Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

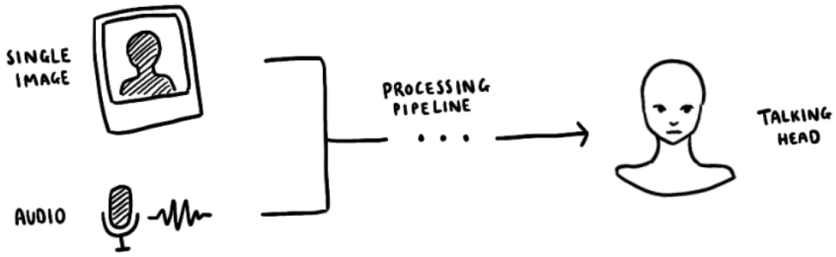
CENTER FOR PROJECT-BASED LEARNING

DEPARTMENT OF
INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Autumn Semester 2022

Towards Real-Time Audio-Driven Emotional Talking Heads

Master Thesis



Tianhong Gan
tiagan@student.ethz.ch

March 2023

Advisors: Dr. Michele Magno, michele.magno@pbl.ee.ethz.ch
Dr. Christian Vogt, christian.vogt@pbl.ee.ethz.ch

Professor: Prof. Dr. Sebastian Kozerke, kozerke@biomed.ee.ethz.ch

Acknowledgments

A large thank you to my advisors for offering valuable academic advice, and to my colleagues and friends, who have allowed this thesis to be possible, offering feedback and support at every step.

Declaration of Originality

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor. For a detailed version of the declaration of originality, please refer to Appendix A

Tianhong Gan,
Zurich, March 2023

Abstract

The synthesis of audio-driven talking-head animations have a large variety of applications, including animated films, gaming and digital humans. State-of-the-art works have made significant improvements in crucial features of the talking-head model (including input simplicity, real-time capability, and emotional portrayal), however none have achieved a model that displays superiority in all features. Additionally, select features also show room for further progress. This thesis drives audio-driven talking head synthesis towards real-time, emotional models with simple input modality. First, features from state-of-the-art architectures are combined to obtain initially a pipeline that synthesises high-fidelity, audio-driven 3D talking heads from a single in-the-wild 2D image, which has previously only been achievable through non-trivial methods of generating or crafting a compatible 3D mesh template for the corresponding 3D pipeline. Additionally, this thesis optimises the talking-head synthesis pipeline (based on VOCA) towards real-time capabilities, with a 5x improvement in rendering time and a 1.6x improvement in audio processing time over the VOCA model, obtaining a possible frame-rate of 14 FPS (from previously 3 FPS) on a Nvidia RTX 2080. Finally, this thesis contributes a small 3D emotion-based talking-head dataset, 3D MEAD, generated from a subset of the recently released (2D) MEAD dataset, and proposes a verified working method for obtaining a potentially much larger 3D emotion-based talking-head dataset given sufficient computing power and time. This provides a milestone for the capability to further improve 3D talking heads.

Contents

Chapter 1

Introduction

Synthesising audio-driven talking-head animations present an important challenge that may positively impact a variety of applications, including film making, animation, gaming, and video streaming. However, using only audio to synthesize realistic, high-fidelity talking-head animations of a target subject, especially in real time, is no trivial task. While state-of-the-art works have individually improved select features (including audio-lip sync, emotional portrayal, real-time capabilities and input simplicity), none have created a pipeline that fulfils all desired properties.

This thesis aims to compose and optimise a pipeline improving all desired properties, driven towards real-time, audio-driven, emotional talking-head synthesis from a single identity input.

1.1. Motivation

State-of-the-art works have individually improved different aspects of audio-driven talking-head synthesis, including emotional portrayal, input simplicity and real-time capabilities. However, no (known) existing work completely satisfies all desired features. Additionally, satisfied features often have room for further improvement.

While huge advancements have been achieved in audio-lip sync accuracy, the ingraining of emotion into facial features is still mostly lacking. State-of-the-art audio-driven emotional talking-heads have managed to achieve results on generating a single emotion of a specified intensity on a per-phrase basis. However, they have still failed to portray believable variance in emotion in longer audio-driven inputs with emotional variation, which would be required in most circumstances.

Additionally, most existing pipelines focus on 2D-2D (2D image/video to 2D talking head) or 3D-3D (3D mesh to 3D talking head) architectures. However, 2D-2D works that rely on a single 2D identity image often achieve less realistic outputs, displaying

unnatural artifacts or movements, which is not optimal when desired for professional use applications. To achieve high-fidelity, believable results in 2D-2D works, a long input video of the target subject is often required, followed by subject-specific re-training. This is tedious, since large amounts of data are required and retraining must be performed for each new target subject (where in applications from animation to film-making, there will be numerous subjects).

Conversely, 3D-3D works generally have no issues with resolution or unnatural artifacts when presented with a single 3D identity mesh. However, the input requirement of a neutral 3D mesh of the target head in order to synthesize an animation is often neither easily nor directly obtainable.

Finally, while good results have been achieved in real time for re-enactment-based (i.e. audio + video with facial tracking) talking-head synthesis, non re-enactment based real-time works driven by audio alone are scarce. The existing architectures are also lacking in quality, in terms of realism (facial movements, audio-lip sync) and expressive capability.

1.2. Objective

The objective of this thesis is to combine (from existing state-of-the-art works) and optimise an initial pipeline that can synthesise, from a single 2D identity image, a talking head in 3D graphics that is animated in near-real-time via audio-driven inputs alone. The talking-head synthesis model architecture will then be modified, incorporating emotion-based inputs, and thus generating more realistic facial movements and expressions.

1.2.1. Research Questions

1. Which combination of state-of-the-art architectures can be combined to synthesise an audio-driven, high-fidelity 3D talking head from a single 2D identity image?
2. How can the initial pipeline be optimized to synthesise talking-heads in real time?
3. How can emotion-based features be compactly extracted from audio inputs, to generate a talking-head exhibiting realistic facial movements and emotions?

1.2.2. Outline

The thesis will be split into three parts. In the first part, a pipeline of existing works will be combined to allow an audio-driven 3D talking head to be generated from a single in-the-wild 2D image. In the second part, the initial pipeline will be optimised to obtain near real-time capabilities. Finally, the third part will incorporate natural emotion with varying degrees, by extracting a new set of emotion-correlated features from the audio input.

The work will be evaluated using qualitative methods, including checks such as visual inspections and whether the work can be generalised across subjects. Additionally, quantitative checks, such as reconstruction error and velocity error.

Chapter 2

Related Work

The synthesization of audio-driven talking heads is currently a widely active topic, with literature existing predominantly in two groups: 2D to 2D (2D image/video target identity to 2D talking head), and 3D to 3D (3D neutral mesh identity to 3D talking head). Various works and architectures present progress in several features, including audio-lip sync accuracy, emotional expression and real-time capabilities. This section will provide a background insight into the advantages and drawbacks of the existing architectures, which have influenced the development of this project.

2.1. 2D to 2D

Most audio-driven talking-head literature falls under the 2D-2D category, where a 2D image [1–4] and/or video [5–11] of the target person is taken in order to generate a 2D synthesized talking-head video (from audio input). The most notable 2D-2D works are summarised below.

	D-L [%]	D-V [%]	D-Rot [%]	D-Pos [%]
MakeItTalk	4.6	0.9	6.1	10.1
Live Speech Portraits	3.6	0.8	3.6	8.9

Table 2.1.: Quantitative evaluation of 2D audio-driven works. D-L represents the normalized Euclidean position difference between predicted and actual points, and D-V represents the normalized velocity difference. D-Rot represents the rotation angle differences, and D-Pos represents normalized translation distances [5].

