



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

An Audio-driving Facial Animation Method based on Multi-modal Relevance

by
Yizhen Ma

Supervisor: Gordon Dickers
Second Marker:

Project submitted as part of the requirements for
the award of MSc Software Engineering (Web Develop-
ment)

November 2021

DECLARATION

I, ...Yizhen Ma..... declare that I am the sole author of this Project; that all references cited have been consulted; that I have conducted all work of which this is a record, and that the finished work lies within the prescribed word limits.

This work has not previously been accepted as part of any other degree submission.

Signed : Yizhen Ma.....

Date : 11/11/2021.....

FORM OF CONSENT

I _____Yizhen Ma_____ hereby consent that my Project, submitted in candidature for the *MSC Software Engineering (Web Development)*, if successful, may be made available for inter-library loan or photocopying (subject to the law of copyright), and that the title and abstract may be made available to outside organisations.

Signed _____Yizhen Ma_____

Date _____11/11/2021_____

ABSTRACT

This dissertation proposes a novel audio-driven facial animation method based on multi-modal relevance, that is, given a piece of audio and a segment of talking video, this method can generate realistic video results of the same speaker conditioned on the audio. To take full advantage of different modality inputs, this paper designs a neural network to model the relevance of sequential inputs, including intra-sequence and inter-sequence across modalities, the calculated relevance scores are transformed into weighted value and contribute to predicting dynamics of human faces. In addition, this method utilizes the intermediate human faces representation of 3DMMs, which convert a face image into a vector in disentangled representation space, by introducing 3DMM, we exclude the impact of audio non-related regions in human faces when performing mapping from audio signal to facial movement. In order to visualize our results, we utilize a neural rendering network to generate photorealistic RGB images of human faces from 3D coarse rendering results. Compared with other methods, this method is directly oriented to sequential input, which is more in line with the characteristic of input data in audio-driven facial animation tasks. In addition, owing to the design of Transformer architecture, this method can perform parallel prediction, which means the entire input sequence can be processed in one step, more importantly, by adding the correlation information of input sequence, which obtained in relevance modelling stage, this method can effectively utilize the information of adjacent frames, thereby generating more accurate lip results, at the same time, gain better performance in lip-audio synchronization. Both qualitative evaluation such as self-reenactment and multi-target reenactment and quantitative evaluation such as landmark error and expression error are conducted and the results prove the effectiveness of this method.

CONTENTS PAGE

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION	5
CHAPTER 2: LITERATURE REVIEW	6
CHAPTER 3: RESEARCH METHODOLOGY	14
3.1 EVALUATION CRITERIA.....	14
3.2 EXPERIMENTAL DESIGN	14
3.2.1 OVERVIEW	15
3.2.2 DECOUPLING OF FACE ATTRIBUTES BASED ON 3DMM	16
3.2.3 SPEECH-TO-EXPRESSION MAPPING BASED ON MULTI-MODAL RELEVANCE.....	18
3.2.4 PHOTOREALISTIC 2D IMAGE GENERATION BASED ON NEURAL RENDERING	22
CHAPTER 4: DESIGN AND IMPLEMENTATION.....	23
4.1 DESIGN	23
4.2 IMPLEMENTATION	25
CHAPTER 5: RESULTS AND DISCUSSION / EVALUATION.....	27
5.1 QUALITATIVE EVALUATION	27
5.2 QUANTITATIVE EVALUATION	29
CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS.....	31
REFERENCES	32
APPENDICES	37
1. ETHICS FORM.....	37
2. SOME CRITICAL CODE	44
3. MORE RESULTS	45

CHAPTER 1: INTRODUCTION

The purpose of facial animation is to make use of information from different driving modalities to generate highly natural and smooth facial animations. Facial animation technology has strong practical significance and application value in scenarios such as human-computer interaction and virtual reality. The current human-computer interaction is usually through voice or text communication, but compared with the process of human-to-human interaction, the experience it brings to people is completely different. In fact, facial expressions, body movements, and tone of voice in human-to-human interaction play a vital role in the interaction process. The same content will give people completely different feelings through different facial expressions. At present, human-computer interaction based on facial animation technology has been applied in many real-world scenarios, such as artificial intelligence anchors and virtual customer service agents.

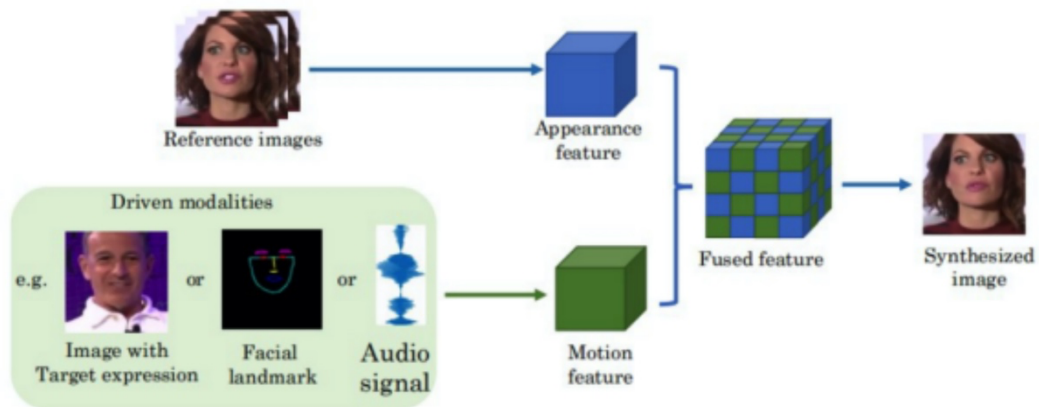


Figure 1. Schematic diagram of face animation generation

Facial animation usually require two types of input, namely reference input and driving input(also called source input), as shown in Figure 1. Among them, the reference input provides information about the character, including the character's identity information, posture information, expression information, etc.; the driving input provides the information we want the character to animate. Our task is to refine the parts we want from the reference input and the driving input. For reference input, we generally hope to extract identity information and posture information of the character; for driving input, we hope to extract facial expression-related information contained therein. For the case where the driving input is video, we can directly extract the corresponding expression information since it inherently lies in the 2D space, but for driving modality with less information such as text or speech, additional

modules are required to transfer it to 2D space. After that, we need to fuse the extracted information and generate video results based on the fused information.

Facial animation generation can improve user experience in industries such as film making, television and games. Since this task usually needs to be applied to the scene of human interaction, it has extremely high requirements for real-time and image quality of the results, and how to generate real, natural, non-mechanized results is still under exploring.

We mainly focus on one specific type of task in facial animation generation, i.e. audio-driven facial animation where the driving input is a piece of audio and the reference is a piece of talking video. Current research in this field usually focus on how to optimize the process of audio-to-expression mapping[38][40][45][50] or employ special designed network to improve the quality of synthesized video[47-49]. In our research, we are going to use a Transformer-based sequence generation model during feature extraction and generation stage, so as to solve the shortcomings of the current sequence model in terms of computational efficiency, long-term and short-term dependency extraction. In addition, current face animation methods do not effectively use the information in the reference video, while latter contains a lot of information about the speaking style of the reference character. We explicitly study the correlation between the driving audio and the facial expression features obtained from reference video through 3D face reconstruction process[12], so as to establish a mapping from the source audio to the 3D space.

The structure of this dissertation is as follows, in Chapter 2, we introduce the development of facial animation and related technologies in our method, including 3D face model, Transformers and facial animation. In Chapter 3, we explain the methodology used in our experiments, including evaluation methodology, design of system architecture and the pipeline of our proposed method. In Chapter 4, we explain the design and implementation of our proposed method in detail, including comparison between several possible options and why make such decisions. In Chapter 5, we evaluate our method using several criteria proposed in Chapter 3. In Chapter 6, we summarize the work we do and point out the future application of this technique.

CHAPTER 2: LITERATURE REVIEW

3D Face Model. The 3D Face Morphing Model (3DMM) was first proposed by Blanz and Vetter at SIGGRAPH in 1999 [1]. They collected a large 3D scan face dataset and employed optical flow to construct dense semantic correspondence of the original scan data.

They used PCA to establish a face parameterized model, which decouples the face in the 3D scene into five types of representations, including shape (identity), expression, texture, lighting, and posture. Through controlling different components of the input face representation, the 3DMM model can generate fine-grained 3D face by controlling the weight of the components. The contribution of 3DMM is that, first of all, it provides a controllable face generation model, we can achieve fine-grained generation by controlling the different components of the parameters; in addition, it provides a decoupled face representation, while this decoupling representation is more maneuverable; finally, it provides a paradigm for constructing 3D face models, which has a profound impact on subsequent research including methods based on traditional methods and methods based on deep learning. [2] first proposed to use an encoder-decoder structure to model the geometric structure of a human face, where the encoder projects a 3D face onto a 2D image, and the decoder fits a classic face model. This allows the decoupling of shape changes caused by identity and expression. Bagautdinov et al. [3] introduced a VAE-based structure to represent global and local shape changes in different layers of the network, and then simulate different levels of details of the facial geometry. In their work, the 3D geometric structure is represented by using a two-dimensional mapping and convolved in the image dimension. This work can represent highly detailed geometric structure information in the hidden space. Furthermore, Lombardi et al. [4] proposed to jointly encode changes in appearance and geometry, so that highly detailed facial rendering can be performed from a new perspective. Ranjan et al. [5] proposed a CoMA model based on the auto-encoder architecture. This is the first model that directly convolves on 3D mesh to model the geometric structure of the face, allowing very concise representation of facial geometry. Li et al. [6] proposed FLAME head model based on a 4D face scan dataset, which can combine jaw movement and linear expression blendshapes to provide nonlinear control on facial expressions. Thomas et al. [7] proposed a new strategy and modeling technique for face registration, which considers symmetry, multi-scale and spatially varying details, and can be applied to neutral faces and faces expression. In addition, they released an open-source software framework for registration and model building. Based on the framework, they released a new version of Basel face model (BFM17), which broadened the age distribution of the original model and added an additional facial expressions model. In view of its powerful expressive ability and stability, BFM17 has gradually become the priority for current 3D face models. Egger et al. [8] summarized the development of 3D face models in 20 years, and reviewed the establishment of 3D face models and the challenges of applying these models (including

capture, modeling, image formation and image analysis). They compared different 3D face models and provide suggestions for future research methods of this field.

Except for 3D face model itself, the research on the application of the 3D face model is also a very active topic. Among them, 3D face reconstruction is a research direction that develops quite rapidly. 3D face reconstruction refers to obtaining a 3D face representation of the person from a face image or a sequence of face images. Since 3D faces are highly manipulable, we can edit 3D faces to change certain attributes. After processing with neural rendering technology, we can achieve the purpose of editing the characters in 2D images, thus avoiding various problems of directly editing 2D faces.

