Intuitive Speech-Driven Face Animation

Finn Rasmus Schäfer
finn.schaefer@tum.de

Haifan Zhang
haifan.zhang@tum.de

Valentina Miller
valentina.miller@tum.de

Yuxin Guo
yuxin.guo@tum.de

Abstract

Facial animation is a key area of research, aimed at generating realistic human-like facial shapes, expressions andS motions. Traditional approaches are often labor-intensive and expensive, while newer methods utilizing deep learning and complex 3D face models have shown promising results. This paper introduces two novel approaches to enhance intuitive speech-driven facial animation pipelines based on deep learning, hence focusing on emotional expressions. The first contribution extends the renowned VOCA (Voice Operated Character Animation) framework [3]. An extension is proposed that allows for overcoming the lack of emotional information through a post-processing layer. Additionally, a dataset with emotional content is generated to retrain the model, aiming for it to learn emotional patterns. However, the results show that the model is not capable of learning to represent emotional expressions. Secondly, EmoFormer is introduced, a 3D facial animation model that employs a novel network structure involving an encoder with disentangled emotion and content spaces and a transformer-based decoder. To address the problem that existing datasets do not contain emotion information, A dataset of 3D emotional talking faces is constructed using EMOCA [4] and RAVDESS [13]. The EmoFormer's superiority in terms of realistic mesh deformations, intuitive lip movements, emotional expressions, and generalization to unseen data indicates that this architecture is well-suited. Hence, EmoFormer allows for intuitive speech-driven facial animations.

1. Introduction

Facial animation is a heavily researched task within computer graphics, focusing on animating human-like 3D facial shapes, facial expressions, and facial motions. The field is motivated by its various applications such as 3D animations, online communication, video games, and user interfaces. However, the task remains unsolved due to the absence of a straightforward representation for human-like attributes [4].

Traditional approaches are mostly of a highly labor-intensive nature. For instance, face rigging is a traditional approach which refers to the manual refinement of the position for all faces' rigs, a highly time consuming process. Hence, these approaches are very expensive [3]. Further research resulted in new approaches, among other things, face capturing by depth sensors. However, without the availability of 3D face models that are complex enough to capture the facial characteristics in detail, only poor results can be achieved through this approach [11]. Altogether, traditional approaches are not sufficient to solve the task of facial animation in an economical and precise manner [11].

The introduction of two complex face models, Faces Learned with an Articulated Model and Expressions (FLAME) and Basel Face Model in 2017, opened up new possibilities [8, 11]. Gathering data by scanning numerous faces and representing them within these model formats allowed for the employment of deep learning [4]. While good results on facial shapes and facial movements can be achieved by those models [8, 11], research on facial expressions is still ongoing. One possible approach is speech-driven facial animation, which stands in contrast to traditional methods like face rigging. By leveraging a fully automated pipeline, speech-driven animation not only simplifies the creation of 3D content but also offers the advantages of producing natural, nuanced expressions directly from speech input, ensuring efficiency, and enabling dynamic interaction in applications such as virtual avatars and real-time experiences. Latest research has shown that by extracting information from the audio data, additional, valuable knowledge can be obtained, resulting in more human-like, and therefore, more intuitively-seeming facial expressions. This can improve deep learning algorithms for facial animation [3]. As there are still limitations in performance for speech-driven facial animation, the most state of the art approaches like VOCA [3] and Facformer [6] lack the incorporation of emotional expressions, consequently yielding less natural outcomes. this work aims to further explore two promising approaches in order to achieve better

and more intuitive results with a special emphasis on emotional content. The main contributions of the paper are as follows:

1. While the speech-driven facial animation framework *VOCA* is able to animate facial shapes and facial movements well, the results miss intuition due to a lack of emotional information in the animations [3]. Hence, the first contribution of this paper is an extension of *VOCA* to allow emotional display.

2. A transformer architecture to disentangle content and emotion information. By individually training the encoder and decoder, this paper expects to achieve satisfactory results.

2. Related works

2.1. Parametric Face Model

The parametric face models has been a pivotal research area within computer vision and computer graphics. These models aim to capture and present the human facial appearance and expression using a set of underlying parameters with lower dimensionality compared to the mesh vertices. Thus they are widely used in a range of fields, including facial animation, virtual reality and game industry.

Early work of parametric face models can be tracked back to Blanz and Vetter's seminal work on 3D Morphable Models (3DMMs) [2]. Their model was established through statistical analysis of facial scans from diverse individuals. This approach represented facial shapes and textures as lower-dimensional vectors, significantly reducing storage and computational demands in contrast to full 3D models. In certain studies [7, 12], the integration of blendshapes has been proposed to address limitations arising from the global attributes of PCA. These introduced blendshape models enable the generation of facial poses through linear combinations of distinct expressions. Consequently, this approach facilitates localized expression control by adjusting blendshape weights.

In recent developments, the FLAME model [11] separates the expression of identity, pose, and facial expression. It's built from thousands of 4D facial scans and notably includes the neck area. In the FLAME model, there are specific blendshapes used to control the neck and jaw movements, making it particularly useful for tasks like facial animations.