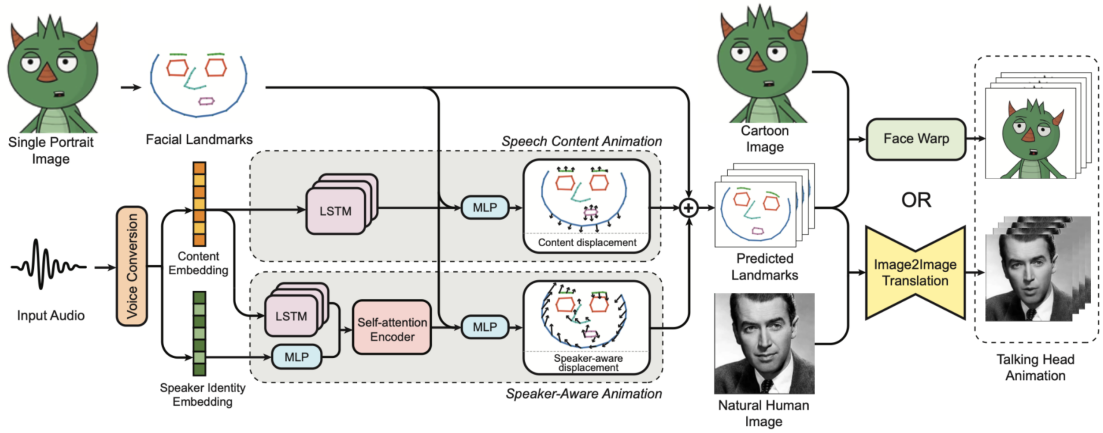


Figure 2.1.: MakeItTalk Model Architecture: From the input audio, extract the content embedding (captures the phonetic and prosodic audio-lip sync information) and the speaker identity embedding (determines expression and head movements). The content embedding is passed through LSTM (captures temporal dependencies) and MLP layers to predict the content displacement (audio-lip sync + nearby facial movements). Both embeddings are passed through LSTM and MLP layers into a self-attention encoder, and a final MLP layer to generate the speaker aware displacement (facial expression and the head motion dynamics). The generated displacements and facial landmarks extracted from an input identity image are summed to generate predicted landmarks, which are used to synthesise the final 2D animation. This is done using an image to image translation technique for realistic input images and a face warp algorithm based on Delaunay triangulation for non-realistic photo inputs. [1]

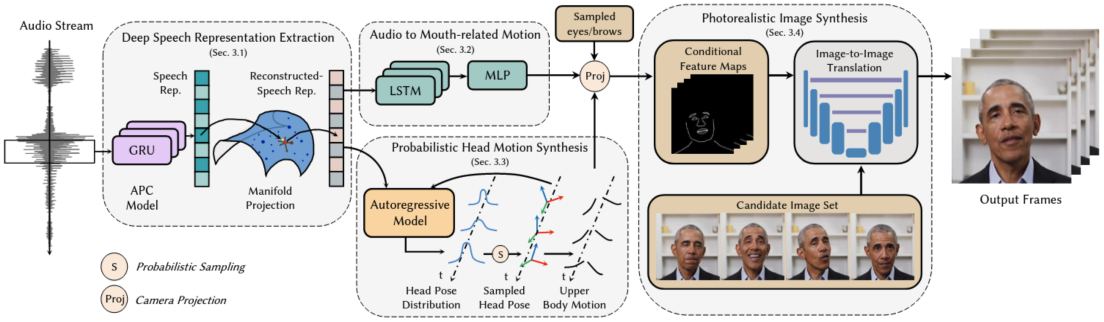


Figure 2.2.: Live Speech Portraits Model Architecture: From the input audio, deep speech representations are extracted using an autoregressive predictive coding model (extract structural speech representations) and manifold projection (improve generalisation across subjects). The extracted representation is passed through LSTM and MLP layers to predict mouth-related motion, and passed through an autoregressive probabilistic model to predict head poses and upper body motion. By projecting the predicted motions, conditional feature maps can be generated, which are fed to an image-to-image translation network along with a candidate image set (extracted from an input video of the target subject) to synthesize photorealistic talking-heads. [5]

2.1.1. MakeItTalk

The MakeItTalk [1] model architecture can be observed in Figure 2.1. MakeItTalk is audio-driven using a single identity image of the target subject, and generalises well to (both realistic and cartoon) new targets without retraining. Whilst this work achieves considerable results using a single identity image, the output resolution is limited to 256 x 256 pixels, and additionally does not capture emotion-related facial movements nor offer real-time capabilities.

The output animation of MakeItTalk achieves a reasonable imitation of real talking heads, however the results are unrealistic, and can easily be distinguished from a natural video (for natural human image inputs). A quantitative evaluation of the work can be seen in Table 2.1.

2.1.2. Live Speech Portraits

The model architecture of Live Speech Portraits [5] can be observed in Figure 2.2. Live Speech Portraits is capable of synthesizing audio-driven 2D talking heads in real-time and at an increased resolution of 512 x 512. However, it requires a 3-5 minute video of the target subject, and retraining is needed for each new target.

The synthesized talking head, again, achieves a reasonable imitation of realistic talking head movements. However, it is often prone to unnatural artefacts where the input

	Lip Vertex Error [mm]	Favourability (Full-Face/Lip-Sync/Upper-Face) [%]
VOCA	3.720	35.2/32.9/44.4
MeshTalk	3.184	64.8/67.1/55.6

Table 2.2.: Quantitative evaluation of 3D audio-driven works. Favourability refers to the percentage, out of 400 pairs of side-by-side clips, chosen to be of better visual quality between VOCA and MeshTalk [12].

video has facial landmarks inaccurately calibrated (i.e. due to partial occlusion), and additionally unnatural movements (for example excessive blinking). A quantitative evaluation and comparison in relation to MakeItTalk can be found in Table 2.1, where Live Speech Portraits achieves marginally better results.

2.2. 3D to 3D

Several works (although less populated compared to 2D-2D) additionally falls under the 3D-3D category. This uses a neutral expression 3D mesh template in order to synthesize a 3D talking-head animation from audio input [12–16]. Some relevant works are discussed below.

2.2.1. VOCA

VOCA takes in a neutral 3D mesh template of the target subject in the openly accessible Faces Learned with an Articulated Model and Expressions (FLAME) [18] topology (see Section 3.1.1), and synthesizes a 3D talking head driven by audio. The output generalises to new targets without retraining and achieves high-fidelity results when it comes to audio-lip sync accuracy. However, the model does not demonstrate any emotive capabilities, and also does not achieve real-time. The model architecture of VOCA can be observed in Figure 2.3.

VOCA captures and releases openly their full 4D face training dataset, VOCASET, which contains approximately 29 minutes of 4D scans captured at 60 FPS with synchronised audio from 12 speakers. VOCA is trained on VOCASET, using the training set (8 subjects, 320 sentences), validation set (2 subjects, 40 sentences), and test set (2 subjects, 40 sentences) with no overlap for subjects or sentences.

Visually, the animated synthesized mesh sequence of VOCA achieves quite remarkable audio-lip sync accuracy, with no unnatural artifacts. However, the resulting talking-heads are consistently neutrally posed, having no emotional expression.

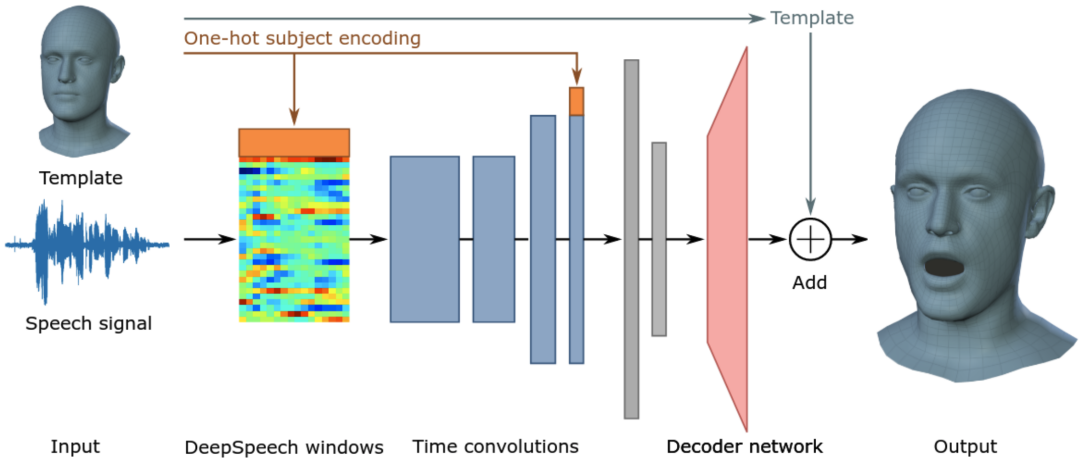


Figure 2.3.: VOCA Model Architecture: From an input audio, DeepSpeech [17] is used to extract the unnormalised log probabilities of each (alphabet) character for 0.02 second frames of the audio input. The extracted frames are resampled (from 50 FPS) to 60 FPS, and modified to form overlapping windows containing adjacent frames (to incorporate temporal information). An encoder (formed of 4 convolutaional layers and 2 fully connected layers) is trained to transform these audio features to a low-dimensional embedding. The audio features and the final convolutional layer are conditioned on subject labels, allowing subject-specific styles to be learned when training across multiple subjects. The decoder (a fully connected layer with linear activation function) maps the low-dimensional embedding into a high-dimensional space of vertex displacements, outputting a $V \times 3$ ($V = 5023$) dimensional array of vertex displacements from the template mesh, which are concatenated to form the final output animation. [13]