Tuan et al. [9] proposed a regression model of 3DMM CNN to regress 3D face shape and texture parameters from images. In addition to focusing on the regression of face shape, ExpNet[10] and FacePoseNet[11] are designed to regress expression and pose parameters of the 3D face solely. Deng et al. [12] proposed an accurate 3D face reconstruction method based on weakly supervised learning. They propose to use image pixel loss and perception loss in the reconstruction stage, and can regress the coefficients of all 3D faces at once. In addition, their model can be well adapted to the presence of facial occlusion and multiple pictures of a single person. In their results, those generated based on multiple pictures of single person is better than that of a single picture. [13] summarized the development process of 3D face reconstruction technology by 2020, and put forward a prospect for the future of 3D face reconstruction technology.

Transformers. Transformer [17] is one type of neural network which mainly based on self-attention mechanism. It was originally proposed in the field of NLP and used to solve the problem of computational efficiency and long-term dependence of traditional sequence models such as LSTM and GRU. Since the day it was proposed, Transformer has shined in the field of NLP, surpassed many state-of-the-art results at that time in multiple tasks, and proved the rationality of its architecture. The follow-up work is basically based on the original Transformer architecture. Among them, BERT [18] pre-trained a Transformer using self-supervised methods on unlabeled text data. The text features obtained by the Transformer have achieved better results than traditional Word2Vec in various tasks. Although BERT is effective, but because the model is too large, and requires too many hardware resources, so it is not suitable for many lightweight scenarios. In view of the high time complexity and excessive memory usage of the original Transformer, Zhou et al. [19] proposed a new Informer architecture. The ProbSparse self-attention mechanism proposed by this architecture is less

time-complex than the original self-attention. In addition, it also proposed a distillation mechanism to shorten the length of the self-attention sequence, thereby reducing the memory consumption of the architecture. Finally, they proposed to predict only once instead of the original step-by-step prediction, which allows the predicted results to be obtained in linear time.

While Transformer is hot in the NLP field, it has also caused CV scholars to think about Transformer. Dosovitskiy et al. [20] first proposed a Transformer-based image classification model, ViT, which novelly regarded image as a sequence of patches, and retained the original structural information of the image by indexing each patch thus construct a pseudo-sequence. In addition, the patch-based image sequence contains less information, which circumvents the problem of excessive memory consumption of Transformer to a certain extent. ViT retains the original NLP Transformer's architectural design as much as possible, and entirely based on attention mechanism. ViT has been trained and tested on multiple image classification datasets. The results show that it has achieved better classification results in scenarios with large amounts of data. Even when the amount of data is small, it can achieve a good image classification result compared with other models. Compared with the way that completely abandons CNN in ViT, DETR [21] seeks a compromise. For the object detection task, DETR first uses CNN to obtain the feature sequence of the image. Since the object detection task does not require ordered inputs, so the feature sequence obtained by CNN can directly input into a Transformer to predict the category and coordinates of the object, which is more efficient and convenient than the traditional methods of setting anchors. Chen et al. [22] proposed a Transformer for general vision problems, namely IPT. The main idea of IPT is to use the strong expressive ability of Transformer model, and use multi-head and multi-tail learning on a large-scale corpus, then draw the idea of contrastive learning to fine-tune the model on different tasks. IPT has achieved better results than current best methods on different underlying CV tasks (such as image super-resolution, denoising, etc.). Swin Transformer [23] as the latest Transformer model, has achieved higher accuracy on image classification problems, at the same time the speed is close to CNN methods. Swin Transformer draws on the idea of hierarchy and local connection in CNN, achieves better prediction accuracy by constructing a hierarchical Transformer and focus on single window through a sliding window, thereby reducing the overall amount of calculation.

In addition to these pseudo-sequence problems that divide images into patches, there are also some CV tasks dealing with image sequences. These tasks are more in line with the

characteristics of Transformer itself, and good results have been achieved on some tasks. Liu [24] et al. proposed ConvTransformer architecture and used it to solve the video prediction problem, that is, predicting the intermediate or subsequent frames based on the existing video input. Its characteristic is to replace the 1D tensor input in traditional Transformer with a 3D tensor, and design a 3D tensor attention calculation method based on convolution. So that the Transformer architecture can be adapted to the problem of image sequences, and it can achieve quite realistic results in the prediction of intermediate frames and the prediction of subsequent frames. In addition to the single-modality problem in CV field, Transformer is currently also used to solve cross-modal or multi-modal problems in CV fields. Zheng et al. [25] used Transformer to solve visual question answering (VQA) task. They proposed to use three Transformers to model the correlation in text modality, image modality (using object detection skeletons), and cross-modality. These three Transformers allow the model to capture the high-order relationship between the inputs, so as to achieve the purpose of expressing spatial alignment. Inspired by the information retrieval model, they also used a relevance modeling method based on text entities and visual entities. It turns out the combination of entity-based and cross-modal relevance perform better than other methods. Similarly, Li et al. [26] used a cross-modal Transformer to model the correlation of different modality inputs in the task of audio-driven 3D dance generation. Their main contribution is that they labeled the AIST dataset, used optical flow method to establish dense semantic correspondence of the dataset, and used it as representation of motion feature. Experiments have proved that dance movements generated based on the cross-modal Transformer can not only make movements in every frame more realistic, and achieve a higher continuity in the overall video. Han et al. [27] summarized the progress of visual Transformer, and concluded several success factors of the visual Transformer and the future direction.

Facial Animation. Face animation refers to the generation of video of the target person speaking under the conditions of one or more pictures of the target person and input of one or more driving modalities. The resulting video contains information come from driving modality(speaking content) and also the information of the target person (identity of the speaker, speaking style, etc.). According to the driving modalities, the task of face animation can be divided into (1) video-driven face animation[31-37]; (2) audio-driven face animation. According to face representations used, tasks can be divided into 1) latent vector representation; (2) facial landmark representation; (3) 3D face representation. Due to the fact that different driving modalities will lead to totally different problems that need to be solved. Here, we

only list the related work that is closest to our task, that is, researches of audio-driven facial animation.

The goal of audio-driven facial animation is to generate a video result that conforms to the audio content of the driving input. A key problem that needs to be solved in this task is the alignment of audio to facial expression space. Considerations include: character identity preservation, audio-visual synchronization, natural movement, etc. Audio-driven face animation can be divided into methods based on latent vector representation, facial landmarks and 3D face according to the difference of intermediate face representation. Here we will discuss them separately.

The method based on latent vector uses latent vector as the representation of facial expressions, and controls the accuracy of the speech-to-face expression mapping through the loss function of image result, without the need to generate a visual intermediate facial representation. The advantage of these methods is that end-to-end training strategy can be used, and the demand for computing resources is relatively small. Jamaludin et al. [38] first proposed a facial animation generation network based on an encoder-decoder structure. They used two sub-networks to extract the expression feature of the audio and the identity feature of the face in one picture, and then obtained a video result through the decoder. It belongs to the early exploratory work of face animation generation based on deep learning. Due to its simple architecture and no additional control conditions, the generated result suffers from problems such as identity leakage, low video resolution and unnatural smoothness, etc. Wiles et al. [40] proposed X2Face network. By learning an intermediate representation of the embedded face, the network effectively separates the information of the character's identity, expression and head posture, thereby training the network in several stages. The problem of this model is the instability of face embedding and the generation network are relatively simple, besides, the results of their model are unstable. Vougioukas et al. [42] proposed to use a frame discriminator and a sequence discriminator in the generation stage to solve the problem of unsmoothness in generated video. On the one hand, they ensure the quality of the generated results, and on the other hand, ensure the stability of the generated video results.

The method based on facial landmarks first considers mapping from audio to accurate face landmarks, and then uses the image-to-image translation network to generate photo-realistic pictures. The method using facial landmarks can add loss constraints on facial landmarks, which can alleviate the stability problems of methods based on latent vectors due to end-to-end training to some extent. Suwajanakorn et al. [39] proposed a single person-

oriented face animation method. They used a large number of Obama speech videos to train their speech-to-lip-landmarks network and rendering network, so as to be able to generate very photorealistic results for single person. Since only single character is used in the training data, their network generalization ability is poor, and the reproducibility is not high. Sefik et al. [41] proposed a network to map audio to the movement of facial landmarks. Their method can generate better lip synchronization results on standard face, but because only the landmarks of the standard face is generated without considering the difference in landmarks of different people, so the practical value of their work is low. Zhang et al. [43,44] proposed the APB2Face and APB2FaceV2 networks. The former uses additional posture and blink coding, and only extracts the expression feature from the audio. The former has achieved good results on their self-built dataset and can generate natural blink results; the latter proposed an adaptive convolution module based on the former architecture to fuse audio features, and used a lightweight network structure to achieve real-time face animation. Zhou et al. [45] proposed a novel facial animation generation network, Makeittalk, based on the intermediate facial landmarks. In their audio processing stage, the audio decoupling method was used to divide the audio into identity and content parts to control different feature levels respectively. Their method can generate natural head and lip movement, but the problem is that their head movement is trained based on a large video dataset of a single person, so there exist problems such as fixed and mechanical movement pattern. Yu et al. [46] proposed a facial animation method based on multi-modal driving input, that is, text and speech co-drive pattern to generate landmark movement of the standard face, then use the de-standardization method to convert the standard face to specific face, and finally generate the resulting video through a generation network based on the face sketch. Prajwal et al. [47] used masks of lower half of the face on the original video frame, and used a pre-trained lip-reading discriminator network to ensure that better lips are generated, because the network only focuses on the generation of the lower half of the face, their proposed network can generate very precise lip movements.