2.2. Speech-Driven 3D Facial Animations

Speech-driven 3D face animation is a field of research that aims to generate realistic facial animation from speech signals. In recent years, numerous research have been conducted on 2D-based talking face generation [9, 14], while we focus on animating 3D face in this project. In 3D facial animation, the procedural methods break down facial movements into small units and create a set of explicit rules for mapping speech to facial motions. For example, JALI [5] draw from psycholinguistics, using 2 visually distinct anatomical actions to control jaw articulation and lower-face muscles. However they require extensive manual labor. Alternatively, various data-driven methods have emerged. VOCA [3] leverages temporal convolutions to produce speaker-independent 3D facial animation, yet the face motions are only present in the lower face. Meshtalk [15] learns a categorical latent space for facial animation that disentangles audio-correlated and audio-uncorrelated information. Although MeshTalk focuses on the upper part of the face, which is lacking in VOCA, the facial emotion remain insufficient. FaceFormer [6] uses a transformer-based model to predict facial movements autoregressively, the problem of lack of emotion is still unsolved because it also use VOCASET dataset which contains no emotional data.

2.3. Transformer in Vision

While transformer [16] has become the highest standard for natural language processing tasks, it was also widely used in computer vision. Unlike RNN that has a sequential processing flow, transformer can process entire sequence in parallel, which allows it to capture long-range contextual information. Lots of the recent works on 3D sequence synthesis have explored the power of transformer. FaceFormer [6] leverages transformer to encode the long-term audio context and predicts a sequence of animated 3D face meshes autoregressively.

3. Methods

In this chapter, two extensions of existing speech-driven facial animation pipelines are introduced that were developed and researched with the intention of allowing for emotion display. Firstly, an extension of the speech-driven facial animation framework *VOCA* is proposed. By adding a post-processing layer to the original pipeline and further, training on the thereby generated data, this approach aims to capture emotions. Secondly, EmoFormer, an extension of FaceFormer [6] is presented. By leveraging the feature of transformer, we formulate speech-driven 3D facial animation as a sequence-to-sequence learning problem and generate face movements conditioned on audio signal and previous predictions.

3.1. VOCA Approach

VOCA proposes a realistic speech-driven animation of faces based on audio inputs and a 3D template of the face, see Fig. 1. The approach is chosen due to its outstanding performance for speech-driven facial animation [3].

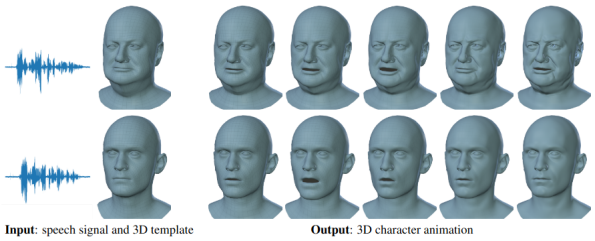


Figure 1. Illustration of *VOCA's* inputs and outputs

3.1.1 VOCA Pipeline

The pipeline proposed by VOCA takes a subject-specific template and a raw audio signal as inputs and maps these to a sequence of target 3D meshes. In order to achieve this, an encoder is employed that extracts audio features from the input as a first step. This step is achieved by utilizing DeepSpeech to obtain low-dimensional embeddings. These embeddings capture the relevant features from any audio signal and therefore, allow for a concise representation of the audio input. A decoder is then employed to map the resulting embeddings to a high-dimensional space of 3D vertex displacements, representing the offsets of the 3D vertices of the original 3D template and therefore, representing the facial movements. Finally, this resulting sequence of offsets is added to the subject-specific template and is displayed as a character animation. For an overview of the *VOCA* pipeline, refer to Fig. 2. The key advantage of the pipeline is the high accuracy for the mouth movements for various identities, even for different languages [3].

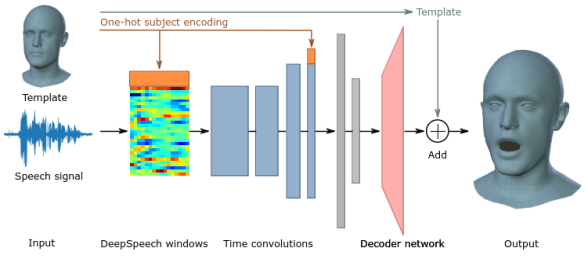
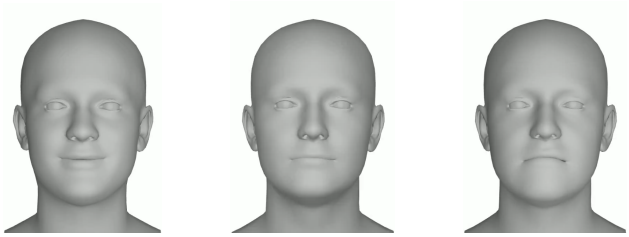