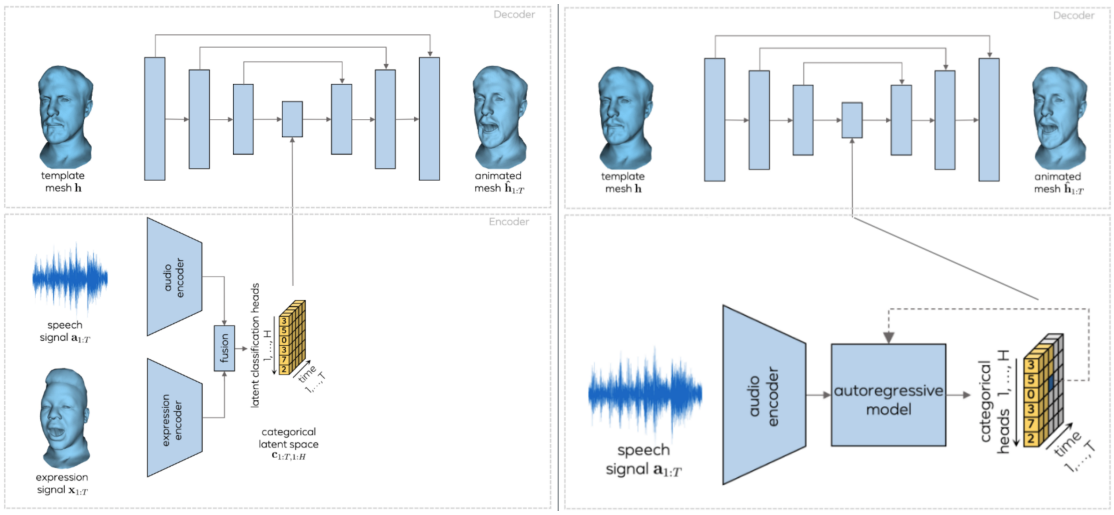


Figure 2.4.: MeshTalk Model Architecture: An expression signal (encoded via 3 fully connected layers, followed by a single LSTM layer to capture temporal dependencies) and audio input (encoded via a 4-layer 1D convolutional network) are fused using a 3-layer MLP, and mapped to a categorical latent expression space. This is passed to a UNet-style decoder that animates an input neutral 3D mesh template according to the encoded expressions (left). Where an expression signal is not present, the encoded audio input is used to learn an autoregressive temporal model (made up of 4 convolutional layers with increasing dilation along the temporal axis) over the categorical latent space. Audio-conditioned categorical expression codes are then sequentially sampled for each position in the latent expression space, and finally passed to the UNet-style decoder (right). [12]

	Single Identity Input	Real-Time	Emotion
MakeItTalk	yes (2D image)	no	no
Live Speech Portraits	no	yes	no
VOCA	yes (3D mesh)	no	no
MeshTalk	yes (3D mesh*)	near	no

Table 2.3.: Capability of related audio-driven works. Here, single identity input refers to a single 2D image or 3D mesh (* = 3D mesh not easily obtainable).

2.2.2. MeshTalk

The model architecture of MeshTalk [12] can be observed in Figure 2.4. Predominantly, the work is based of re-enactment of a given expression sequence with an independent audio sequence (which is irrelevant with regard to this thesis, as audio-driven works are the focus). However, in the case where an expression signal is absent, a separate pipeline is provided to predict the corresponding expressions from audio input alone, which is of interest.

In the latter pipeline, MeshTalk achieves similar functionality to VOCA, with higher audio-lip sync accuracy, and achieves near real-time capabilities on higher-end setups. However, the input neutral 3D mesh template (required for each target) uses a topology that is not easily generated for new targets, and thus only provided mesh templates can be used. This means it is difficult to synthesize talking head animations of new targets, without a provided template mesh. Additionally, only a tiny subset of the training dataset (13 out of 250 subjects) is released, which means the achieved results are unlikely reproducible. The model also does not have emotive capabilities.

Visually, MeshTalk achieves a result akin to VOCA, where a quantitative comparison of the works can be seen in Table 2.2. MeshTalk achieves a marginally higher audio-lip sync accuracy, and additionally higher favourability in the observed user study.

2.3. Comparison + Discussion

Comparing 2D-2D pipelines to 3D-3D pipelines, it can be seen that 2D works are generally limited to resolutions of under 512 x 512 pixels, whilst 3D models can easily be adapted to higher resolutions (by modifying the associated renderer). 2D models are also often prone to unnatural artifacts, where 3D works are generally more resilient. Conversely, the target identity input for 3D works (neutral 3D mesh) is significantly more complex to generate for new subjects, where for 2D works, a simple image or video can be used.

Table 2.3 summarises all discussed works and their capabilities. Whilst several different audio-driven works with their respective architectures have each made notable

progress in different directions, there is no optimal work that displays capability in all of single image, emotional, and real-time applications.

Additionally, it can be noted that works focusing on emotion-based talking heads, which focus on correctly capturing facial expression for different emotive sentences, is currently an area of interest with much room for improvement. This is in large part due to the absence of a suitable training dataset, especially in the 3D domain. Existing works are generally only suitable for re-enactment/facial-tracking based applications [9, 19, 20] (lip movements driven by the audio input, however facial expression driven by a separate reference video input), or otherwise only allow a small selection of emotion at constant intensity per phrase [15].

Finally, works featuring real-time talking heads are also often based on re-enactment [20–22], or otherwise require retraining per new target using a long input video, as seen in Live Speech Portraits [5].

Chapter 3

Background

3.1. 2D Image to 3D Talking Head

This section discusses the relevant background information required for the combination of the initial pipeline, of synthesizing an audio-driven 3D talking head from a single 2D identity input image.

3.1.1. FLAME

FLAME [18] is an openly accessible 3D statistical head model, learned from over 33,000 accurately aligned 3D head scans to enable the fitting and customization of identity and expression dependent shape variation in animatable 3D heads models with ease. The registration pipeline can be seen in Figure 3.1.

The topology of the FLAME model is refined from the full-body SMPL [23] template, and adjusted to contain holes for the mouth and eyes. The final mesh model consists of $N = 5023$ vertices and $K = 4$ joints, and is configurable via a function $M(\vec{\beta},\vec{\theta},\vec{\psi}) : \mathbb{R}^{|\vec{\beta}|\times|\vec{\theta}|\times|\vec{\psi}|} \rightarrow \mathbb{R}^{3N}$ that takes coefficients describing three parameters:

- Shape $\vec{\beta} \in \mathbb{R}^{|\vec{\beta}|}$,
- Pose $\vec{\theta} \in \mathbb{R}^{|\vec{\theta}|}$,
- Expression $\vec{\psi} \in \mathbb{R}^{|\vec{\psi}|}$,

and returns N vertices of the output mesh (with additional translation and rotation parameters describing the position of the mesh).

The model function is a standard skinning function $W(\bar{\mathbf{T}},\mathbf{J},\vec{\theta},\mathcal{W})$, used to rotate the vertices of $\bar{\mathbf{T}} \in \mathbb{R}^{3N}$ (template mesh in zero pose) about the joints $\mathbf{J} \in \mathbb{R}^{3K}$, and linearly smoothed by the $\mathcal{W} \in \mathbb{R}^{K\times N}$ blendweights. It is defined more formally by Equation 3.1:

$$M(\vec{\beta},\vec{\theta},\vec{\psi}) = W(T_{\mathrm{P}}(\vec{\beta},\vec{\theta},\vec{\psi}),\mathbf{J}(\vec{\beta}),\vec{\theta},\mathcal{W}), \tag{3.1}$$

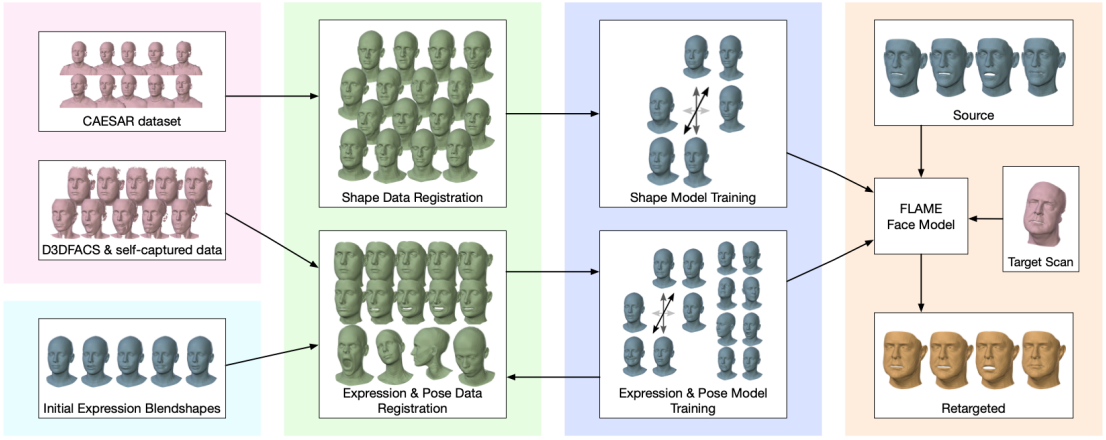


Figure 3.1.: FLAME Registration Pipeline [18]: The CAESAR dataset is used for shape data registration, and the D3DFACs dataset along with self captured data and initial expression blendshapes are used for expression and pose data registration. The registration pipeline registers meshes while regularizing to a FLAME model, and trains a FLAME model using the registrations.

where T_P is given by Equation 3.2:

$$T_P(\vec{\beta},\vec{\theta},\vec{\psi}) = \bar{\mathbf{T}} + B_S(\vec{\beta};\mathcal{S}) + B_P(\vec{\theta};\mathcal{P}) + B_E(\vec{\psi};\mathcal{E}) \tag{3.2}$$

Here,

- $B_S(\vec{\beta};\mathcal{S}) : \mathbb{R}^{|\vec{\beta}|} \rightarrow \mathbb{R}^{3N}$ is a shape blendshape function, accounting for identity related shape variation,
- $B_P(\vec{\theta};\mathcal{P}) : \mathbb{R}^{|\vec{\theta}|} \rightarrow \mathbb{R}^{3N}$ is a corrective pose blendshape function, for correcting pose deformations,
- $B_E(\vec{\psi};\mathcal{E}) : \mathbb{R}^{|\vec{\psi}|} \rightarrow \mathbb{R}^{3N}$ is an expression blendshape function, to capture facial expressions.