The 3D face-based methods first consider how to obtain accurate facial expression parameters in 3D space from audio. According to the types of final result, it can be divided into directly generating 4D mesh results or using rendering methods to obtain 2D video results. Karras et al. [48] proposed a low-latency audio-driven facial animation generation method based on face meshes. Their method directly learns the mapping from audio waveforms to 3D face mesh vertices, and uses a low-dimensional vector to represent expression changes that cannot be learned by audio. Cudeiro et al. [49] proposed the VOCA network to learn the

mapping of audio to 3D face vertex transformation, which can control the difference in speaking style through additional input. The 3D face model in their work is FLAME[6], and released an open source dataset of 4D face scans. The main problem with VOCA is that the original model is only trained on 12 speakers, which can only be tested on the template characters preceded during training, thus limiting the scope of application of the model. Song et al. [50] proposed a 3DMM-based audio-driven facial animation model. They used a pre-processing network to extract audio content features, and then used a mapping network to map the audio to the expression space of 3DMM. After obtaining the recombined features, Rasterization and cropping are used to obtain an image of the lip area, then a neural rendering network is trained to obtain a refined lip result, final video is obtained after stitching with the original image. Their method can well cope with large pose changes. Yi et al. [51] innovatively proposed to study the correlation between audio and posture changes, and used neural networks to establish their mapping. In addition, in view of the weak generalization ability of current neural rendering networks, they refer to memory mechanism and proposes a memory network to retain speaker information during training, so as to speed up and optimize the inference process through the calculation of speaker similarity during testing. Similarly, the AudioDVP proposed by Wen et al. [52] is very similar in network structure to [51]. The difference is that they do not consider the relationship between audio and head pose, and only calculate the reconstruction loss of the lower half of the face during the rendering process. Since the head pose is obtained from the original frame, the lower half of the generated face can be pasted back to the original image to ensure sufficient authenticity. The problem with AudioDVP is that the neural rendering module requires a long speaking video for fine-tuning during testing, so real-time rendering cannot be achieved. Richard et al. [53] proposed a cross-modal 3D facial animation model MeshTalk. They proposed a cross-modal loss in the classification hidden space of face animation, which combines audio-related and audio-unrelated parts separately, thereby generating more accurate face mesh animation results. MeshTalk has some versatility problems, including a certain delay in audio and video. In addition, the model cannot achieve real-time performance on some hardware with low computing power.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 EVALUATION CRITERIA

The goal of audio-driving facial animation is to generate photorealistic and temporal coherent speaking video, and most importantly, requires lip motion in sync with the content of audio. We use both qualitative and quantitative criteria to analysis the results and evaluate the performance of our system.

For qualitative evaluation, we adopt two strategies to evaluate our system, the first is self-reenactment, where source audio and reference audio are the same, the other is multi-target reenactment, which compare the results using the same audio to reenact different character. In self-reenactment, the video result should be similar to the original video since both audio input and character are the same. In multi-target reenactment, the lip movement should be similar across different generated results since the driving audio source are the same.

We further carried out quantitative evaluation. We calculate the L2 difference of 3D expression coefficients between generated results and ground-truth frames, since expression coefficients of 3DMM reflect the whole facial expression rather than audio-correlated region, to focus on lip shape accuracy, we use additional facial landmark L2 difference between generated results and ground truth images. In addition, another factor we care for is the audio-visual synchronization of generated video, therefore we adopt SyncNet[54], which calculate the audio-visual synchronization score using a two-stream neural network. For the details of evaluation implementation and results, please refers to Chapter 5.

3.2 EXPERIMENTAL DESIGN

The main purpose of this project is to generate realistic talking videos while take full advantage of information in all input modalities. Generally, in the task of audio-driven talking video generation, the inputs usually consist of a piece audio and single image, sometimes accompanied with a short video to fine-tune the network to learn person-specific speaking style. In training stage, both the audio stream and video stream are split into frames, and the network learns to map feature of specify audio frame to corresponding target video frame feature, which can be considered as a frame level one-to-one mapping. Actually, this pattern lacks consideration of relevance between adjacent frames of both audio and visual modalities, which can be seen as an intrinsic feature of the input sequence, besides, even some methods already consider using sequence model to extract temporal audio feature to improve temporal

stability, they still not explicitly consider the relevance in single modality or relevance between audio and video stream, i.e. the multi-modal relevance.

In our hypothesis, we assume explicit modelling of both single modality relevance and multi-modal relevance between audio and visual modalities will be beneficial for the mapping from audio to visual feature. Based on this hypothesis, we design a system to 1) model relevance of different modalities, 2) compute audio-visual mapping based on the relevance scores of different modalities, 3) visualize results by 3D rendering and use a neural rendering network to verify the effectiveness of our method.

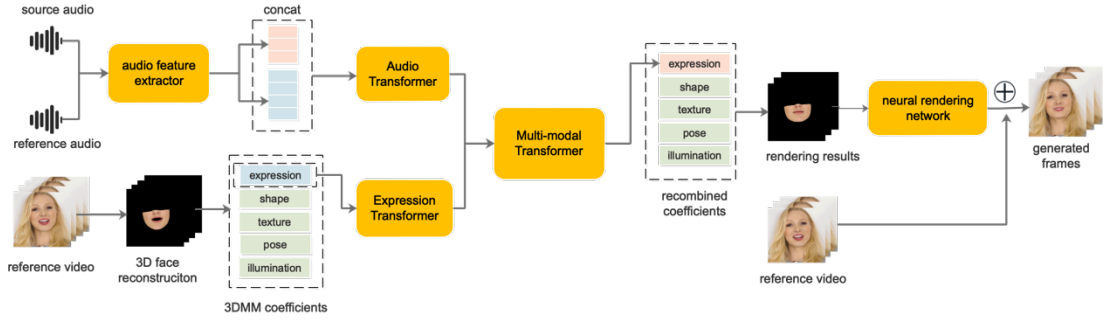


Figure 2. Flowchart of our proposed audio-driven talking video generation model based on multi-modal relevance, except for the usual audio to expression mapping module, we introduce an additional branch which accept video frames as input. Both audio and image input are being mapped to 3DMM space, self-attention and cross-attention of intermediate 3DMM representation are being calculate in Transformers. The predicted expression coefficients replaced those extracted from video frames, and render to photorealistic face images by a neural rendering network.

3.2.1 OVERVIEW

The framework of our proposed method is showed in Figure 2. The framework of our proposed method has three main components, decoupling of face attributes based on 3DMM(described in Section 3.2.2), speech-to-expression mapping based on multi-modal relevance(described in Section 3.2.3) and photorealistic 2D image generation based on neural rendering (described in Section 3.2.4)

The model formulation is as follows; given a piece of audio clip A_{src} (hereinafter called source audio or driving audio) and a short talking video V_{ref} (hereinafter called reference video). Our goal is to generate a new talking video $\hat{V}$ of the same person in V_{ref} while the content of speech keeps the same with A_{src} .

The source audio is separated into m frames $a_{src} = \{a_{src_1}, a_{src_2}, \dots, a_{src_m}\}$. Similarly, audio stream and visual stream are separated from the reference video and converted into a sequence of audio frames $a_{ref} = \{a_{ref_1}, a_{ref_2}, \dots, a_{ref_n}\}$ and a sequence of video frames

$v_{ref} = \{v_{ref_1}, v_{ref_2}, \dots, v_{ref_n}\}$, both have a length of n . For each reference video frame, we use 3D face reconstruction method[12] to obtain the reconstruction coefficients which contains expression and other components. For i th video frame, after 3D face reconstruction, a vector X_i is obtained of R^{257} . For source audio sequence a_{src} and the reference audio sequence a_{ref} , we use librosa[55] to extract mfcc feature of those audio clips whose dimension is 13, and then concatenate those mfcc features along time dimension, result in an input vector $mfcc$ of shape $(m + n, 13)$.

After obtained independent representation of both audio and video input, to fuse information of audio modality and visual modality, we adopt a multi-modal relevance computing architecture to predict visual output in according with reference audio. The relevance is computed by Transformer[17] model, which is a neural network all about attention. There are two types of relevance to be computed, single-modal relevance and multi-modal relevance. Single-modal relevance modeling is performed to obtain the sequence-dependent representation of source audio, then the multi-modal correlation is performed among source audio feature, reference audio feature and reconstructed expression coefficients of reference video frames. After that, the predicted expression coefficients are combined with other facial coefficients of reference frames in order, which are then feed into 3DMMs to generate 3D meshes. Next, using OpenGL, the 3D meshes can be rasterized and what we get are coarse 2D rendering frames. Finally, a neural rendering network is used to render these coarse frames to obtain photorealistic video frames, and the video frames are spliced to obtain the predicted video result. The following sections describe each procedure in detail.

3.2.2 DECOUPLING OF FACE ATTRIBUTES BASED ON 3DMM

One of the core issues of audio-driven facial animation task is how to represent human faces. Due to the high complexity of human faces, it is quite difficult to find a suitable feature that represents the face; in addition, various attributes of human faces are highly entangled, brings about more difficulties for some specific tasks which needs to analysis specific semantic features of human faces. A recent study[34] have shown that a human face can be represented as a point on a high-dimensional manifold. Therefore, dimensionality reduction algorithms are often needed to reduce the computational complexity.

Monocular RGB reconstruction based on 3DMM is a method to solve the problem of face representation. 3DMM itself is a 3D face statistical model, which solves the problem of recovering 3D faces from 2D images; in addition, 3DMM provides parameterized and

disentangled space which is able to manipulate specific attributes of given face model, including face shape, expression, texture, posture, etc. It provides the possibility to independently study a certain component of the face. With the development of neural networks, the 3DMM coefficients solved by neural networks have better accuracy and higher computational efficiency than traditional methods that are calculated through optimization algorithms.

Except for 3DMM, there are other options for human face representation, for example, facial landmarks. Compared with these methods, 3DMM has more advantages:

- 1) **Easy to manipulate.** 3DMM is more maneuverable than other face representation due to its disentangled parameters space and parametric characteristics such as better control of attributes such as lighting, expression, and posture.
- 2) **Better identity preserving.** Among audio-driven talking face generation experiments, some studies based on 2D face representation usually suffer from the problem of identity leakage, which means the generated faces are far from the original person and unidentical, the reason can be attributed to inadequate separation from facial movement-related parts and facial structural parts, resulting facial structural information is changed during the process of reenactment. For 3DMM representation, disentanglement of moving parts of face (lip, facial muscles) and motionless parts (shape, contour) have already been done during the face reconstruction procedure. Since we usually do not modify the identity component of the 3D face in the reference video, the face animation generation task based on the 3D face can provide a better identity preserving effect.

We use 3DMM face representation in our experiment. Except for the above mentioned advantages of 3DMM, explicit decoupling of facial motion-related features from other motionless features is crucial for the multi-modal relevance modeling procedure in the next section, since for relevance modeling, it learns the potential relevance between different frames in the input sequence, so that frames at different positions in the input sequence will affect the prediction of the current frame, current research indicated that only some acoustic feature of speech have connection with movement of specific areas of faces, including some non-speech-related facial features in input, for example, geometry of face, will cause the network fail to converge.

Here we use [12] to reconstruct 3D face from 2D input image, since their method is fast, accurate and robust to occlusion which is pretty common in real-life data. The specific process is as follows, given the input image sequence $v_{ref} = \{v_{ref_1}, v_{ref_2}, \dots, v_{ref_n}\}$, which length is n , we input each image separately into the 3D face reconstruction network to obtain

a vector of 257 dimension $X_i = (\alpha_i, \beta_i, \delta_i, \gamma_i, p_i) \in R^{257}$ specific to the current image v_{ref_i} . $\alpha_i, \beta_i, \delta_i, \gamma_i, p_i$ respectively represent geometric, expression, texture, lighting and pose coefficient vectors of 3D parameterized face model. This set of parameters is adapted to specific face model which is settled during training phase. Here the expression basis are built from a subset of FaceWarehouse[56] basis, and others bases are from BFM17[7].