Figure 2. An overview of the *VOCA* Architecture

However, the absence of any upper face movements, apart from eye blinking, implies that the pipeline is incapable of any emotional display [3]. This lack of emotion makes the approach less human-like and, hence, less intuitive as the display of emotion while speaking is a natural aspect for human beings. Aiming to allow for emotional display, this work explores two ways to overcome this constraint. First, a layer of post-processing is added to the pipeline. In this layer, emotions are extracted from the audio input and statically added to the output of the *VOCA* pipeline. Further, this technique is the basis of the second proposal to overcome the lack of emotions. It suggests generating a dataset with emotional content with the help of the post-processing step. The *VOCA* architecture is then retrained on the obtained dataset with the intention that the model learns the patterns of emotions and consequently, displays emotions in its output.

3.1.2 Static Approach

Similarly to the *VOCA* pipeline, this approach takes a subject-specific template and a raw audio signal as input and maps these to a sequence of target 3D meshes. However, in contrast to the original pipeline, the output of the *VOCA* model is further processed. For this, the audio input should contain emotional cues in speech. As with the help of a speech emotion recognition model by SpeechBrain[1], the emotion of the speaker is classified during the post-processing. The model distinguishes between nine different emotions, namely angry, excited, fear, sad, surprised, frustrated, happy, disappointed and neutral. For each of the nine emotions an emotion template based on FLAME is generated, for an example refer to Fig. 3. Hence, the recognized



(a) Happy (b) Neutral (c) Sad

Figure 3. Exemplary emotional templates

emotion can be mapped to its FLAME parameters and then be added as an offset to the output of the *VOCA* model.

3.1.3 Retraining

The *VOCA* dataset, as introduced in Sec. 2, on which the proposed architecture is trained does not include emotional display. Hence, by retraining the pipeline on a dataset that includes emotional content, this work aims to achieve a model which is capable of emotional display. In order to achieve such a model, first a dataset must be prepared that includes such content.

Data Generation The same approach, as proposed as a static solution, can be used to generate a dataset with emotional content and hence, a dataset from which the model can learn emotions. By linking each input audio to the, by

[1]https://speechbrain.github.io/, accessed on the 20th of August 2023

emotions post-processed, output of the *VOCA* model, a new data set is obtained. As inputs the emotional speech and song audios of the RAVDESS dataset [13] are used. More precisely, the *VOCA* model is utilized to obtain the correct mouth movements from emotional audio input. The resulting sequences, which do not include any emotional display, are then post-processed by the static approach, namely by post-processing the obtained sequences with the extracted emotions of its corresponding audio input. The resulting dataset is formatted in the same way as the *VOCA* dataset which implies that it can be used for the training without any additional processing which is further explored in Sec. 4.1.2.

Loss function In accordance with the *VOCA* pipeline [3], the loss applied for the training consists of two parts. Firstly, the position term is introduced to minimize the distance between the predicted outputs and the ground truth:

$$E_{\mathrm{p}} = ||\mathbf{y_i} - \mathbf{f_i}||_F^2 \tag{1}$$

Secondly, the velocity term is introduced to minimize the differences between consecutive frames of predicted outputs and training vertices in order to achieve smooth transitions over time:

$$E_{\mathrm{v}} = ||(\mathbf{y_i} - \mathbf{y_{i-1}}) - (\mathbf{f_i} - \mathbf{f_{i-1}})||_F^2 \tag{2}$$

3.2. EmoFormer

We propose a 3D facial animation model that can generate emotional face movements from audio signals. It takes raw audio sequence as input and output tracked FLAME mesh.

3.2.1 Dataset construction

A key challenge is that there is no publicly available 3D talking face data with emotion. To address this issue, we use the state-of-the-art deep learning method to generate 3D face sequences from existing 2D emotional audio-visual dataset. Specifically, we employ EMOCA [4] to reconstruct emotional 3D face sequences of the RAVDESS [13] dataset. The RAVDESS is a mutlimodal database of emotional speech and song. The database is consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent with 8 different emotions. EMOCA is an approach to reconstruct animatable 3D face from in-the-wild images or videos. By introducing a deep perceptual emotion consistency loss, EMOCA is capable of recovering facial expressions that convey the correct emotional state. A large 3D emotional talking face dataset (RAVDESS-EMOCA) is constructed using this method, each data sample consists of the original RAVDESS video and a corresponding 3D face mesh sequence generated by EMOCA.

3.2.2 Network Structure

The EmoFormer uses a encoder-decoder structure. As shown in Fig. 4, the encoder consists of a pre-trained wav2vec feature extractor, a content encoder and an emotion encoder. They extract features from raw audio input and disentangle them into temporal features (content features) and time invariant features (emotion features). These features are then used by decoder to generate animated mesh. The decoder consists of a single standard transformer decoder layer, motion encoder and decoder to map between vertices dimension and feature dimension.