FLAME achieves results that are more expressive than the FaceWarehouse model [24] and the Basel Face Model [25], with mesh-to-scan fittings that are more accurate, as observed in Figure 3.2.

3.1.2. DECA

DECA [21] is a monocular 3D face reconstruction model, which can be used to reconstruct detailed 3D faces from single in-the-wild images in real time (approximately 120 FPS on a Nvidia Quadro RTX 5000). The corresponding model architecture can be seen in Figure 3.3.

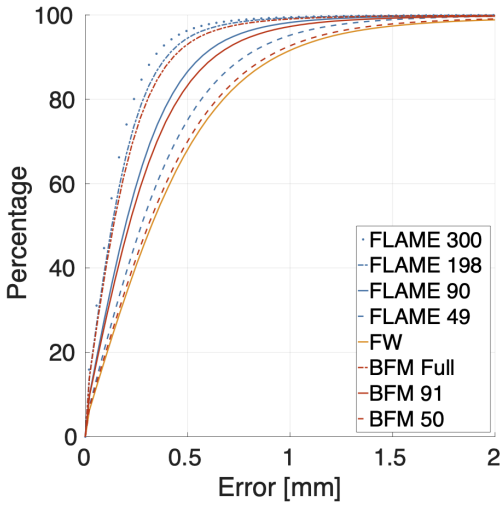


Figure 3.2.: Cumulative scan-to-mesh distance computed over all model-fits of the neutral BU-3DFE scans dataset. [18] Here, FLAME X refers to the a FLAME model with X shape parameters, FW is the FaceWarehouse Model [24], and BFM is the Basel Face Model [25].

From a single 2D image input, 2D and 3D facial landmarks are extracted. DECA reconstructs and outputs a coarse 3D reconstruction in FLAME's (see Section 3.1.1) model space, along with it's corresponding texture (albedo map). This coarse reconstructions follows FLAME topology, with $N = 5023$ vertices.

The coarse reconstruction is then augmented with a detailed displacement map to additionally obtain a detailed 3D reconstruction output. The detailed reconstruction follows a different custom topology with over 50,000 vertices, and is made animatable via the disentanglement of person-specific details (including moles, pores and expression-independent wrinkles) and expression dependent wrinkles. An overview of DECA's input-output pipeline can be seen in Figure 3.4.

DECA is trained on a total of approximately 2 million images (from datasets including VGGFace2, BUPT-Balancedface and VoxCeleb2) and achieves state-of-the-art shape reconstruction accuracy on two benchmarks, as shown in Tables 3.1 and 3.2.

3.2. Real-Time Optimisation

This section discusses the relevant background information required for the real-time optimisation of the initial pipeline.

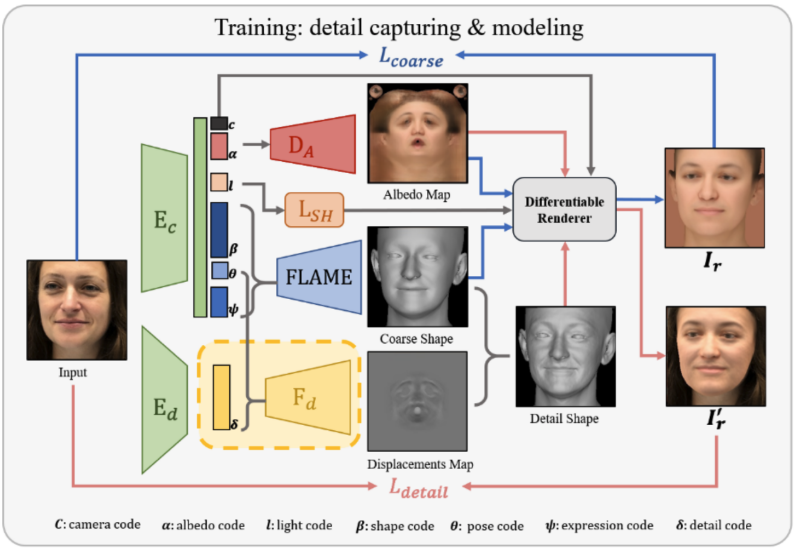


Figure 3.3.: DECA Model Architecture: DECA takes as input a 2D image and trains an encoder E_c, which consists of a ResNet50 [26] network and a fully connected layer, used to regress a low-dimensional latent code containing parameters $\mathcal{C}$, α, $\boldsymbol{l}$, and FLAME parameters β, θ and ψ as defined in the figure. The latent code is decoded to obtain the coarse shape and albedo map used to synthesise a 2D image I_r. The difference between the synthesized image and the input image is minimised. A second encoder, E_d, with the same architecture as E_c, encodes the input image into a 128-dimensional latent code containing the subject-specific details. This is concatenated to the FLAME parameters to obtain a new latent code, which is decoded to obtain a detailed UV displacement map, used to augment the coarse FLAME parameters to obtain a detailed reconstruction. L_{detail} is optimized to disentangle person-specific details and expression-dependent wrinkles, to reconstruct faces with mid-frequency details. [21]

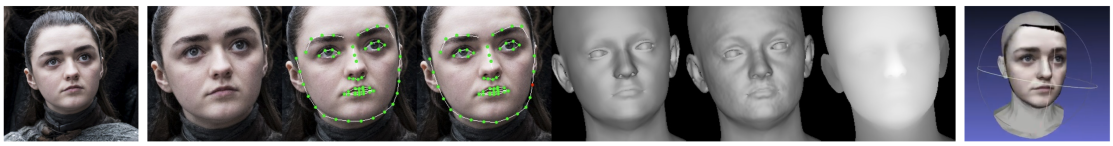


Figure 3.4.: DECA Input/Output

Method	Median (mm)	Mean (mm)	Std (mm)
3DMM-CNN	1.84	2.33	2.05
PRNet	1.50	1.98	1.88
Deng et al.19	1.23	1.54	1.29
RingNet	1.21	1.54	1.31
3DDFA-V2	1.23	1.57	1.39
MGCNet	1.31	1.87	2.63
DECA	**1.09**	**1.38**	**1.18**

Table 3.1.: Reconstruction error (distances between all vertices in the reference scan and the closest points on the reconstructed mesh) on the NoW [27] benchmark. [21]

Method	Median (mm)		Mean (mm)		Std (mm)	
	LQ	HQ	LQ	HQ	LQ	HQ
3DMM-CNN	1.88	1.85	2.32	2.29	1.89	1.88
Extreme3D	2.40	2.37	3.49	3.58	6.15	6.75
PRNet	1.79	1.59	2.38	2.06	2.19	1.79
RingNet	1.63	1.59	2.08	2.02	1.79	1.69
3DDFA-V2	1.62	1.49	2.10	1.91	1.87	**1.64**
DECA	**1.48**	**1.45**	**1.91**	**1.89**	**1.66**	1.68

Table 3.2.: Reconstruction error (distances between all vertices in the reference scan and the closest points on the reconstructed mesh) on the Feng et al. [28] benchmark. LQ refers to 1344 low quality images extracted from videos, and HQ refers to 656 high-quality images taken in controlled scenarios. [21]