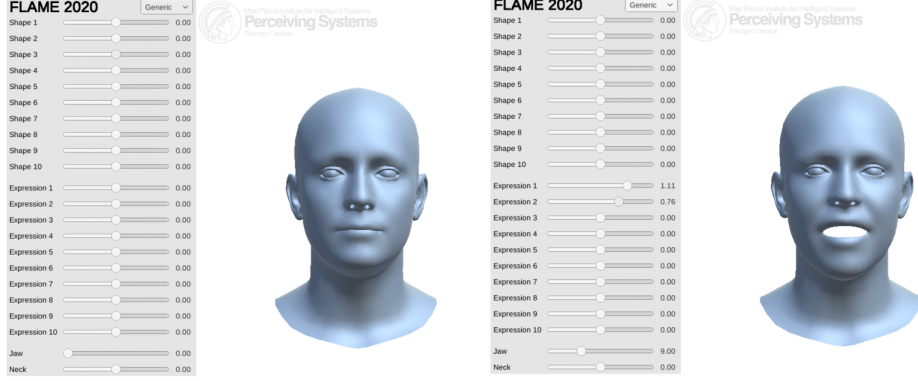


Figure 3. Comparison between results of neutral expression and one random expression, the interactive system is provided by[57] and the face model is FLAME[6]. There are several bars in the left pane to controls the coefficients of 3D face model bases, the right image is rendering results under current parameters settings. The only difference between these two results is the value of first and second expression coefficients and jaw coefficients. Jaw coefficients is unique for FLAME face model, while contained in generic expression components for other face model.

It is worth noting that the expression coefficients in 3DMM mainly controls the dynamic change of lip area of a 3D mesh, so mapping from speech to 3D face expression coefficients can achieve the purpose of face animation. Figure 3 gives an intuition of changing expression coefficients leads to the structural change of lip area.

3.2.3 SPEECH-TO-EXPRESSION MAPPING BASED ON MULTI-MODAL RELEVANCE

In the task of audio-driven facial animation generation, how to align the audio to expression space is the most crucial problem. Experiments suggest that current sequence model used in this mapping stage has insufficient computational efficiency and problems in short-term and long-term dependency extraction, which leads to unsmoothness in generated video. To solve the above shortcomings, after the audio feature extraction stage, we adopt a Transformer-based sequence generation model to achieve the mapping from audio to expression. In addition, the current face animation method mostly does not effectively use the image-level information in reference video, while the latter contains a lot of information about the speaking style of the character.

We explicitly model the relationship between the modal characteristics of the source audio and the reference video, so as to establish a mapping from the source audio to the expression that is more in line with the speaker's style.

There are three types of relevance to be considered, audio-audio single-modal double-sequence relevance, expression-expression single-modal single-sequence relevance and audio-expression cross-modal double-sequence correlation. Each correlation modeling is carried out through a Transformer model, which lead to three Transformer models, namely Audio Transformer $f_{audio}(mfcc)$, Expression Transformer $f_{exp}(\beta)$, and Multi-modal Transformer $f_{cross}(h^{mfcc}, h^\beta)$ in our entire model $f_{trans} = (f_{audio}, f_{exp}, f_{cross})$.

The core idea of Transformer is attention, which conduct by self-attention layers, each self-attention layer includes a multi-head self-attention layer and a feed-forward layer; the relevant representation is calculated by involving relevance computation in multi-head attention layer. Specifically, for input sequence $X = (x_1, x_2, \dots, x_n) \in R^{n \times m}$ where n and m refers to sequence length and dimension of feature respectively. The general calculation formula for hidden representation is:

$$X' = \frac{q \times k}{\sqrt{m}} \times v, \quad X' \in R^{n \times m},$$

where q, k, v all represent input sequence in self-attention manner. $\times$ here represents matrix multiplication.

The expression-expression relevance calculation uses the expression components β of 3D coefficients X , obtained by 3D face reconstruction of the reference video, which work as the input to f_{exp} to obtain the implicit sequence-aware representation of the input expression coefficients sequence. The purpose of calculating self-relevance of facial expression coefficients is to add sequence information to the input expression parameters to obtain expression hidden vectors. By adding expression self-relevance constraints, we introduce the dynamic change information of expression which is implicit in the reference video into the training and prediction stages, which can improve the accuracy and coherence of the prediction results, and make the speaking style of the generated results more similar to it in the reference video. The output of f_{exp} is $h_{1:n}^x$, which is the hidden representation of input expression sequence $\beta_{1:n}$. The calculation formula of f_{exp} is as follows:

$$f_{exp}(\beta_{1:n}) = h_{1:n}^\beta$$

Audio-audio correlation is mainly used to establish the correlation between source audio and reference audio. The intuition behind modelling the relevance of source and reference

audio is that, on the one hand, the movement of the articulating organ is not only determined by the current phoneme, but also determined by the forward and backward phonemes, that is, the coordinated pronunciation of the forward and backward has a great influence on the movement trajectory of the articulating organ corresponding to the current phoneme. What makes the predication of acoustic lip movement more suitable for N-to-1 mapping rather one-to-one mapping, that is why we model the relevance between audio frames of the same input sequence. On the other hand, modeling the relevance between source audio and reference audio provides additional reference information for audio to expression mapping. Given concatenated audio features $mfcc_{1:m+n}$, The calculation formula of audio-audio autocorrelation is as follows:

$$f_{audio}(mfcc_{1:m+n}) = h_{1:m+n}^{mfcc}$$

Audio-expression cross-modal correlation is used to improve the accuracy of audio-to-expression mapping and the continuity of the output parameter sequence, since reference audio and reference video come from the same media resource, which make them own inherent synchronization properties. We hope to calculate the relevance score based on the correlation between source audio and reference audio, then rely on the relevance score and the synchronization of reference audio and reference video to enhance the accuracy of inference results. It is implemented by a Multi-modal Transformer f_{cross} , which completes audio-to-expression mapping based on the correlation between multiple modalities. The audio-expression calculation formula is as follows.

$$f_{cross}(h_{1:m+n}^{mfcc}, h_{1:n}^{\beta}) = \hat{\beta}_{1:m}$$

Before computing the multi-modal relevance, we first concatenate audio and video hidden vector $h_{1:m+n}^{mfcc}, h_{1:n}^{\beta}$ to get a composite vector $x_{1:m+2n}^{mfcc,\beta}$, then self-attention is computed on this vector. After that, we obtain a hidden vector which contains multi-modal information $h_{1:m+2n}^{mfcc,\beta}$, since we only need to predict expression coefficients in according with reference audio with length m , we reserve part of output which consider source audio as q and reference features as k during self-attention calculation stage, which result in partial of hidden vector $h_{1:m}^{mfcc,\beta}$, then this hidden vector pass through a linear layer, to obtain the predicted expression coefficients $\hat{\beta}_{1:m}$.

All Transformers in our experiment follows the architecture of [17], each Transformer has 10 Transformer encoding layers which consist of a multi-head attention layer and a

positional feed-forward layer, a learnable positional encoding is used in our experiment. Each multi-head attention layer has 10 attention heads, the hidden size of layer is 800.

We measure the L2 discrepancy between predicted expression coefficients $\hat{\beta}_{1:m}$ computed by multi-modal relevance and expression coefficients $\beta_{1:m}$ extracted from visual stream, the expression loss we design is as follows

$$\mathcal{L}_{\text{exp}}(X) = \frac{1}{m} \sum_{i=1}^m \|\beta_i - \hat{\beta}_i\|_2$$

While expression loss adds constraints on motion of human faces, it lacks control of lip motion, which lead to a low loss value while lip motion is not in accordance with audio content. The reason behind this might be expression coefficients reflect the dynamic change of total face, and the weights of basis which control non-lip region is relatively high, therefore a low L2 expression loss do not mean a better accuracy of lip motion. That is why we add a lip landmark loss to encourage the lip shape accuracy. After obtained 3D coefficients of human portraits, a 3D mesh $\mathcal{X}$ in line with this coefficient can be obtained by incorporating a 3D face model, then the facial landmark $\ell(\mathcal{X})$ of 3D mesh $\mathcal{X}$ can be computed by project function. We denote the landmarks extracted by face detection method[58] as s_i , then the landmark loss can be formulated as:

$$\mathcal{L}_{\text{mark}}(\mathcal{X}) = \frac{1}{m} \sum_{i=1}^m \|\ell(\mathcal{X}_i) - s_i\|_2$$

It is worth mentioning that, we only impose this constraint on lip landmarks, which is a subset of human facial 68 landmarks.

Furthermore, to improve the continuity of adjacent frames in generated results, we impose a MSE smooth loss constraint on predicted expression coefficients.

$$\mathcal{L}_{\text{smooth}}(\mathcal{X}) = \frac{1}{m-1} \sum_{t=1}^{m-1} \|\hat{\beta}_{(t+1)} - \hat{\beta}_t\|_2$$

In conclusion, we use three loss function to generate accurate lip motion and temporal smooth results, the total loss function during mapping stage can be formulated as:

$$\mathcal{L}_{\text{trans}} = \lambda_1 \mathcal{L}_{\text{exp}} + \lambda_2 \mathcal{L}_{\text{mark}} + \lambda_3 \mathcal{L}_{\text{smooth}}$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weights of these individual loss functions, in our experiment, they are set to 1, 100 and 0.01 respectively. We use Adam optimizer with default setting for 20 epochs with a batch size of 4, the learning rate is fixed at 0.001.

3.2.4 PHOTOREALISTIC 2D IMAGE GENERATION BASED ON NEURAL RENDERING

Neural rendering refers to a technology that uses deep neural networks for image or video generation under conditions that allow explicit or implicit control of the attributes of the 3D scene[16] (such as lighting, camera parameters, pose, geometric structure, etc.). Neural rendering leverage neural networks to simulate complex physical rendering processes to obtain high quality 2D images. In our task, this technique is used to generate visualized 2D results.

Although the demand for high-quality rendering is increasing, there is still no way to directly obtain realistic 2D images from 3D face meshes, limited by the current expressive capabilities of 3D face models. One alternative method and also which used by many existing works rely on physical rasterization to get a coarse 2D face image. After that, a neural network is used to optimize the coarse result to obtain a fine-grained 2D face image. We will use neural rendering to optimize the preliminary results to generate photo-realistic images.