EmoFormer Encoder The design of our EmoFormer encoder follows the state-of-the-art self-supervised pre-trained speech model, wav2vec2.0 [1]. Specifically, the encoder is composed of an audio feature extractor and a transformer encoder. In order to disentangle emotion and context information from the input audio, we model the encoder as two independent subencoders inspired by EVP [10]. The audio feature extractor transform the preprocessed audio into feature vectors. The two transformer encoder are both multi-head self-attention and feed-forward layers, embedding the extracted feature vectors into emotion and context spaces respectively.

EmoFormer Decoder The EmoFormer decoder consists of a single layer transformer decoder. Additionally, a set of motion encoder-decoder is used to reduce the number of parameters. They are a single fully connected layer which map between the vertices dimension and decoder feature dimension. The decoder itself utilize content feature and emotion feature as input. The content features are fed into encoder-decoder attention, while the emotion features are added to previous predictions, before they are fed into decoder to ensure that the emotion features affect each frame equally. The first frame $\mathbf{y_0}$ is directly generated from emotion features, so that it defines the neutral state of the whole sequence.

Inspired by faceformer [6], two bias masks are used in the transformer decoder when calculating self attention and encoder-decoder attention. Given the temporally encoded facial motion $\hat{\mathbf{F}}_\mathbf{t}$, self-attention first linearly projects $\hat{\mathbf{F}}_\mathbf{t}$ into queries $\mathbf{Q}^{\hat{\mathbf{F}}}$ and keys $\mathbf{K}^{\hat{\mathbf{F}}}$ of dimension $\mathbf{d_k}$, and values $\mathbf{V}^{\hat{\mathbf{F}}}$ of dimension $\mathbf{d_v}$, then calculates the attention:

$$\mathrm{Att}(\mathbf{Q}^{\hat{\mathbf{F}}},\mathbf{K}^{\hat{\mathbf{F}}},\mathbf{V}^{\hat{\mathbf{F}}},\mathbf{B}^{\hat{\mathbf{F}}}) = \mathrm{softmax}\left(\frac{\mathbf{Q}^{\hat{\mathbf{F}}}(\mathbf{K}^{\hat{\mathbf{F}}})^T}{\sqrt{\mathbf{d_k}}} + \mathbf{B}^{\hat{\mathbf{F}}}\right)\mathbf{V}^{\hat{\mathbf{F}}} \tag{3}$$

Where the $\mathbf{B}^{\hat{\mathbf{F}}}$ represents the temporal bias and is defined as:

$$\mathbf{B}^{\hat{\mathbf{F}}}(i,j) = \begin{cases} \lfloor (i-j)/\mathbf{p} \rfloor, & j \le i, \\ -\infty, & \text{otherwise.} \end{cases} \tag{4}$$

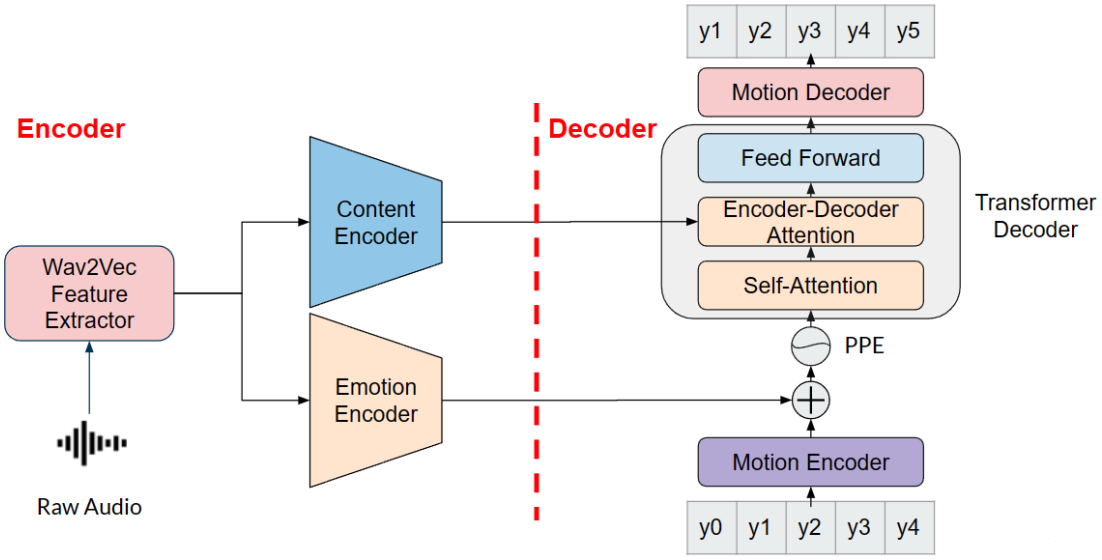


Figure 4. Network structure of EmoFormer. $\mathbf{y_i}$ is the coordinates of vertices at frame i. We use the encoder-decoder structure. On the encoder side, a pre-trained wav2vec feature extractor extract features from audio input, which is then feed into two different encoder. Content encoder encode temporal features and emotion encoder extract features that are time invariant. The decoder is a standard transformer decoder with separate motion encoder and decoder, it takes the output of content encoder and prediction of previous frame as input. The output of emotion encoder are added into the predictions before feeding into decoder so that the time invariant features can influence all frames equally.