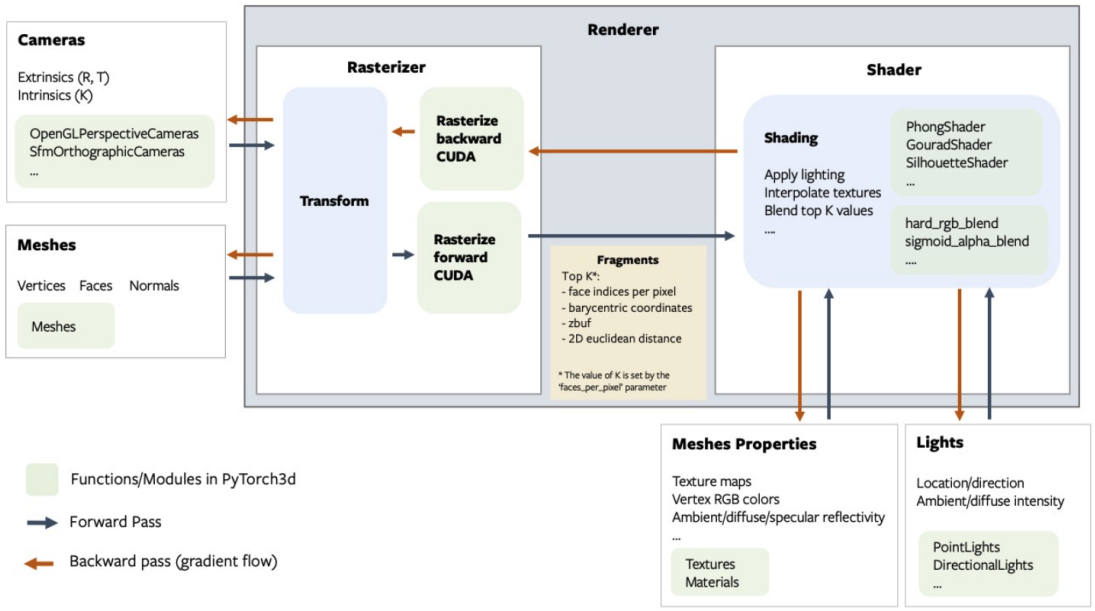


Figure 3.5.: PyTorch3D Differentiable Renderer [29]

3.2.1. PyTorch3D Renderer

PyTorch3D is a library built on top of PyTorch specifically for deep learning with 3D data. PyTorch3D's modular differentiable rendering API (see Figure 3.5), allows the design and implementation of an efficient renderer that works at near real-time (with render speeds of under 150 ms for resolutions of 250 x 250 pixels) for most uncomplicated scenes [29].

3.3. Emotion Integration

This section discusses the relevant background information required for the integration of emotion into the final audio-driven talking-head synthesis pipeline.

3.3.1. HuBERT

Hidden-Unit Bidirectional Encoder Representations from Transformers (HuBERT) [30] is an approach for self-supervised speech representation learning for natural language processing. It takes in a raw audio waveform and outputs a corresponding speech representation matrix. The speech representation matrix is a compact numeric representation of the audio capturing the combined linguistic (content of speech, i.e. text) and acoustic (emotion, accent, age) models.

The information from the raw output matrix can be further processed to extract the

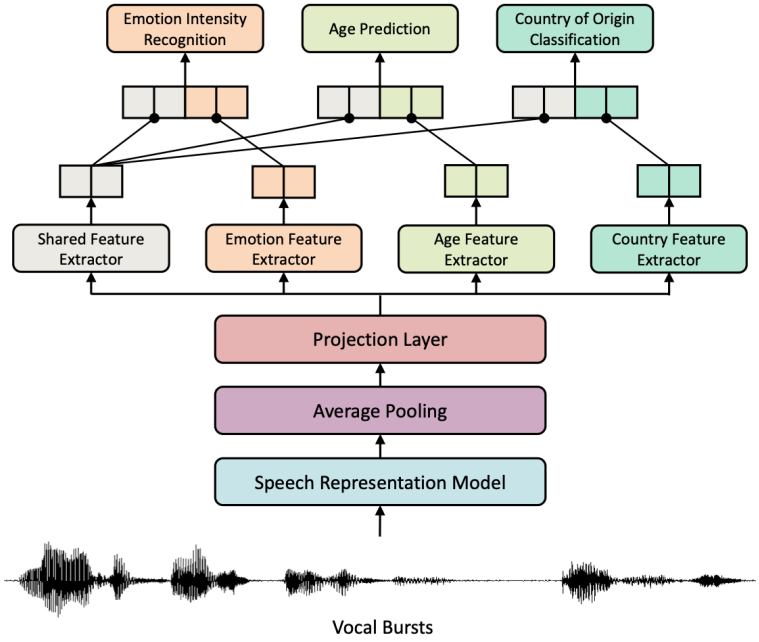


Figure 3.6.: Burst2Vec Model Architecture: Burst2Vec uses HuBERT to extract the speech representation model (to compute temporal acoustic representations) from input vocal bursts (audio input). The speech representation model is passed through an average pooling layer (to obtain vector representations of the audio input) and then through a projection layer (which projects the vector to a lower dimensional space). Feature extractors process the projected representations to obtain task-specific (emotion, age or country) and shared (relevant to all) features. During inference, shared and task-specific representations are concatenated and passed through a projection layer to form a combined representations for each output head. [31]

emotional label and intensity (using Burst2Vec [31] - see Figure 3.6), which achieves a 30% gain in performance compared to baselines using pre-extracted features, and scored highest in the ICML ExVo 2022 Multi-Task Challenge [31].

3.3.2. MEAD

Multi-view Emotional Audio-visual Dataset (MEAD) [32] is a 2D talking face corpus featuring 8 different emotions (neutral, anger, contempt, disgust, fear, happiness, sadness and surprise) at 3 intensity levels, and captured at 7 different view angles in a strictly controlled environment (see Figure 3.7). The corpus yields approximately 40 hours of audio-visual data for each viewing angle.

The captured sentences are selected to be emotionally consistent and cover a variety of

Figure 3.7.: MEAD Talking Head Corpus [32]

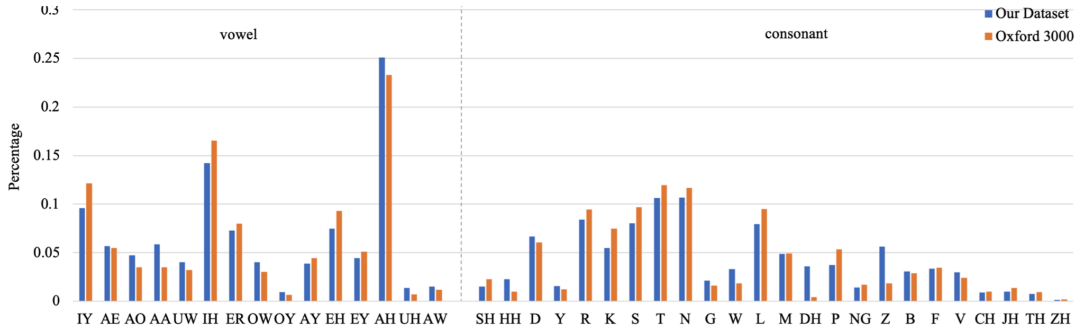


Figure 3.8.: Vowel and consonant distribution (MEAD in blue, Oxford frequently used 3000 words in orange) [32]

phonemes (with a range similar to the 3000 most frequently used words - see Figure 3.8). For each emotion, there are 3 common sentences, 7 emotion related sentences and 20 generic sentences.

The dataset records 60 fluent English speakers, aged 20 to 35 and with previous acting experience, guided to portray natural and accurate emotions by a team led by a professional actor.

MEAD is the largest emotional talking head dataset (to the extent of the knowledge at the time during which this thesis is conducted), containing a wide variety of emotions and sentences, enabling progress in synthesizing 3D emotional talking heads that previously was not possible before the creation of the dataset.

Chapter 4

Method + Implementation

4.1. 2D Image to 3D Talking-Head

The first thesis objective is to select an optimal combination of state-of-the-art architectures that may be combined, in order to synthesise a high-fidelity, audio-driven 3D talking head from a single in-the-wild 2D image. This section explains how the initial pipeline is selected and formed, along with the associated design choices.

4.1.1. 2D Image to 3D Mesh

To synthesize a 3D talking head from a 2D image input, the first step is to extract facial landmarks from the input 2D image, in order to generate a corresponding 3D head mesh of the target subject. To this effect, DECA (see Section 3.1.2), a state-of-the-art 3D face reconstruction model, is used. DECA generates a coarse 3D head model in FLAME topology (see Section 3.1.1) corresponding to a target subject in a given in-the-wild 2D image input. The corresponding mesh texture is also extracted and saved. The resulting 3D head mesh includes all default FLAME parameters (translation, rotation, pose, shape and expression) in a non-zero state, saved as an OBJ file.

DECA is selected for this section of the pipeline, as it obtains state-of-the-art results (see Tables 3.1 and 3.2), with higher accuracy compared to other similar works. Additionally, DECA returns a coarse mesh in the commonly used and accessible FLAME topology, which proves to enable 3D mesh topology alignment to become much more feasible.

4.1.2. 3D Mesh to 3D Talking Head

To synthesize an audio-driven 3D talking head from an input 3D head mesh, VOCA (see Section 2.2.1) is used. VOCA takes in an input audio file, which is used to drive a neutral 3D head mesh in FLAME topology (with only the shape parameter set to non-zero values), to finally synthesize a 3D talking-head sequence corresponding to the

input audio.

VOCA is selected for this section of the pipeline as it achieves high-fidelity results and accurate audio-lip sync capacity. Additionally, VOCA releases their full training dataset, VOCASET, for use in training and validation. Finally, and msot importantly, all 3D head meshes used by VOCA (including the training dataset and the neutral 3D head mesh required for input per inference) are presented in the FLAME topology, which allows for easy topology alignment.