We adopt the neural rendering network of [52], which renders the lower face of coarse 2D input, the lower face is obtained by a face mask to cover the lower face region. After rendering process, we obtain a photorealistic 2D lower face. Then the photorealistic lower face is pasted to the original face image according to the mask region. This process is depicted in Figure 4.

It is worth mentioning that, the neural rendering network is person-specific, which means the training corpus should only contain single person. We create a training corpus on single relatively long video, For the input video, we extract target video frame I and corresponding synthetic frame $\hat{I}$, which was synthesized by 3D face reconstruction and physical rendering methods mentioned in the previous section, lead to a training corpus $\{(I_t, \hat{I}_t)\}_{t=1}^M$, where M denotes the length of input frames.

The neural rendering network consists of a U-Net based generator $\mathcal{G}$ and a discriminator $\mathcal{D}$, which were trained together in an adversarial manner. The generator is composed of an encoder and a decoder. The encoder downsamples input image to a low-dimensional latent representation by several convolutional layers accompanied with batch normalization layers and Leaky ReLU. The decoder is used to upsamples the latent representation and generate photorealistic lower faces. The architecture of decoder is similar to PatchGAN[], which utilize transpose convolution, batch normalization, dropout and ReLU.

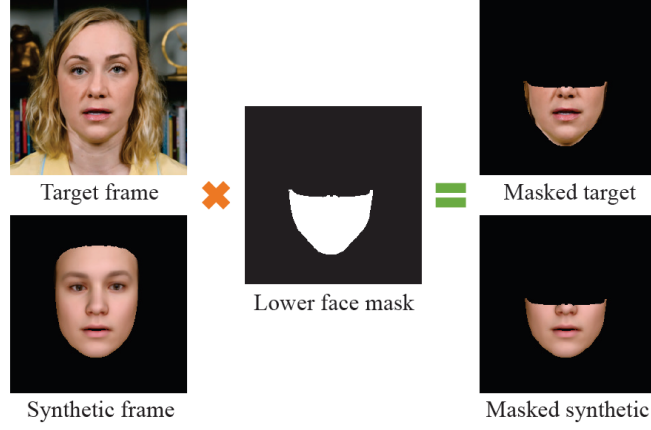


Figure 4. we employ a lower face mask to force neural rendering network focusing on the generation of lip nearby region, the idea is from AudioDVP[52]. Applying face masks for both synthetic frame and target frame, we obtain pair of training data which enable our neural rendering network to refine coarse input

The generator and discriminator of our neural rendering network are optimized by the following loss function:

$$\mathcal{L}(\mathcal{G}, \mathcal{D}) = \mathcal{L}_{\text{rec}}(\mathcal{G}) + \lambda \mathcal{L}_{\text{adv}}(\mathcal{G}, \mathcal{D})$$

The photometric reconstruction loss $\mathcal{L}_{\text{rec}}(\mathcal{G})$ encourage the generator to generate photo-realistic face image, which can be formulated as follows:

$$\mathcal{L}_{\text{rec}}(\mathcal{G}) = \|I_t - \mathcal{G}(\mathbf{T}_t)\|_1$$

Here $T_t = \{\hat{I}_t\}_{i=t-Nw}^{t+Nw}$ refers to a stack of tensor, which is helpful for maintain coherence in generated results.

The GAN adversarial loss can be formulated as follows, here we use the vanilla GAN loss:

$$\mathcal{L}_{\text{adv}}(\mathcal{G}, \mathcal{D}) = \log \mathcal{D}(I_t) + \log(1 - \mathcal{D}(\mathcal{G}(\mathbf{T}_t)))$$

The neural rendering network is trained separately from Transformers f_{trans} , we use Adam optimizer, the network is initialized with normal distribution, the training process takes 300 epochs with batch size set to 8 and learning rate of 0.002.

CHAPTER 4: DESIGN AND IMPLEMENTATION

4.1 DESIGN

Virtual avatar is a technology widely discussed recently, it has been widely used in practical applications such as 3D games, virtual idols, and virtual meetings. In these applications, how to drive the virtual avatar so that it can give feedback similar to real person is a major problem. Audio-driven facial animation, as one of the solutions to this problem, uses the

correlation between voice and facial expressions to complete the driving process of the virtual avatar. There are many researches on audio-driven facial animation according to different type of inputs. In our project, we mainly consider the situation where the input is audio and video (rather than single image). In this context, there are mainly two types of data, voice data and face data in video. With regard to voice data, we need to pay more attention to semantic-related information, while less attention to who the speaker is. Similarly, as for face data, we mainly pay attention to dynamic information (expression) of the face, especially the movement of the lips, and less attention to the identity information of the person. The pipeline of audio-driven facial animation usually includes speech feature extraction, facial feature extraction, multi-modal feature fusion (or audio-expression mapping), and face image generation.

In terms of input data, traditional methods lack effective use of input video information. Specifically, the traditional method usually takes a neural network and takes speech as input and takes facial images as supervision information, the neural networks accomplish mapping from speech features to facial expressions, which can be seen as a one-to-one mapping. In fact, input voice and video own sequence characteristics which can be helpful for audio-expression mapping. How to effectively use this sequence information to assist in generating audio-lip synchronization and realistic video results is the main challenge of our research. To solve above challenge, we designed a many-to-many training strategy, that is, the network input is no longer a single audio-face pair in input sequence, but takes the entire sequence as the input of the neural network in each step. The purpose of this is to make full use of the learning ability of the neural network and take advantage of the correlation information in the input sequence to generate more accurate and sequential results. In order to obtain supervision information, we divide the input sequence into two parts, the first part is used as the seed input to calculate sequential relevance during training, while the other part is used as driving input(audio) and supervision information(face).

To build sequential dependency, we use Transformer network, since Transformer shows advantages over sequence networks such as RNN and LSTM in many tasks[19][20], the superiority of Transformer includes the ability to extract long-term dependency in sequence and parallel computing over entire sequence to improve the speed of inference. Therefore, we use the Transformer network to extract sequence-related features in the meanwhile complete cross-modal feature mapping.

In terms of procedures design, we use librosa python library to extract the mfcc features of speech as the feature representation of speech. This is mainly due to the fact that mfcc features have been proven to capture the dynamic changes of speech in a short duration in many current researches, which can represent semantic information of speech. As for face representation, we use BFM face model based on 3DMM. Compared with 2D face representation (such as facial landmarks). On the one hand, the expression basis of 3DMM reflects the change of the character’s lip shape, which can store dynamic information of the face. on the other hand, 3DMM provides a decoupled representation space for faces, allowing operations on specific components of the face without affecting other components, which avoids the problem of identity loss come from 2D face representation in mapping stage. In audio-expression mapping stage, we use multiple Transformers to extract features with sequential information, and perform cross-modal mapping based on these features to obtain smooth results. In image generation stage, we use OpenGL and a neural rendering network to generate photorealistic face images. OpenGL as a physical rendering technology, is used to render 3D mesh and generate a face image with a specific camera position, but limited by the low quality of the 3D mesh texture, the resulting rendering result is far from photorealistic. In order to obtain realistic face image results, we use the neural rendering network proposed by [52], which can generate photorealistic face images based on the coarse face images obtained by the physical rendering.

4.2 IMPLEMENTATION

Here we describe the implementation detail of our experiment. The structure of this section is as follows. First, we explain the dataset used in this experiment and why we choose it. Then we explain the preprocessing procedure to process data. Finally, we describe the implementation details of both training and evaluation.

In our all experiments, the input of our model contains a piece of audio used to drive specific person (hereinafter referred to as source audio), and a short speaking video which contains only one speaker throughout the entire video (hereinafter referred to as reference video). The dataset used in our method is LRW[59], which is a public large-scale visual speech recognition dataset of more than 1000 identities, Figure 5 shows frames extracted from this dataset among different identities, since this dataset is designed for lip reading task, it contains paired audio and visual data in synchronization with each other, furthermore, they process their collected data with a multi-stage pipeline including shot detection, face

detection, face tracking etc. to remove irrelevant factors in videos which make it suitable for our task.



Figure 5. LRW is a large-scale speaking video dataset which contains speaking videos of thousands of identities, which usually used for tasks concerning of multi-modal such as lip-reading or audio-driven facial animation

The raw data in LRW dataset are speaking videos while the input to our system are paired audio file and video frames, which requires further preprocessing, to separate audio and visual information contained in reference video, audio stream and visual stream are extracted independently by ffmpeg[60] video processing util, result in an audio file(hereinafter referred to as reference audio) and video frames(also called reference frames). All videos in LRW dataset last for 1.16s, we extract video frames at a rate of 25, result in 29 frames for each video. For audio file extraction, we use the default parameters of ffmpeg and obtain audio last for 1.16s, which keep the same with video duration.

To obtain audio feature representation, we extract 13-dim mfcc feature of all audio files and concatenate them over time dimension, result in a sequence of audio frames whose length keeps the same with video frames. Since no 3D labels in our dataset, to train our network, we adopt 3D Face Reconstruction to obtain 3D labels from 2D images. All video frames are fed to our 3D Face Reconstruction network to extract 3DMM coefficients of 257 dimension and being regarded as ground-truth for audio-to-expression mapping.

Our model trained in a self-supervision manner since no 3D ground-truth information contained in our dataset. After obtained audio feature representation and visual feature representation in data preprocessing stage, both of them are fed into our Transformers to compute multi-modal relevance. During training phase, source audio, reference audio and reference frames are all come from the same video, a N-supervision strategy is adopted in our experiment to split source and reference sequence. Specifically, for a input video with 29 frames,

we take the former 20 frames as reference input and the rest 9 frames as source input, lead to 3 sequences of input with length 9, 20, 20 for source audio, reference audio and reference frames respectively, as for the rest 9 video frames not used in reference input, since they are in synchronization with source audio, which can be used as the ground-truth for our audio-to-expression mapping in a self-supervision manner. During evaluation phase, reference audio and reference frames come from the same video, while the source audio is provided additionally.

Specifically, source audio features of length m and reference audio features of length n are concatenate along sequence dimension, result in an audio feature sequence of length $m + n$. The combined audio feature sequence is sent to Audio Transformer to compute self-relevance along sequence dimension, thus generate relevance-based audio features. Self-relevance of expression features is computed in parallel by feeding reference frame features of length n into Expression Transformer, Next, both output feature sequences are concatenate along time dimension then feed to Multi-modal Transformer to fuse multi-modal information based on relevance scores computed by multi-head attention of consecutive features, after computing multi-head attention of input sequences, the linear layer in Multi-modal Transformer carry out feature mapping, which maps source audio to predicted expressions parameters. During training, we supervise the predicted expression coefficients sequence with, while during inference, the predicted expressions are in accordance with source audio.