The $\mathbf{p}$ is an adjustable parameter period, i and j are the indices of $\mathbf{B}^{\tilde{\mathbf{F}}}$. In this way, the frames that closer to current frame are most likely to affect the prediction of current frame during calculation of the self-attention.

The calculation of encoder-decoder attention is similar, assuming that the output of self-attention is $\tilde{\mathbf{F}}_\mathbf{t}$, which has the encoded history context of face motions. $\mathbf{A}_{k\mathbf{T}}$ represents the content feature which contains long-term audio context. Both of them are fed into encoder-decoder attention, where $\mathbf{A}_{k\mathbf{T}}$ is transformed into two separate matrices: keys $\mathbf{K}^\mathbf{A}$ and values $\mathbf{V}^\mathbf{A}$, and $\tilde{\mathbf{F}}_\mathbf{t}$ is transformed into queries $\mathbf{Q}^{\tilde{\mathbf{F}}}$. The output of encoder-decoder attention is then a weighted sum of $\mathbf{V}^\mathbf{A}$:

$$\text{Att}(\mathbf{Q}^{\tilde{\mathbf{F}}},\mathbf{K}^\mathbf{A},\mathbf{V}^\mathbf{A},\mathbf{B}^\mathbf{A}) = \text{softmax}\left(\frac{\mathbf{Q}^{\tilde{\mathbf{F}}}(\mathbf{K}^\mathbf{A})^T}{\sqrt{\mathbf{d_k}}} + \mathbf{B}^\mathbf{A}\right)\mathbf{V}^\mathbf{A} \tag{5}$$

With $\mathbf{B}^\mathbf{A}$ represents alignment bias, which is used to align the motion feature from motion encoder with the content feature:

$$\mathbf{B}^\mathbf{A}(i,j) = \begin{cases} 0, & ki \le j < k(i+1), \\ -\infty, & \text{otherwise.} \end{cases} \tag{6}$$

4. Experimental Results

This section presents an evaluation of the proposed methods.

4.1. VOCA Approach

4.1.1 Static Approach

By post-processing the output of the *VOCA* pipeline, the model is enabled to display emotions corresponding to the audio input. For example, one can make the animation be excited or fearful. Hence, this approach allows for more intuitive facial animations. However, the post-processing of the *VOCA* output implies multiple disadvantages. The most relevant one is the lack of smooth transitions. Since a certain offset will be added corresponding to the detected emotion per frame, a change in emotion will lead to abrupt changes in the animation which make the approach unrealistic. Therefore, the static approach remains a workaround, and does not provide a true solution to the challenge.

4.1.2 Training on emotional data

By training on emotional data, this work aims to obtain a model that is capable of animating expressions corresponding to the emotional content detected in the input audio. In particular, the goal is to realize these expressions on top of

the accurate lip movements achieved by *VOCA*.

Training parameter The following hyperparameters resulted in the best obtained model: a learning rate of 0.0035, a decay rate of 0.9, an adam beta value of 0.9, velocity loss of 2.25 and an acceleration loss of 0.5. Furthermore, the model is trained for 150 epochs on a batch size of 16. The corresponding trainings and validation loss can be seen in Fig. 5. When regarding the resulting animations, it becomes

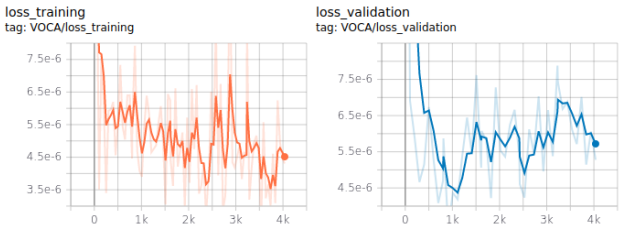


Figure 5. Training loss of the most promising model

clear that the model is not able to disentangle emotion and content of the speeches, especially, when comparing it with the results of the original *VOCA* model. The training on emotional data rather leads to the emotions messing up the lip movements so that in the end nothing works well anymore. Hence, the results suggest that the *VOCA* pipeline is not able to capture the contextual and emotional information of a speech input and reflect it accordingly in an animation.

4.2. EmoFormer Approach

4.2.1 Training

We performed an end-to-end training on 1000 audio sequence in our RAVDESS-EMOCA dataset. During the training, we adopt the autoregressive scheme for the transformer decoder, in which the decoder uses its previous predictions to predict the current frame. Once the prediction of the complete sequence is done, we calculate the Mean Square Error (MSE) between the decoder output $\hat{\mathbf{Y}}_\mathbf{t} = (\hat{\mathbf{y}_1}, \ldots, \hat{\mathbf{y}_T})$ and the ground truth $\mathbf{Y_t} = (\mathbf{y_1}, \ldots, \mathbf{y_T})$:

$$\mathcal{L}_{\text{MSE}} = \sum_{\mathbf{t=1}}^{\mathbf{T}} \sum_{\mathbf{v=1}}^{\mathbf{V}} ||\hat{\mathbf{y}}_{\mathbf{t,v}} - \mathbf{y_{t,v}}||^2 \tag{7}$$

where $\mathbf{V}$ represents the number of vertices of the mesh. The model updates its parameter by minimizing the MSE-loss Eq. (7).