MeshTalk (which achieves a higher audio-lip sync accuracy) was trialed for use in this section of the pipeline, however had an incompatible mesh topology, which would require non-trivial methods to achieve alignment.

4.1.3. Aligning 3D Mesh Topology

To join the two sections of the pipeline discussed above, the topology of the image-to-mesh output and the 3D talking-head synthesis mesh input must be aligned. This means that both meshes must have the same number of vertices, the same ordering of vertices, and additionally must be completely aligned in translation and rotation. This is a non-trivial process if the topology of a 3D mesh is not clearly defined, and no method of fitting a new model to the topology is provided. To this effect, DECA is used for the image-to-mesh part, and VOCA used for the 3D talking-head synthesis part, as defined above, as both works use the known and accessible FLAME mesh topology.

Even with the same topology, aligning the two meshes is non-trivial. The input of VOCA requires a 3D mesh in PLY format, which is zero-posed in all FLAME parameters aside from shape. Conversely, the 3D mesh generated by DECA is saved in OBJ format, and additionally captures all of the translation, rotation, pose, shape and expression FLAME parameters in a non-zero state (since these are captured corresponding to the non-zero parameters of the in-the-wild 2D image). To enable compatibility, several steps are taken.

First, the PyMeshLab library is used to convert the generated OBJ file to PLY format. In this step, it is important to ensure vertex order (i.e. the order in which vertices are listed in the file) is maintained constant, as this directly affects topology. An inconsistent vertex order will generate unnatural artifacts in the synthesized talking head, as the model and inferences would be trained and made on arbitrary topologies with no clear pattern.

Following this, the 3D mesh must be converted to zero pose (i.e. assigning all FLAME parameters to be zero, aside from the 'shape' parameter that captures the identity face shape of the target from the 2D image). This is done using a custom script written by combining and altering some of the existing functions in the open source FLAME library. More specifically, extracting the FLAME parameters from DECA's mesh (by fitting the

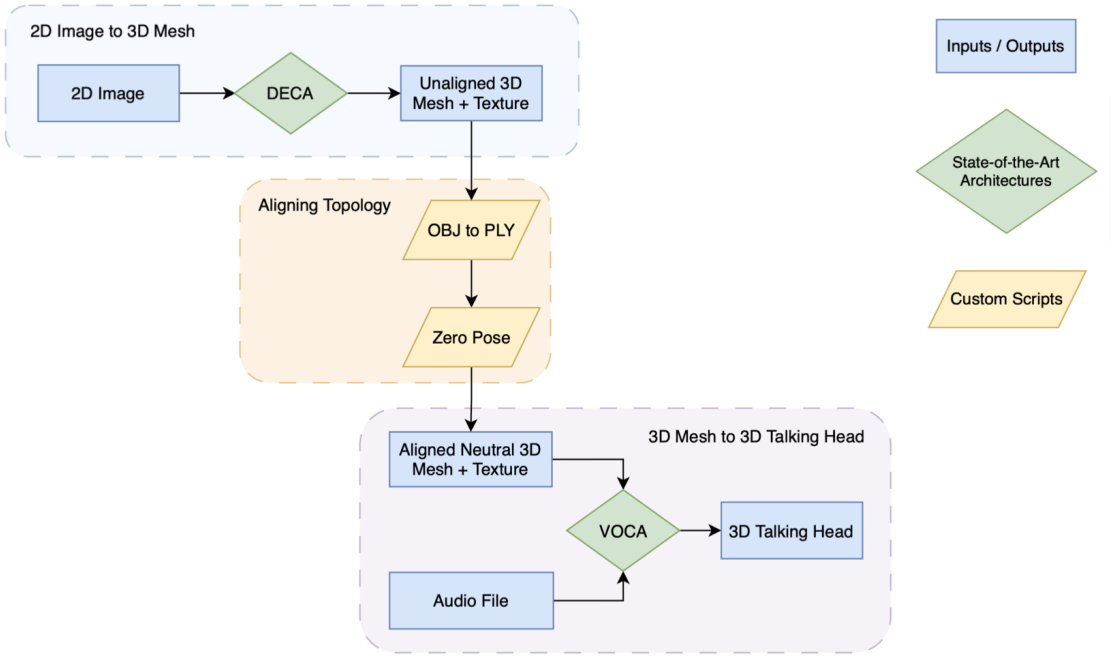


Figure 4.1.: 2D image to 3D talking head pipeline

FLAME model to the mesh) and generating a new mesh that zero-poses all parameters aside from the identity shape parameter. The script takes in and outputs meshes in PLY format.

Using these steps, the 3D mesh output of DECA is made compatible with the required input topology for VOCA.

In summary, the composed pipeline (see Figure 4.1) makes use of DECA [21] to generate a neutral 3D mesh template (in FLAME [18] topology) of the target person from a single in-the-wild image. The 3D mesh template, aligned to the correct mesh topology, and is then driven by VOCA [13] to synthesize a 3D talking-head animation from an audio input.

4.2. Real-Time Optimisation

The second objective of the thesis is to optimise the pipeline, obtained in the first thesis objective, to achieve near real-time capabilities.

The standard for real-time when it comes to visual media or animation varies largely depending on the media and target audience. For example, most movies shown on TV are shot at 24 FPS, whilst the acceptable frame rate for video games is 30 FPS. For

competitive e-sports, oftentimes frame rates of 120 FPS and even up to 240 FPS are also used. For the purposes of this thesis, a frame rate of 30 FPS will be defined as real-time, as any frame rate higher than this is not hugely noticeable or distracting for most forms of media.

4.2.1. Identifying Bottlenecks

The initial pipeline obtained in the first thesis objective can be seen in Figure 4.1. To push the pipeline towards real-time operation, the bottlenecks (sections contributing most to preventing real-time operation) must first be identified. To this effect, each stage of the pipeline, as shown in Figure 4.1, should be individually timed, and the respective obtainable FPS calculated (as the total number of frames divided by total time for the stage). For this, assume the resulting talking-head animation is generated and rendered frame-by-frame (i.e. streaming each frame as it is being rendered rather than waiting for the full sequence to generate).

4.2.2. Audio Feature Extraction

One identified bottleneck is the audio processing section of the pipeline (see results Section 5.2.1). This includes the audio feature extraction (using DeepSpeech) and additional pre-processing of the audio features (see Section 2.2.1 for more details on how audio is processed). To improve the real-time capabilities of the initial pipeline, the DeepSpeech model is replaced by the HuBERT model (see Section 3.3.1), which can extract the relevant audio features in a shorter time.

From the output speech representation matrix of the HuBERT model, the unnormalised log probabilities of each alphabet character can be obtained by extracting the logits (with dimension [1, N, 32], where N represents the number of 0.02 second frames, and 32 represents the number of characters) to match the input format of VOCA (with dimensions [N, 1, 29], where the definition of N remains the same, and 29 is the number of characters).

To replace the function of DeepSpeech with HuBERT, the output of HuBERT is first reshaped to [N, 1, 32], then the correct characters selected and ordered to form an array matching the output array of DeepSpeech, with shape [N, 1, 29].

The HuBERT model is evaluated against the DeepSpeech model in terms of real-time capabilities by timing the respective parts of the pipeline and comparing the total audio processing time achieved by each of the two models, as well as the resulting FPS attainable, calculated as the total number of frames in the processed clip divided by the time taken for processing.

Additionally, to ensure that any improvement to the real-time capabilities does not impact output fidelity of the talking-head synthesis, the reconstruction loss and velocity loss (see Section 4.4) will be compared between a model trained using the original Deep-

Speech audio feature extraction and using the new HuBERT method. The synthesised output talking-heads will also be visually compared between the two models.

4.2.3. Rendering Optimisation

From the results in Section 5.2.1, the largest bottleneck lies in the rendering time of the synthesized talking-head animation. In the initial pipeline, VOCA uses the Pyrender library to create their renderer, which, as observed in the results Section 5.2.1, attains far from real-time capabilities. Alternatively, the PyTorch3D library is used by other real-time works to attain real-time or near real-time capabilities.

To drive the rendering section towards real-time operation, a custom renderer is built using the PyTorch3D library (replacing VOCA's Pyrender renderer), which takes in and renders the sequence of meshes generated by the VOCA pipeline. The new renderer is structured as a class, defined in three parts, as described in the following paragraphs.

The main renderer function takes in, as input, the directory path of the generated object sequence, the path to the generated texture files (including the vertex texture coordinates, the associated Material Template Library (MTL) and the texture image), and the batch size (the number of objects in the sequence to render in each iteration - this is limited by GPU memory capacity). The function renders each object in the generated object sequence, and returns an array of rendered images corresponding to each object. The renderer is a Phong renderer, set with camera and lights such that the talking head is front-facing and adequately lit using a point light source emitting mainly diffuse and ambient light.