Our neural rendering network is trained individually, to obtain paired coarse and fine-grained 2D input, we first extract face images in LRW dataset, and then use 3D face reconstruction and OpenGL renderer to obtain 2D coarse rendering results. The network is trained to recover find-grained results from coarse input with lower face mask.

We implement our method using PyTorch, all experiments are conducted on single PC with 2 NVIDIA GeForce RTX 2080Ti.

CHAPTER 5: RESULTS AND DISCUSSION / EVALUATION

In this Chapter, we demonstrate the evaluation details of our proposed method, in our experiment, we use both qualitative method and quantitative method.

5.1 QUALITATIVE EVALUATION

Since the quality of facial animation is a relatively subjective judgement, to eliminate subjective factor to some extent, we incorporate two methods to evaluate the result by

comparison. First is self-reenactment, where source audio and reference audio are the same, it can reflect the quality of audio to expression mapping in our model by comparing with the original video, the result video generated by our method should be similar to the original video which can be seen as ground truth. The other is multi-target reenactment, which use the same driving video to animate multiple characters, it can reflect the stability of our method, the lip movement across multiple characters should similar to each other since the driving audio are the same for these characters, furthermore, multi-target reenactment can also reflect the accuracy of lip motion given the visual stream correspond to audio stream.

In Figure 6, we depict the results of self-reenactment generated by our proposed method. There are two rows for each identity, the upper row represents images cropped from original video, where lip motions are inherent in line with audio content. the lower row represents images generated by our method driving by its audio content.



Figure 6. self-reenactment results of our proposed method, there are two rows concerning of each identity, the first are images cropped from source video while the second are images generated by our method given audio signal. It can be seen that our generated lip movement are quite similar to the original video with high quality.

In Figure 7, we show the results of multi-target reenactment, to compare the lip motion across different identities, we select speaking videos of 4 characters as reference video, which all collected from YouTube. In this figure, top row shows source video frames that are consistent with source audio, the source audio is used to drive different characters together with individual reference video of each character. Since the driving audios are the same for all identities, the lip motion among identities should be similar, while speaking style is remained

for each individual due to the original speaking video is provided during inference time. For more multi-target reenactment results, please refers to the Appendix.



Figure 7. multi-target reenactment results of our proposed method, here the first row shows source images in accordance with driving audio. The lip movement can be seen as the ideal target. Other rows are images generated by our method given the same audio as the first row. It can be seen that the lip movement generated by our method are quite realistic.

5.2 QUANTITATIVE EVALUATION

The main goal of our proposed method is to map audio to lip motions with better accuracy by introducing multi-modal relevance. During evaluation stage, lip motion accuracy is the most important evaluation criteria. To quantitative analysis the performance of lip motions, we calculate LSE-E(Lip Sync Error – Expression), which compute MSE between expression coefficients $\hat{\beta}$ predicted from audio and 3D expression coefficients reconstructed from video β which can be regarded as ground-truth lip motion. The calculation of LSE-E can be formulated as:

$$\text{LSE-E}(\hat{\beta}, \beta) = \frac{1}{m} \sum_{i=1}^m \|\beta_i - \hat{\beta}_i\|_2$$

where m refers to the length of generated frames. This metric reflects the accuracy of lip shape in frame level.

Since 3D expression coefficients involve the movement of non-lip region, which do not directly reflect the lip motion accuracy, we further compare LSE-L(Lip Sync Error - Landmark), which calculate the MSE error of lip landmarks as an additional criterial, it can be formulated as:

$$\text{LSE-L}(\mathcal{X}, s) = \frac{1}{n} \sum_{j=1}^n \|\mathcal{X}_j - s_j\|_2, n \in \mathcal{R}$$

Where n refers the index of landmarks, $\mathcal{R}$ represent the landmarks which in association with lip region.

The results of above mentioned criterial is showed in Table 1, a lower value is better for both of them.

Table 1. quantitative evaluation of our proposed method, we outperform all other methods in several metrics, LSE-D, LSE-E and LSE-L while obtain relatively good results in LSE-C and FID.

Method	LSE-D ↓	LSE-C ↑	FID ↓	LSE-E ↓	LSE-L ↓
Speech2Vid	13.56	1.928	11.15	0.0739	0.7362
LipGAN	11.16	3.650	2.726	0.0534	0.5268
Wav2Lip	6.512	7.490	3.189	0.0273	0.3606
AudioDVP	6.732	6.919	2.947	0.0298	0.3877
Ours	6.217	6.845	3.093	0.0260	0.3554
Real Videos	5.324	6.931	—	—	—

Audio-lip synchronization is another key objective for facial animation. Chung et al. [64] proposed a two-stream network called SyncNet to measure audio-lip synchronization of video. SyncNet accept an audio and video segments as input, and output two values namely LSE-D(Lip Sync Error - Distance) and LSE-C(Lip Sync Error - Confidence), the former measure the distance between lip and audio representation in a shared space, which can be used to evaluate the matching degree between audio and video, a lower value is better. The latter criteria measure the correlation between audio-visual streams, the lower confidence score means there exists several portiones of audio-visual pairs are out-of-sync. Further details of theses metrics can be found in [54]. Here we employ pretrained SyncNet to verify the audio-lip synchronization of generated videos.

We also report the Fréchet Inception Distance (FID)[61] score which is a common metric used to evaluate the image quality in image generation task.

We compare the results of our method with several state-of-the-art audio-driving talking face generation methods on LRW test dataset, including Speech2Vid[38], LipGAN[62], Wav2Lip[47] and AudioDVP[52], the results demonstrate our method outperforms all of them in LSE-D, LSE-E and LSE-L which prove the effectiveness of our proposed method in generating audio-visual synchronized talking videos.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

In this dissertation, we propose a novel neural network architecture to accomplish the task of audio-driven facial animation. Our system is designed for specific situation in video portraits, when a short talking video is provided. To take full advantage of input modalities, especially the inherent synchronization in reference video. We proposed to use multi-Transformer architecture to compute the multi-modal relevance of input, which can accelerate both the training and evaluating process by its intrinsic parallel computing characteristics, based on this, a precise audio to expression mapping is performed to reenact human portraits. By incorporating a neural rendering network, photorealistic portrait videos can be generated. Both qualitative and quantitative experiments prove the effectiveness of our method. Nevertheless, our approach has several limitations for practical use, for example, our system requires a relatively long video clip when animating new character, this is the demand for both multi-modal relevance learning and neural rendering, the duration of reference video determine the quality of relevance computing and the plausibility of generated results, which is inevitable for our designed system. Another limitation is that we have to pursuit the balance between lip accuracy and the smoothness of lip motion, since human lip motions are highly unpredictable and have strong randomness, even providing audio signal as condition, pursue the accuracy of lip shape in every moment excessively will bring about problem such as inconsecutive movement of lip, an alternative solution and exactly what we do is lower the accuracy to some extent while ensure a level of smoothness in generated results.

Our proposed system can be used to generate video portraits for both 2D space and 3D space, under the condition that the avatar can be fitting to a 3D Face Model. The technique of audio-driven facial animation has broad prospects for development, since it provides the ability to generate motion segment of 3D avatar, which can serve as the underlying technology for VR/AR applications such as virtual meeting, distance education and so on.

REFERENCES

- [1] Blanz, Volker, and Thomas Vetter. "A morphable model for the synthesis of 3D faces." *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 1999.
- [2] Abrevaya, Victoria Fernández, Stefanie Wuhler, and Edmond Boyer. "Multilinear auto-encoder for 3d face model learning." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.
- [3] Bagautdinov, Timur, et al. "Modeling facial geometry using compositional vaes." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.
- [4] Lombardi, Stephen, et al. "Deep appearance models for face rendering." *ACM Transactions on Graphics (TOG)* 37.4 (2018): 1-13.
- [5] Ranjan, Anurag, et al. "Generating 3D faces using convolutional mesh autoencoders." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
- [6] Li, Tianye, et al. "Learning a model of facial shape and expression from 4D scans." *ACM Trans. Graph.* 36.6 (2017): 194-1.
- [7] Gerig, Thomas, et al. "Morphable face models—an open framework." *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018.
- [8] Egger, Bernhard, et al. "3d morphable face models—past, present, and future." *ACM Transactions on Graphics (TOG)* 39.5 (2020): 1-38.
- [9] Tuan Tran, Anh, et al. "Regressing robust and discriminative 3D morphable models with a very deep neural network." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
- [10] Chang, Feng-Ju, et al. "Expnet: Landmark-free, deep, 3d facial expressions." *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018.
- [11] Chang, Feng-Ju, et al. "Faceposenet: Making a case for landmark-free face alignment." *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017.
- [12] Deng, Yu, et al. "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

- [13] Zollhöfer, Michael, et al. "State of the art on monocular 3D face reconstruction, tracking, and applications." *Computer Graphics Forum*. Vol. 37. No. 2. 2018.
- [14] Blanz, Volker, et al. "Reanimating faces in images and video." *Computer graphics forum*. Vol. 22. No. 3. Oxford, UK: Blackwell Publishing, Inc, 2003.
- [15] Thies, Justus, Michael Zollhöfer, and Matthias Nießner. "Deferred neural rendering: Image synthesis using neural textures." *ACM Transactions on Graphics (TOG)* 38.4 (2019): 1-12.
- [16] Tewari, Ayush, et al. "State of the art on neural rendering." *Computer Graphics Forum*. Vol. 39. No. 2. 2020.
- [17] Vaswani, Ashish, et al. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).
- [18] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
- [19] Zhou, Haoyi, et al. "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting." *arXiv preprint arXiv:2012.07436* (2020).
- [20] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).
- [21] Carion, Nicolas, et al. "End-to-end object detection with transformers." *European Conference on Computer Vision*. Springer, Cham, 2020.
- [22] Chen, Hanqing, et al. "Pre-trained image processing transformer." *arXiv preprint arXiv:2012.00364* (2020).
- [23] Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." *arXiv preprint arXiv:2103.14030* (2021).
- [24] Liu, Zhouyong, et al. "ConvTransformer: A Convolutional Transformer Network for Video Frame Synthesis." *arXiv preprint arXiv:2011.10185* (2020).
- [25] Zheng, Chen, Quan Guo, and Parisa Kordjamshidi. "Cross-Modality Relevance for Reasoning on Language and Vision." *arXiv preprint arXiv:2005.06035* (2020).
- [26] Li, Ruilong, et al. "Learn to Dance with AIST++: Music Conditioned 3D Dance Generation." *arXiv preprint arXiv:2101.08779* (2021).
- [27] Han, Kai, et al. "A Survey on Visual Transformer." *arXiv preprint arXiv:2012.12556* (2020).
- [28] Vig, Jesse, and Yonatan Belinkov. "Analyzing the structure of attention in a transformer language model." *arXiv preprint arXiv:1906.04284* (2019).

