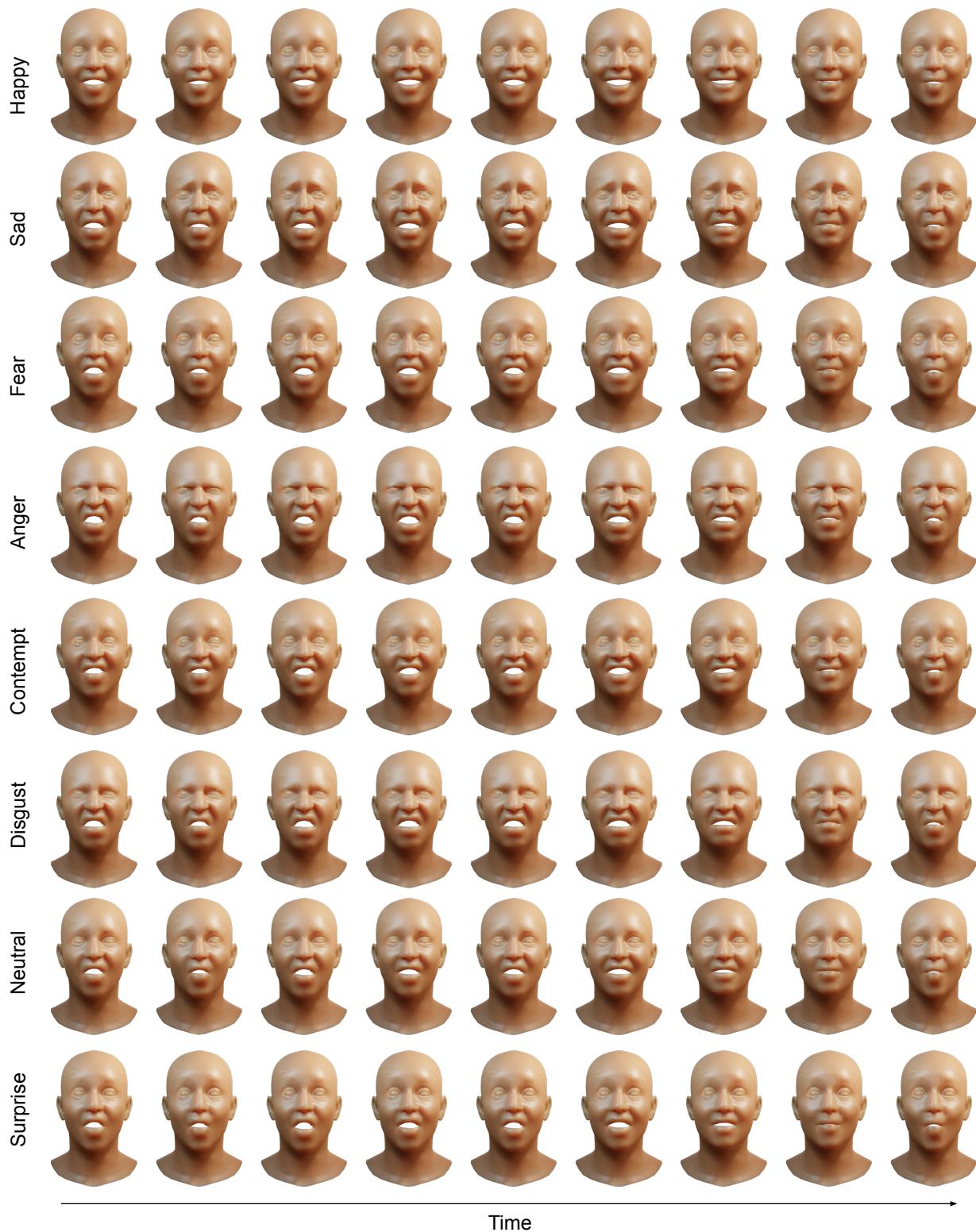


Figure 18. Additional results with the same input audio but different input emotion.