Additionally, a separate 'get texture' function (called from within the main renderer function) pre-processes the texture required by the main renderer. This function uses the texture file path taken as input in the main renderer and returns the corresponding textures defined as a TexturesUV object (a PyTorch3D object, for rendering with texture), or otherwise defines a plain white object texture in the case where texture is not defined/desired.

The final function uses the ffmpeg library in order to generate the final video animation of the rendered object sequence combined with the original driving audio.

To evaluate the new renderer, the timings achievable from the newly implemented PyTorch3D renderer will be compared to VOCA's original Pyrender renderer. This is done by assessing the total render time for a talking-head animation, along with the final attainable FPS. Additionally, the rendering results will be visually compared between the two renderers, to ensure any improvement in time does not result in any loss of output fidelity.

4.3. Emotional 3D Dataset

The VOCA model used in the custom pipeline built thus far is trained using VOCASET, which does not capture any emotional variation of facial features (as VOCA is focused on achieving high audio-lip sync accuracy, the expression parameters are set to zero-pose in the training data).

In order to retrain while taking into consideration emotion-based parameters, one must first prepare an emotional dataset. Such a suitable dataset can be defined as one which contains a variety of sentences evenly covering the phoneme domain, which portrays accurately an adequately large set of different emotions (including neutral/no emotions), and captured, preferably with a number of different speakers, in the 3D FLAME topology.

Finding a suitable emotional dataset

The input of VOCA requires 3D meshes in the FLAME topology, thus the input dataset for training must also follow this topology. This poses an initial issue, as out of the available datasets captured in FLAME topology, none capture the emotional variation and information needed to retrain an emotion-based model.

Additionally, converting a different 3D dataset into FLAME topology is a non-trivial problem, with no given method of conversion, and may pose a whole other project. Not to mention, the existence of such a optimally suitable 3D dataset does not actually exist (the slow progress of emotion-based audio-driven talking heads is largely due to the lack of a suitable dataset [32]).

In the 2D domain, there is also a deficiency in suitable datasets, with RAVDESS 2018 [33] being one of the predominantly used datasets for training emotion-based talking heads, and MEAD being one of the newer, larger emotion-based datasets. Although RAVDESS appears to contain a variety of emotions (calm, happy, sad, angry, fearful, surprise, and disgust) with 24 different speakers, it appears to speak only 2 sentences, which may hinder the audio-lip sync accuracy if used. Hence, MEAD (see Section 3.3.2), which portrays a similar variety of emotions, however with a much larger variety of sentences, is selected for use.

Converting to 3D

Whilst a suitable dataset in the 2D domain has been selected, the required form of input must be in the 3D domain, and furthermore in FLAME topology. In order to achieve this, a custom conversion pipeline is developed to create a new 3D MEAD, based on the 2D MEAD dataset. This is done by passing each frame of the the 2D video sequence through DECA, to obtain the corresponding 3D sequences in FLAME topology.

More specifically, as in the initial pipeline, DECA is used to convert each 2D frame to a 3D mesh, where the generated OBJ mesh file is converted to PLY format, and finally zero-posed (as emotional facial expressions are very relevant, the zero-pose is only applied to the translation and rotation parameters of FLAME, leaving the values of shape, expression and pose). The designed method is inspired by a similar method used in [15], where the 2D RAVDESS dataset is converted to 3D frame-by-frame using 3DMovieMaker.

Selecting a suitable subset

As the size of the MEAD dataset far surpasses that of the VOCASET dataset on which VOCA is trained, it is desirable to carefully select a subset of MEAD to process (this is also required due to the limited storage memory and limited time available for processing).

To select the optimal subset, one must first observe the dataset splits utilised by VOCASET, then select a subset of MEAD which imitates this split. This is non-trivial, since comparatively, MEAD features each sentence recorded with 8 emotions, 3 intensities, and 7 viewing angles, whereas VOCASET has each sentence recorded only once (neutral emotion, front-facing view). This is complex as it is important to have some common sentences featuring every emotion (to learn emotion variance when pronouncing the same sentence), whilst also ensure a wide variety of different sentences are utilised without too much repetition (to ensure the quality of the audio-lip sync).

VOCASET includes a training set (8 subjects, 320 sentences), validation set (2 subjects, 40 sentences), and test set (2 subjects, 40 sentences), with no overlap for subjects or sentences. In total, this contains 29 minutes of 4D scans captured at 60 FPS and synchronised audio from 12 speakers. Comparatively, the full MEAD dataset (see Section 3.3.2) features 60 subjects, with approximately 40 hours of audio-visual data recorded for each of the 7 viewing angles. A single front-facing viewing angle is selected for simplification, to match VOCASET (which also uses only a front-facing view).

According to the MEAD paper [32], for each emotion, there are 3 common sentences, 7 emotion related sentences and 20 generic sentences. However upon observation, the following sentence structure is observed.

For emotional sentences (happy, contempt, disgust, fear, happy, sad, surprised):

- 13 sentences present across all emotions (common)
- 17 unique sentences per emotion (unique)

Additionally for sentences with neutral emotion:

- 13 sentences persent across all emotions (common)

- 12 sentences with 2 sentences shared with each emotion except contempt (shared)
- 15 unique neutral sentences (neutral)

The full transcript of sentences has been transcribed and can be seen in the appendix (Section B.2).

The size of the selected subset is limited by the limited time available (during this thesis) for generating the subset. Given no subset is selected, 40 hours of audio-visual data (including all 8 emotions and 3 intensities for the 60 subjects) for the front-facing viewing angle must be converted from 2D video to 3D mesh sequences. The 2D image (frame) to 3D mesh generation step requires approximately 23 seconds per frame (see Table 5.1), although may be slightly faster when initializing DECA only once to generate each clip (rather than once per frame).

This yields a total generation time of 40*60*60 seconds of video * 60 FPS * 23 seconds to generate a single frame = up to 55200 hours (= 328 weeks) required to generate the full dataset, which is obviously infeasible. Therefore a subset of the dataset must be selected to fulfil a generation time of at most 1.5 weeks (approximately 252 hours). This yields the capability of generating at most 252*60*60 seconds / (60 FPS * 3 seconds per video sentence * 23 seconds to generate a single frame) = approximately 219 sentences (around 11 minutes of data).

After analysing the sentences present in the MEAD dataset in more detail, it can be noted that the available data is slightly disorganised, with lots of mislabels (sentences labelled incorrectly), missing clips and broken clips. Additionally, emotions are often obviously 'faked' (as the dataset is recorded with amateur actors). Thus, when selecting the chosen subset, all sequences for a subset of subjects are first manually sorted to list the corresponding errors, and sequences that feature errors are avoided when selecting the subset to generate.

The dataset subset has thus been selected and split as shown in the appendix Section B.1, where in total, 10 subjects (6 training, 2 validation, 2 test) are selected with a total of 214 sentences (150 training, 32 validation, 32 testing), of which 144 are unique (108 training, 18 validation, 18 testing). The subset is selected to match the distribution of the VOCASET dataset splits, taking into account the time constraints present for generating the dataset.

The selected subset is converted from the captured videos in 2D MEAD to sequences of 3D meshes using the method specified in the above section.

As the dataset is generated using an unconventional method (rendered using DECA frame by frame from a 2D dataset to 3D), to ensure the quality of the produced dataset, it is important to manually visually process the output dataset to ensure no unnatural artifacts or issues exist (since classical metrics such as reconstruction and velocity loss

cannot be used). All the 3D mesh sequences generated are rendered using the PyTorch3D renderer built as part of this thesis. Each rendered clip is manually viewed and checked, to observe the overall quality of the generated dataset.

Observing the 3D generated dataset, it can be seen that emotions are not very accurate, and differences between intensities are sometimes not apparent (this was also true for the 2D counterpart, however more severely so for the 3D generated dataset). Additionally, many of the clips feature jitters around the nose, especially in the emotions of fear and surprise. Due to these observations, it is decided that for every emotional sentence, only one out of three clips for the three available intensities are selected for use, in order to avoid clips with jitters and additionally eliminate confusing and ambiguous information presented by multiple emotional intensities rarely accurately portraying variation in emotional intensity.

Aligning dataset to VOCA model input

Having generated a suitable 3D dataset for training, it is important to determine the dataset input format required to train the VOCA model. The required input files for the original VOCA model are named and the definitions inferred via testing and observation as follows:

- data_verts.npy : A numpy array file containing every frame (converted 3D mesh) of every sentence of every subject in one large vertex array of dimension [N, 5023, 3], where N is the total number of frames, and 5023 is the number of vertices per mesh, with each vertex having 3 coordinates.
- init_expression_basis.npy : A numpy array file of shape [15069, 100], corresponding to the weights (initialised as 100 PCA components computed over the vertex displacements of the training data, initialised as zeroes) of the linear activation function (outputs 5023 × 3 dimensional array of vertex displacements from the input subject specific template) of the fully connected layer in VOCA's decoder.
- raw_audio_fixed.pkl : A pickle file containing the dictionary storing the raw audio for every subject and sentence, with dtype=int16 and a sample rate of 16000.
- processed_audio_deepspeech.pkl : A pickle file containing the dictionary storing the DeepSpeech processed audio (shape [N, 16, 29], where N is the number of frames, 16 is the window size for overlapping frames to incorporate temporal information, and 29 is represents each of the 26 characters of the alphabet plus special characters) for every subject and sentence.
- subj_seq_to_idx.pkl : A pickle file containing the dictionary mapping of each frame in every sentence for every subject to the corresponding index in the vertex array data_verts.npy.
- templates.pkl : A pickle file containing the dictionary of neutral 3D mesh templates of all 12 subjects (for VOCASET).