- [29] Voita, Elena, et al. "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned." *arXiv preprint arXiv:1905.09418* (2019).
- [30] Chefer, Hila, Shir Gur, and Lior Wolf. "Transformer Interpretability Beyond Attention Visualization." *arXiv preprint arXiv:2012.09838* (2020).
- [31] Thies, Justus, et al. "Face2face: Real-time face capture and reenactment of rgb videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
- [32] Kim, Hyeongwoo, et al. "Deep video portraits." *ACM Transactions on Graphics (TOG)* 37.4 (2018): 1-14.
- [33] Gafni, Guy, et al. "Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction." *arXiv preprint arXiv:2012.03065* (2020).
- [34] Jiang, Zi-Hang, et al. "Disentangled representation learning for 3D face shape." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- [35] Siarohin, Aliaksandr, et al. "First order motion model for image animation." *Advances in Neural Information Processing Systems* 32 (2019): 7137-7147.
- [36] Zakharov, Egor, et al. "Few-shot adversarial learning of realistic neural talking head models." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019.
- [37] Yao, Guangming, et al. "Mesh Guided One-shot Face Reenactment Using Graph Convolutional Networks." *Proceedings of the 28th ACM International Conference on Multimedia*. 2020.
- [38] Jamaludin, Amir, Joon Son Chung, and Andrew Zisserman. "You said that?: Synthesising talking faces from audio." *International Journal of Computer Vision* 127.11 (2019): 1767-1779.
- [39] Suwajanakorn, Supasorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. "Synthesizing obama: learning lip sync from audio." *ACM Transactions on Graphics (ToG)* 36.4 (2017): 1-13.
- [40] Wiles, Olivia, A. Koepke, and Andrew Zisserman. "X2face: A network for controlling face generation using images, audio, and pose codes." *Proceedings of the European conference on computer vision (ECCV)*. 2018.
- [41] Eskimez, Sefik Emre, et al. "Generating talking face landmarks from speech." *International Conference on Latent Variable Analysis and Signal Separation*. Springer, Cham, 2018.

- [42] Vougioukas, Konstantinos, Stavros Petridis, and Maja Pantic. "End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs." *CVPR Workshops*. 2019.
- [43] Zhang, Jiangning, et al. "APB2Face: Audio-guided face reenactment with auxiliary pose and blink signals." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
- [44] Zhang, Jiangning, et al. "APB2FaceV2: Real-Time Audio-Guided Multi-Face Reenactment." *arXiv preprint arXiv:2010.13017* (2020).
- [45] Zhou, Yang, et al. "MakeltTalk: speaker-aware talking-head animation." *ACM Transactions on Graphics (TOG)* 39.6 (2020): 1-15.
- [46] Yu, Lingyun, et al. "Multimodal inputs driven talking face generation with spatial-temporal dependency." *IEEE Transactions on Circuits and Systems for Video Technology* (2020).
- [47] Prajwal, K. R., et al. "A lip sync expert is all you need for speech to lip generation in the wild." *Proceedings of the 28th ACM International Conference on Multimedia*. 2020.
- [48] Karras, Tero, et al. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." *ACM Transactions on Graphics (TOG)* 36.4 (2017): 1-12.
- [49] Cudeiro, Daniel, et al. "Capture, learning, and synthesis of 3D speaking styles." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- [50] Song, Linsen, et al. "Everybody's Talkin': Let Me Talk as You Want." *arXiv preprint arXiv:2001.05201* (2020).
- [51] Yi, Ran, et al. "Audio-driven talking face video generation with natural head pose." *arXiv preprint arXiv:2002.10137* (2020).
- [52] Wen, Xin, et al. "Photorealistic Audio-driven Video Portraits." *IEEE Transactions on Visualization and Computer Graphics* 26.12 (2020): 3457-3466.
- [53] Richard, Alexander, et al. "MeshTalk: 3D Face Animation from Speech using Cross-Modality Disentanglement." *arXiv preprint arXiv:2104.08223* (2021).
- [54] Chung, Joon Son, and Andrew Zisserman. "Out of time: automated lip sync in the wild." *Asian conference on computer vision*. Springer, Cham, 2016.
- [55] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. 8).

- [56] Cao, Chen, et al. "Facewarehouse: A 3d facial expression database for visual computing." *IEEE Transactions on Visualization and Computer Graphics* 20.3 (2013): 413-425.
- [57] FLAME Interactivate System. <https://flame.is.tue.mpg.de/interactivemodelviewer.html>
- [58] King, Davis E. "Dlib-ml: A machine learning toolkit." *The Journal of Machine Learning Research* 10 (2009): 1755-1758.
- [59] Chung, Joon Son, and Andrew Zisserman. "Lip reading in the wild." *Asian conference on computer vision*. Springer, Cham, 2016.
- [60] Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, 2006(146), 10.
- [61] Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." *Advances in neural information processing systems* 30 (2017).
- [62] Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. 2019. Towards Automatic Face-to-Face Translation. In Proceedings of the 27th ACM International Conference on Multimedia. ACM, 1428–1436.

APPENDICES

1. ETHICS FORM

SECTION A: About You (Principal Researcher)

1	Full Name:	Yizhen Ma			
2	Tick all boxes which apply:	Member of staff:	☐	Honorary research fellow:	☐
3	Undergraduate Student	☐	Taught Postgraduate Student	☐	Postgraduate Research Student
					☒
4	Institute/Academic Discipline/Centre:	University of Wales Trinity Saint David			
5	Campus:	Swansea			
6	E-mail address:	2006703@student.uwtsd.ac.uk			
7	Contact Telephone Number:	+86 18105565949			
For students:					
8	Student Number:	2006703			
9	Programme of Study:	MSc Software Engineering (Web Development) - Swansea (September - Full Time)			
10	Director of Studies/Supervisor:	Gordon Dickers			

SECTION B: Approval for Research Activity

1	Has the research activity received approval in principle? (please check the Guidance Notes as to the appropriate approval process for different levels of research by different categories of individual)	YES	☒	NO	☐
					Date
2	If Yes, please indicate source of approval (and date where known): Approval in principle must be obtained from the relevant source prior to seeking ethical approval	Research Degrees Committee	☒		
3		Institute Research Committee	☐		
4		Other (write in)	☐		

SECTION C: Internal and External Ethical Guidance Materials

	Please list the core ethical guidance documents that have been referred to during the completion of this form (including any discipline-specific codes of research ethics, and also any specific ethical guidance relating to the proposed methodology). Please tick to confirm that your research proposal adheres to these codes and guidelines.	
1	UWTSd Research Ethics & Integrity Code of Practice	☐

2	UWTSD Research Data Management Policy	☐
3		☐

SECTION D: External Collaborative Research Activity

1	Does the research activity involve collaborators outside of the University?	YES	☐	NO	☒	
2	If Yes, please provide the name of the external organisation and name and contact details for the main contact person and confirmation this person has consented to their personal data being shared.as part of this collaboration.					
3	Institution					
4	Contact person name					
5	Contact person e-mail address					
6	Has this individual consented to sharing their details on this form?	YES	☐	NO	☐	
7	Are you in receipt of a KESS scholarship?	YES	☐	NO	☐	
8	Is your research externally funded	YES	☐	NO	☐	
9	Are you specifically employed to undertake this research in either a paid or voluntary capacity?	Voluntary	YES	☐	NO	☐
10		Employed	YES	☐	NO	☐
11	Is the research being undertaken within an existing UWTSD Athrofa Professional Learning Partnership (APLP)	If YES then the permission question below does not need to be answered.	YES	☐	NO	☐
12	Permission to undertake the research has been provided by the partner organisation	(If YES attach copy) If NO the application cannot continue	YES	☐	NO	☐

Where research activity is carried out in collaboration with an external organisation

13	Does this organisation have its own ethics approval system?	YES	☐	NO	☐
14	If Yes, please attach a copy of any final approval (or interim approval) from the organisation				

SECTION E: Details of Research Activity

1	Indicative title:	Multi-Modal Relevance Based Facial Animation		
2	Proposed start date:	May 2021	Proposed end date:	Nov. 2021
3	Introduction to the Research (maximum 300 words)			

	Ensure that you write for a <u>Non-Specialist Audience</u> when outlining your response to the three points below: • <i>Purpose of Research Activity</i> • <i>Proposed Research Question</i> • <i>Aims of Research Activity</i> • <i>Objectives of Research Activity</i> Demonstrate, briefly, how <u>Existing Research</u> has informed the proposed activity and explain • <i>What the research activity will add to the body of knowledge</i> • <i>How it addresses an area of importance.</i>
4	Purpose of Research Activity 1. Propose a new deep learning method for facial animation generation task to improve the quality of previous methods 2. Introducing a novel structure and a new strategy which make full use of the input data. (this box should expand as you type)
5	Research Question 1. How to improve the video smoothness in facial animation task? 2. How to make full use the correlation between inputs to improve the audio-visual synchronization in result videos? (this box should expand as you type)
6	Aims of Research Activity 1) Evaluating the impact of proposed architecture on results regards to smoothness and naturalism. 2) Assessing the effect of multi-modal correlation strategy in improving audio-visual synchronization (this box should expand as you type)
7	Objectives of Research Activity 1) To explore the relevance between speech and facial expression 2) To explore how attention mechanism improve the performance in the scope of facial animation (this box should expand as you type)
8	Proposed methods (maximum 600 words) Provide a brief summary of all the methods that may be used in the research activity, making it clear what specific techniques may be used. If methods other than those listed in this section are deemed appropriate later, additional ethical approval for those methods will be needed.
9	For data collection, we won't collect any data by ourselves. We will apply for public available talking video dataset and request permission from the organization for using their data in our experiments. For more technical details, we use several methods to process our data. (1) Deep sequence model Due to the shortcomings of the traditional sequence model in terms of computational efficiency and long-term dependency extraction, the resulting video is not smooth. We consider using a transformer-based sequence generation model in the audio feature extraction stage to solve the above shortcomings. (2) 3D morphable models to represent human face

	3D face has better transformation characteristics than a 2D face representation (such as face landmarks), such as better control of attributes like lighting, expression, and posture. In addition, 3D morphable models provides a disentangled representation space for human faces, which provides a better result space to our sequence generation problem. (3) Neural rendering technology Neural rendering refers to a technology that uses deep neural networks for image or video generation under conditions that allow explicit or implicit control of the attributes of the 3D scene (such as lighting, camera parameters, pose, geometric structure, etc.). In our task, this technique is used for visualization. (this box should expand as you type)
--	---