4.2.2 Results

We first test our model with RAVDESS-audios that are unseen in the training phase, then we compare the result to the ground truth and to faceformer prediction. As shown in Fig. 6, our result is closer to the ground truth in terms of upper face movement and jaw movement, with which the emotion is expressed.

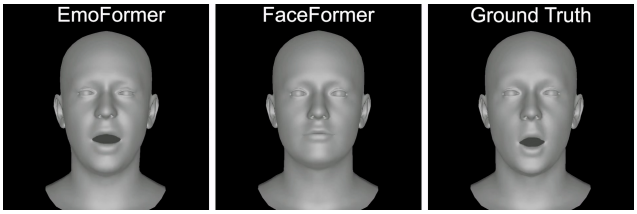


Figure 6. Testing result on unseen RAVDESS-audio with strong fearful emotion

We also test the generalization capability of our model by using self-recorded audio. For this test, we recorded the same sentence with 3 different emotion. The results are represented in Fig. 7. As it suggests, our model is capable to extract emotions from raw audio input and express them with correct facial movement.

(a) Angry (b) Neutral (c) Happy

Figure 7. Testing generalization with audios with same content but different emotions

User Control Our model provides a great opportunity for user adjustment by predicting vertices displacement from a static FLAME template. It allows us to change the speaker's identity by simply replace the FLAME template as post-processing (Fig. 8). Similarly, due to a independent emotion encoder, we can change the emotion of an audio content as our wish by manually editing the emotion vector (output of emotion encoder).

The results of change emotion are presented in Fig. 9. We manually change the emotion vector from neutral to happy, resulting in a corresponding change of prediction as anticipated. This behaviour provides additional evidence that the encoder effectively disentangles the underlying features, confirming that the emotion encoder's output encapsulates emotional features.

4.2.3 Ablation Study

For better understanding of our network, we performed ablation study by changing or removing some of its compo-

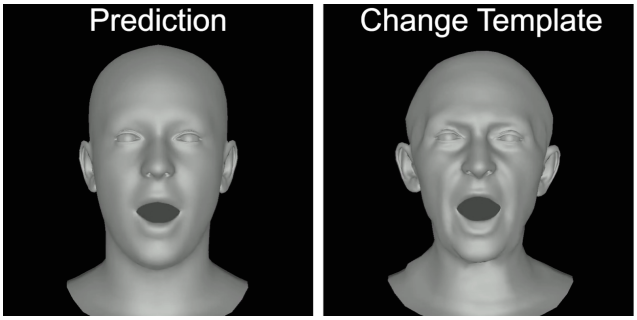


Figure 8. Change speaker's identity by changing FLAME template as post-processing

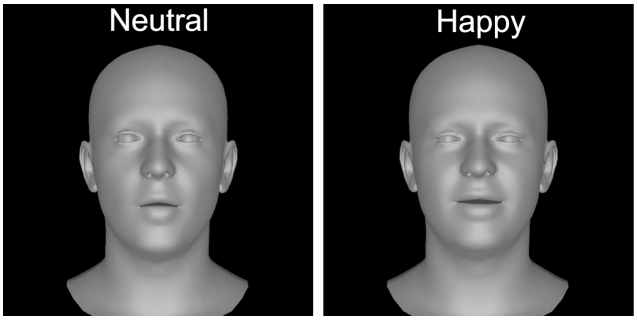


Figure 9. Change the emotion from neutral to happy

nent. Since we are using the parametric FLAME mesh as our basement, a straightforward consideration arises: instead of predicting the actual mesh, we can adjust our network to predict the FLAME parameters, which has significantly lower dimensionality. Thus our first ablation approach is to change the output of our model into FLAME parameters. We decide to only use the first 50 expression parameters and 3 parameters for jaw movement, because the shape parameters are not recognisable from audio. We also adjust the loss function as:

$$\mathcal{L} = \sum_{i=1}^{50} \mathbf{E_i} + \lambda \sum_{i=1}^{3} \mathbf{J_i} \tag{8}$$

Where $\mathbf{E}$ represents the expression parameter and $\mathbf{J}$ the jaw pose. We choose L1 loss to maintain the sparse characteristic of FLAME parameters, since the jaw pose has different order of magnitude as the expression parameters, we introduced a hyperparameter λ to control the integration of the two losses. The modification described above leads to a less favorable outcome, wherein our network predicts an average facial state accompanied by random mouth movements. We posit that the cause of this phenomenon could be that the FLAME parameters are too sensitive in contrast to vertices displacements. Additionally, the incorporation of these two loss functions transforms the task into a multi-task learning problem, necessitating careful fine-tuning of the parameter λ to prevent the network from exhibiting a bias towards a singular task.