10	Location of research activity Identify all locations where research activity will take place.
11	(this box should expand as you type)
12	Research activity outside of the UK If research activity will take place overseas, you are responsible for ensuring that local ethical considerations are complied with and that the relevant permissions are sought. Specify any local guidelines (e.g. from local professional associations/learned societies/universities) that exist and whether these involve any ethical stipulations beyond those usual in the UK (provide details of any licenses or permissions required). Also specify whether there are any specific ethical issues raised by the local context in which the research activity is taking place, for example, particular cultural and/or legal sensitivities or vulnerabilities of participants.
13	Wuhan China (this box should expand as you type)

14	Use of documentation not in the public domain: Are any documents NOT publicly available?	NO	☒
		YES	☐
15	If Yes, please provide details here of how you will gain access to specific documentation that is not in the public domain and that this is in accordance with prevailing data protection law of the country in question and England and Wales. (this box should expand as you type)		

SECTION F: Scope of Research Activity

1	Will the research activity include:	YES	NO
2	Use of a questionnaire or similar research instrument?	☐	☒
3	Use of interviews?	☐	☒
4	Use of diaries?	☐	☒

5	Participant observation with their knowledge?	☐	☒
6	Participant observation without their knowledge?	☐	☒
7	Use of video or audio recording?	☐	☒
8	Access to personal or confidential information without the participants' specific consent?	☐	☒
9	Administration of any questions, test stimuli, presentation that may be experienced as physically, mentally or emotionally harmful / offensive?	☐	☒
10	Performance of any acts which may cause embarrassment or affect self-esteem?	☐	☒
11	Investigation of participants involved in illegal activities?	☐	☒
12	Use of procedures that involve deception?	☐	☒
13	Administration of any substance, agent or placebo?	☐	☒
14	Working with live vertebrate animals?	☐	☒
15	Other primary data collection methods, please explain in this box For example, 'focus groups'. Please indicate the type of data collection method(s) in this box and tick the accompany box.	☐	☐
16	Details of any other primary data collection method: We will apply for public available dataset with the institution's permission (this box should expand as you type)		

If NO to every question, then the research activity is (ethically) low risk and **may** be exempt from **some** of the following sections (please refer to Guidance Notes).

If YES to any question, then no research activity should be undertaken until full ethical approval has been obtained.

8	Disclosure and Barring Service			
9	If the research activity involves children or vulnerable adults, a Disclosure and Barring Service (DBS) certificate must be obtained before any contact with such participants.	YES	NO	N/A
10	Does your research require you to hold a current DBS Certificate?	☐	☒	☐

SECTION J: Data Protection and Storage

In completing this section refer to the University's Research Data Management Policy and the extensive resources on the University's Research Data Management web pages (<http://uwtsd.ac.uk/library/research-data-management/>).

1	Does the research activity involve personal data (as defined by the General Data Protection Regulation 2016 "GDPR" and the Data Protection Act 2018 "DPA")?	YES	NO
	<i>"Personal data" means any information relating to an identified or identifiable natural person ('data subject'). An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors</i>	☐	☒

	specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.		
2	If YES, provide a description of the data and explain why this data needs to be collected:		
	(this box should expand as you type)		
3	Does it involve special category data (as defined by the GDPR)?	YES	NO
	“Special category data” means sensitive personal data consisting of information as to the data subjects’ – (a) racial or ethnic origin, (b) political opinions, (c) religious beliefs or other beliefs of a similar nature, (d) membership of a trade union (within the meaning of the Trade Union and Labour Relations (Consolidation) Act 1992), (e) physical or mental health or condition, (f) sexual life, (g) genetics, (h) biometric data (as used for ID purposes),	☐	☒
4	If YES, provide a description of the special category data and explain why this data needs to be collected:		
	(this box should expand as you type)		

5	Will the research activity involve storing personal data and/or special category data on one of the following:	YES	NO
6	Manual files (i.e. in paper form)?	☐	☒
7	University computers?	☐	☒
8	Private company computers?	☐	☒
9	Home or other personal computers?	☐	☒
10	Laptop computers/ CDs/ Portable disk-drives/ memory sticks?	☐	☒
11	“Cloud” storage or websites?	☐	☒
12	Other – specify:	☐	☒
13	For all stored data, explain the measures in place to ensure the security of the data collected, data confidentiality, including details of password protection, encryption, anonymisation and pseudonymisation:		
	(this box should expand as you type)		
14	All Data Storage		
15	Will the research activity involve any of the following activities:	YES	NO
16	Electronic transfer of data in any form?	☐	☒
17	Sharing of data with others at the University?	☐	☒
18	Sharing of data with other organisations?	☐	☒

19	Export of data outside the European Union or importing of data from outside the UK?	☐	☒
20	Use of personal addresses, postcodes, faxes, emails or telephone numbers?	☐	☒
21	Publication of data that might allow identification of individuals?	☐	☒
22	Use of data management system?	☐	☒
23	Data archiving?	☐	☒
24	If YES to any question, please provide full details, explaining how this will be conducted in accordance with the GDPR and DPA (and/or any international equivalent):		
	<i>(this box should expand as you type)</i>		
25	List all who will have access to the data generated by the research activity:		
	The principal researcher. <i>(this box should expand as you type)</i>		
26	List who will have control of, and act as custodian(s) for, data generated by the research activity:		
	The principal researcher <i>(this box should expand as you type)</i>		
27	Give details of data storage arrangements, including security measures in place to protect the data, where data will be stored, how long for, and in what form. Will data be archived – if so how and if not why not.		
	All data will be encrypted and kept in password protected cloud storage on the University Office 365 system which will not be shared. Any USB sticks used to store or transfer data will be password protected, and will be reformatted at the end of the project in order to destroy the data. The data will be stored until the completion of the project and then deleted. <i>(this box should expand as you type)</i>		
28	Please indicate if your data will be stored in the UWTSD Research Data Repository (see https://researchdata.uwtsd.ac.uk/). If so please explain. <i>(Most relevant to academic staff)</i>		
	<i>(this box should expand as you type)</i>		
29	Confirm that you have read the UWTSD guidance on data management (see https://www.uwtsd.ac.uk/library/research-data-management/)	YES	☒
		NO	☐
30	Confirm that you are aware that you need to keep all data until after your research has completed or the end of your funding	YES	☒
		NO	☐

SECTION K: Declaration

31	The information which I have provided is correct and complete to the best of my knowledge. I have attempted to identify any risks and issues related to the research activity and acknowledge my obligations and the rights of the participants. In submitting this application I hereby confirm that I undertake to ensure that the above named research activity will meet the University's Research Ethics and Integrity Code of Practice which is published on the website: https://www.uwtsd.ac.uk/research/research-ethics/	
	Signature of applicant: Yizhen Ma	Date: 2021/11/11

2. SOME CRITICAL CODE

The Multi-Transformer model we used in our method

```

self.au_pre_fc = nn.Sequential(
    nn.Linear(d_au_vec, d_embed_vec),
    nn.BatchNorm1d(d_embed_vec),
    nn.ReLU(True),
    nn.Linear(d_embed_vec, d_embed_vec),
    nn.BatchNorm1d(d_embed_vec),
    nn.ReLU(True)
)

self.image_transformer = TransformerEncoderV3(d_input_vec=d_embed_vec, n_layers=10, n_head=10, d_k=d_embed_vec,
                                              d_v=d_embed_vec, d_model=d_embed_vec, d_inner=800,
                                              n_position=n_seed, persistent=persistent)
self.audio_transformer = TransformerEncoderV3(d_input_vec=d_embed_vec, n_layers=10, n_head=10, d_k=d_embed_vec,
                                              d_v=d_embed_vec, d_model=d_embed_vec, d_inner=800,
                                              n_position=n_seed + n_position, persistent=persistent)
self.cross_transformer = TransformerEncoderV3(d_input_vec=d_embed_vec, n_layers=10, n_head=10, d_k=d_embed_vec,
                                              d_v=d_embed_vec, d_model=d_embed_vec, d_inner=800,
                                              n_position=n_seed * 2 + n_position, persistent=persistent)

self.cross_post_fc = nn.Sequential(
    nn.Linear(d_embed_vec, d_coeff_vec),
    nn.BatchNorm1d(d_coeff_vec),
    nn.ReLU(True),
    nn.Linear(d_coeff_vec, d_coeff_vec)
)

```

Part of hyper-parameters default setting of our model

```

''' train arguments'''
parser.add_argument('--device', type=str, default='cuda')
parser.add_argument('--device_ids', type=str, default='0,1,2')
parser.add_argument('--stage', type=str, default='train')
parser.add_argument('--batch_size', type=int, default=32)
parser.add_argument('--lr', type=int, default=2e-4)
parser.add_argument('--num_epoch', type=int, default=200)
parser.add_argument('--num_workers', type=int, default=15)
parser.add_argument('--pin_memory', default=False, action='store_true')
parser.add_argument('--load_pretrained_G', default=False, action='store_true',
                    help='if load pretrained model(unet) parameters')
parser.add_argument('--lr_decay', default=False, action='store_true', help='if using learning rate decay')
parser.add_argument('--test_freq', type=int, default=10, help='test after how many epochs')

# sequence length of input & output
parser.add_argument('--n_seed', type=int, default=200, help='the number of images input to encoder')
parser.add_argument('--n_position', type=int, default=50, help='the number of images going to be predicted')

''' checkpoints '''
parser.add_argument('--pretrained_G_path', type=str, default='checkpoints',
                    help='pretrained model load path')
parser.add_argument('--ckpt_dir', type=str, default='checkpoints',
                    help='model save directory')
parser.add_argument('--ckpt_last_freq', type=int, default=5000,
                    help='how many steps of saving model')
parser.add_argument('--ckpt_epoch_freq', type=int, default=2,
                    help='how many epoch of saving model')
parser.add_argument('--test_speed', default=False, action='store_true')

''' tensorboard '''
parser.add_argument('--use_tensorboard', default=False, action='store_true')
parser.add_argument('--tb_write_freq', type=str, default=1, help='how many epochs write to tensorboard')
parser.add_argument('--log_dir', type=str, default='runs', help='tensorboard log save directory')

''' visualization '''
parser.add_argument('--output_jpg', default=False, action='store_true')
parser.add_argument('--jpg_freq', type=int, default=20, help='interval of result visualization')
parser.add_argument('--jpg_dir', type=str, default='output/images', help='jpg file save directory')
parser.add_argument('--output_video', default=False, action='store_true')
parser.add_argument('--video_freq', type=int, default=100, help='interval of saving video result')
parser.add_argument('--video_dir', type=str, default='output/videos', help='video file save directory')

```

3. MORE RESULTS

The first row shows the desired lip movement we want different identities to animate, the next few rows show the animated results generated by our method.