The second ablation attempt aims to assess the role of our emotion encoder. We remove our emotion encoder by manually setting the emotion vector to zero. This leads to the similar outcome as the first attempt: an average facial expression with random mouth movements. Potential factors contributing to this outcome could originate from the dataset we used. Removing the emotion features introduces a situation where a single content corresponds to two distinct ground truth meshes, thereby amplifying the complexities of the training process. It is plausible that our network could have converged towards a local minimum, signifying the prediction of average face with random mouth movements. The supplementary videos are available[2].

5. Comparison

Since our approaches use different loss functions and these values are relatively unintuitive when it comes to interpretation, we compare the presented approaches using a common technique for benchmarking. We do this separately once without a weighting of our desired goals and once with a weighting for our desired project goals (Eq. (9)). The first static comparison can be seen as an overall comparison. The second one has a higher sense for how intuitive the approaches are when it comes to emotions, which was the main task of the project. For this reason we compare 4 major categories:

- Realistic mesh deforming: This category is used to measure how smooth and realistic the overall face movement is. A high score indicates a smooth face movement without unrealistic face deformation. A score in the mid area indicates that there is mesh deforming, but this is not really realistic, there are for example vertices jumping around. A low score indicates a completely deformed mesh that can not be identified as a human head anymore.

- Intuitive speaking: Although the lip movement is part of the mesh deformation, we decided to give this part of the face it's own performance indicator. Especially the lip movement is crucial for a good overall result. If the lips do not synchronize with the audio, this leads to a unintuitive and wrong result. A high score indicates that the lip movement is realistically fitting to the audio input. A score in the mid area indicates that there is lip movement but the lip movement does not synchronize with the audio all the time or there is sometimes unrealistic movement like jumps of the lips. A low score indicates that there is absolutely not realistic

[2]https://drive.google.com/drive/folders/1PsZuDQ9VN05-lUdkiVASaFRk7J1VQXtZ?usp=sharing

movement. Lips are not moving in any correlation to the audio or even not moving at all.

- Intuitive emotion: This indicates if the emotion that is part of the audio can be seen in the displayed mesh. A high scoring indicates, that the displayed emotion are realistic, intuitive and affecting the whole face, a score in the mid point area indicates that there are the right emotions displayed, but they leak to the intuitive part. A low score indicates that there are no reasonable or wrong emotions displayed in the mesh.

- Generalization: Because these approaches should be usable out of the box and should be able to perform intuitive emotional face animation also on unseen data, the generalization is also a performance indicator. As for the other indicators too, a high score means there is really good generalization and the result is intuitive, a mod score indicates that there is good performance on seen data, okay or good results on unseen data, but the results are not really intuitive and a low score indicates that the result is not able to learn any correlation and can't be used on unseen data.

$$\text{Score} = S_{\text{mesh}} + S_{\text{speaking}} + 2.5S_{\text{emotion}} + 0.8S_{\text{generalization}} \quad (9)$$

As represented in Eq. (9), our selection of weights closely adheres to the objectives of our project. Our primary focus centered on the generation of a mesh exhibiting intuitive emotions. Thus we assigned the highest weight (2.5) to the intuitive emotion performance indicator. The essential elements for result comparison were realistic mesh deformation and intuitive lip movement during speech. However, in our case, we were willing to strike a balance between lip movement and emotions. This rationale led us to assign weights of 1.0 to both realistic mesh movement and intuitive speech. Due to the limitation that RAVDESS dataset only contains two different sentences, we held a realistic expectation of imperfect generalization. As a consequence, a comparatively lower weight was assigned to this performance indicator with the value of 0.8. Tus the optimal outcome would correspond to a score of 530.

As represented in Table 1, it becomes evident that our approaches consistently outperform the baseline approaches. Specifically, VOCA and FaceFormer exhibit shortcomings in expressing emotions, a deficiency highlighted by our analyses. Notably, EmoFormer demonstrates the ability to effectively express emotions, consequently achieving the highest score among all the approaches.

The different methods have its advantages and disadvantages. a notable drawback of VOCA on RAVDESS is the absence of realistic mesh in some cases. The mesh of the upper face becomes static and ends up in an unnatural movement. Therefore, the score in mesh deformation and because of the correlation also the score for intuitive speaking are significantly lower then compared to the others. Nevertheless, within regions of the mesh where unrealistic deformations are absent, a reasonable degree of emotional expression is still evident. Overall this approach is not tending to generalize well, primarily due to its failure in effectively learning the desired features. The overall score appears satisfactory, and there is potential for enhancement through further investigation and exploration. In contrast, VOCA Blendshapes perform reasonably well. This variant maintains smooth and realistic mesh deformation, while also preserving the quality of lip movement observed in the original VOCA approach. It introduces emotions that are beyond the scope of the conventional VOCA approach, as the template remains static and lacks interactivity, thereby lacks intuitiveness. The generalization score is significantly lower because every speaker gets the same emotional style, which deviates from the real world. Hence, VOCA Blendshapes presents a sensible design choice that is both practical and follows the principle of "Keep It Simple and Straightforward" (KISS). In contrast to the EmoFormer approach, the results it produces lack the desired level of intuitiveness and demonstrate poorer generalization. Thus the VOCA Blendshapes could be used for emotional manipulating meshes. The EmoFormer Approach excels in mesh deformation, displaying remarkable performance overall. However, occasional minor errors emerge in certain short sequences when dealing with unseen data. The lip movement functions effectively. However, in a few short sequences, there is a noticeable lack of synchronization between the lip movement and the audio. As a result, a high score, although not perfect, was assigned to this performance indicator. EmoFormer exhibits the ability to infuse intuitive and lifelike emotions into the mesh, with particularly impressive outcomes observed in the upper face movement, where most of the emotion can be derived of. delivers really good results. In general, the model delivers really good results, its generalization capacity is commendable, but due to the limitation of training dataset, certain sequences exhibit instances of generalization issues.

6. Summary and Conclusion

All in all, all our approaches extended existing ones and created a more usable result for intuitive speech driven face animation. We indicated that the correlation between speech and upper face movement is quite hard to learn for an encoder decoder based network and that a method including transformers works significantly better for realistic generalization in this case. All explored methods can be used and extend the existing base work of FaceFormer and VOCA. The provided results are resonable and can be benchmarked.

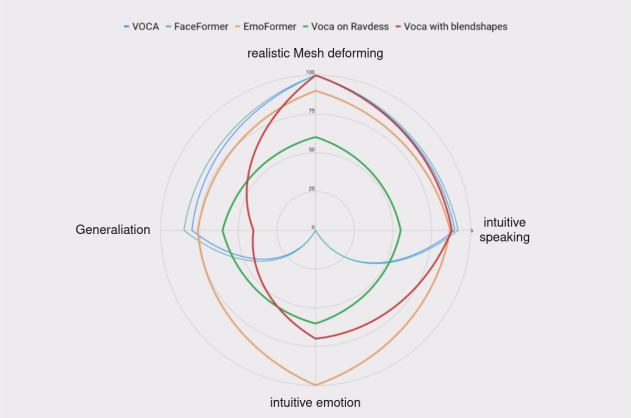


Figure 10. This Diagram holds an overview about the different approaches. Each category is rated between 0 and 100 points, where 100 is perfect and 0 is bad. Our own approaches are displayed with a more width line style. A more detailed conclusion of this comparison can be seen in Tab. 1

Since there are not really speech driven face animation models that aim to display realistic emotions there cannot be a comparison to state of the art approaches. With further research or simply new data that is suitable for this task, we could expect the EmoFormer approach to generalize better.

References

[1] 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.

[2] V Blanz and T Vetter. A morphable model for the synthesis of 3d faces. In *26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH 1999)*, pages 187–194. ACM Press, 1999.

[3] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael Black. Capture, learning, and synthesis of 3D speaking styles. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10101–10111, 2019.

[4] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022.

[5] Pif Edwards, Chris Landreth, Eugene Fiume, and Karan Singh. Jali: an animator-centric viseme model for expressive lip synchronization. *ACM Transactions on graphics (TOG)*, 35(4):1–11, 2016.

[6] Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph.*, 32(6):158–1, 2013.

[8] T. Gerig, A. Morel-Forster, C. Blumer, B. Egger, M. Luthi, S. Schoenborn, and T. Vetter. Morphable face models - an open framework. In *2018 13th IEEE International Conference on Automatic Face amp; Gesture Recognition (FG 2018)*, pages 75–82, Los Alamitos, CA, USA, may 2018. IEEE Computer Society.

[9] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021.

[10] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[11] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6), nov 2017.

[12] Yunzhu Li, Benyuan Sun, Tianfu Wu, and Yizhou Wang. Face detection with end-to-end integration of a convnet and a 3d model. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 420–436. Springer, 2016.

[13] 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.

[14] Hai X Pham, Samuel Cheung, and Vladimir Pavlovic. Speech-driven 3d facial animation with implicit emotional awareness: A deep learning approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 80–88, 2017.

[15] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Approach	**Realistic Mesh Deforming**	**Intuitive Speaking**	**Intuitive Emotions**	**Generalization**	$\sum$	**Weighted** $\sum$
VOCA	100	90	0	80	270	254.0
FaceFormer	100	92	0	85	277	260.0
EmoFormer	90	87	100	76	353	487.8
VOCA on RAVDESS	60	55	60	60	235	313.0
VOCA Blendshapes	100	88	70	40	298	395.0

Table 1. Comparison of approaches with ratings for different categories. Weights for the weighted sum score are 1.0 for realistic mesh deforming and intuitive speaking, 2.5 for intuitive emotions, and 0.8 for generalization. The chosen weights are explained in Sec. 5.