Using the generated 3D mesh sequences, the required input files can finally be generated. 'data_verts.npy' and 'subj_seq_to_idx.pkl' may be generated using some simple python functions that iterate over the generated mesh sequences. The raw and processed audio files can also be generated in a similar matter, by iterating over the input audio files, calling the audio processing function, and appending to a dictionary for every subject and sentence. The 'templates.pkl' file is generated by selected the first frame within the 2D video, and converting this into a zero-posed (including expression) 3D mesh of the target subject.

Finally, since it is unclear from VOCA's description of the 'init_expression_basis' file whether the weights provided for VOCASET are based specifically on the VOCASET dataset, or on an general set of weights for 3D talking head models (as the file itself seems to be copied and used in some other models, upon investigation), three methods of generating the file are conducted, including:

- Training on the weights provided by VOCASET (1)
- Training on an array of weights initialised to zero (2)
- Training on the array of weights saved from training on the zeroed-out weights (3)

Evaluation

An evaluation for the effectiveness of the 3D dataset generation method can be measured as a percentage of accurately converted clips, as metrics such as reconstruction loss and velocity loss cannot be measured between a 2D input and 3D output. Additionally, a sanity test evaluation is performed to check whether results obtained from retraining the original VOCA model using the newly generated 3D MEAD dataset has the potential to be adequately realistic (i.e. whether the new dataset is feasible for use). The results are evaluated against the model trained on VOCASET, using the training and validation losses (based on reconstruction error and velocity error - see Section 4.4).

4.4. Evaluation Metrics

The final pipeline can be evaluated quantitatively and qualitatively. Quantitative evaluation of the pipeline can be measured using the following metrics:

- Reconstruction error: given by Equation 4.1, is the average Euclidean distance between the coordinates of the predicted vertices and the reference ones.
- Velocity error: given by Equation 4.2, is the average Euclidean distance between every two adjacent frames of all predicted vertices and reference ones (to indicate the accuracy of dynamic vertex motion).

$$ReconstructionLoss = ||\mathbf{y}_i - \mathbf{f}_i||^2 \tag{4.1}$$

$$VelocityLoss = ||(\mathbf{y}_{i} - \mathbf{y}_{i\text{-}1}) - (\mathbf{f}_{i} - \mathbf{f}_{i\text{-}1})||^2 \tag{4.2}$$

$$Loss = ReconstructionLoss + VelocityLoss \tag{4.3}$$

Here, $\mathbf{y} \in \mathcal{R}^{N\times 3}$ are the actual vertex positions, $\mathbf{f} \in \mathcal{R}^{N\times 3}$ are the predicted vertex positions and N is the number of vertices.

Additional qualitative evaluation metrics that may be considered include:

- Generalisation across subjects: whether the results are generalisable to different subjects.
- Speaker styles: whether the results accurately portray different speaking styles.

Chapter 5

Results

5.1. 2D Image to 3D Talking Head

Following the method specified in the corresponding method section, a functional pipeline is achieved, which takes in an audio waveform and a single in-the-wild 2D image of the target subject, and synthesizes a purely audio-driven, high-fidelity 3D talking head. This has previously only been achievable through non-trivial methods of finding and combining several works to generate a compatible 3D mesh template for a corresponding 3D-3D pipeline.

Figure 4.1 shows the combination of state-of-the-art architectures used to achieve this. Figure 5.1 shows the resulting outputs for each stage of the pipeline from 2D image to neutral 3D mesh, and Figure 5.2 shows a still from the synthesized talking-head animation. As can be observed, both the generated 3D mesh, as well as the talking head synthesized using the generated mesh, are high-fidelity outputs that accurately portrays the input 2D image. The accuracies of both outputs follow the accuracies of the DECA

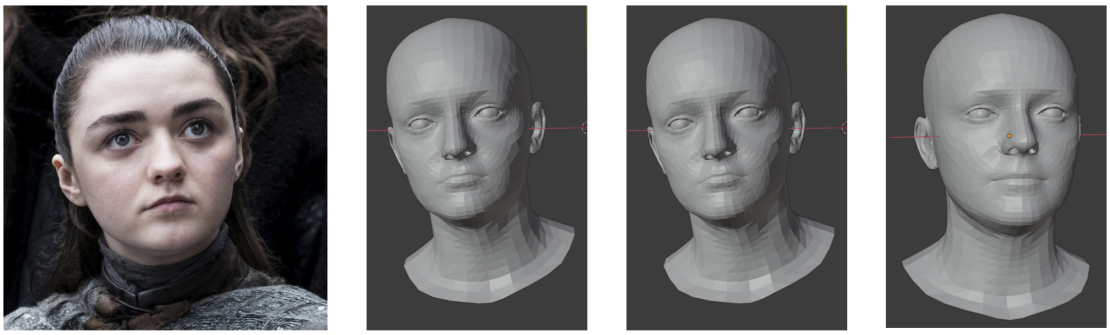


Figure 5.1.: 2D image to 3D neutral mesh.

Figure 5.2.: Frame from the synthesized 3D talking head animation, without texture (left) and with texture (right).

	2D Image to 3D Mesh (DECA)	Aligning Topology		3D Mesh to 3D Talking Head (VOCA)		
		OBJ to PLY	*Zero-Pose*	*DeepSpeech Audio Processing*	*Object Sequence Generation*	*Rendering*
Time [s]	16.04	0.03	6.83	12.15	10.54	100.44
Approx. FPS	13			25	28	3

Table 5.1.: The pipeline timings (on a Nvidia RTX 2080) for each stage of the pipeline. The attainable FPS is calculated as total number of frames (= 300) / time taken for section.

and VOCA models (described in Sections 3.1.2 and 2.2.1, respectively).

5.2. Real-Time Optimisation

In this section, the results for the method defined in the corresponding real-time optimisation Section 4.2 is documented and described.

5.2.1. Identifying Bottlenecks

The time taken and associated FPS for synthesizing a 3D talking head from a 5 second audio input for each stage of the initial pipeline can be seen in Table 5.1. From this, the bottlenecks in the pipeline can be identified and are highlighted in red.

	Audio Feature Extraction [s]	Additional Processing [s]	Total Time [s]	FPS
DeepSpeech	6.42	5.73	12.15	25
HuBERT	1.35	6.03	7.38	41

Table 5.2.: Audio processing time using the HuBERT model compared to the DeepSpeech model, for processing of a 5 second (approximately 300 frame) textured sequence. The attainable FPS is calculated as total number of frames / total render time.

It can be seen that for the image to mesh section (highlighted in grey), a total time of 22.9 s is required. To achieve a real-time talking head, the image to mesh generation part of the pipeline is not as crucial, as the 3D mesh, in most cases, would only need to be generated once, and can be reused. This does not influence the real-time capabilities of the 3D talking-head synthesis part. To this regard, as long as the mesh generation is within a reasonably short time-span, it should be acceptable. Thus, although the time achieved in the first section (image to mesh) does not reach 30 FPS, the section completes within 23 seconds, which is a reasonably short time-span, and is thus irrelevant to the real-time capability limitations.

In the VOCA (3D talking-head synthesis) stage, the resulting bottlenecks can be identified as the DeepSpeech audio processing step, and, predominantly, the rendering step, achieving frame-rates of approximately 25 FPS and 3 FPS, respectively. For the rendering step in particular, this is nowhere near real time, and must be optimised.

5.2.2. Audio Feature Extraction

As observed in Table 5.2, by replacing the DeepSpeech model with the HuBERT model for audio feature extraction, the resulting time required for audio processing improves by almost double from 12.15 s for DeepSpeech to 7.38 s for HuBERT. HuBERT extracts the required audio features in a significantly faster time, although takes slightly longer for the additional processing (including reshaping, resampling and making windows), as it must reshape it's output array into a format matching DeepSpeech, for input into VOCA. The new time allows a FPS of approximately 41 FPS to be achieved, compared to 25 FPS when using DeepSpeech, which is an improvement of over 1.6 times.

To ensure the output fidelity has not been affected by the improvement in real-time capability, the results obtained using the newly integrated HuBERT model is compared to DeepSpeech.

Visually, the quality of the results are comparable if not better than the original model (slightly more facial movements, slightly more expressive). The training and validation

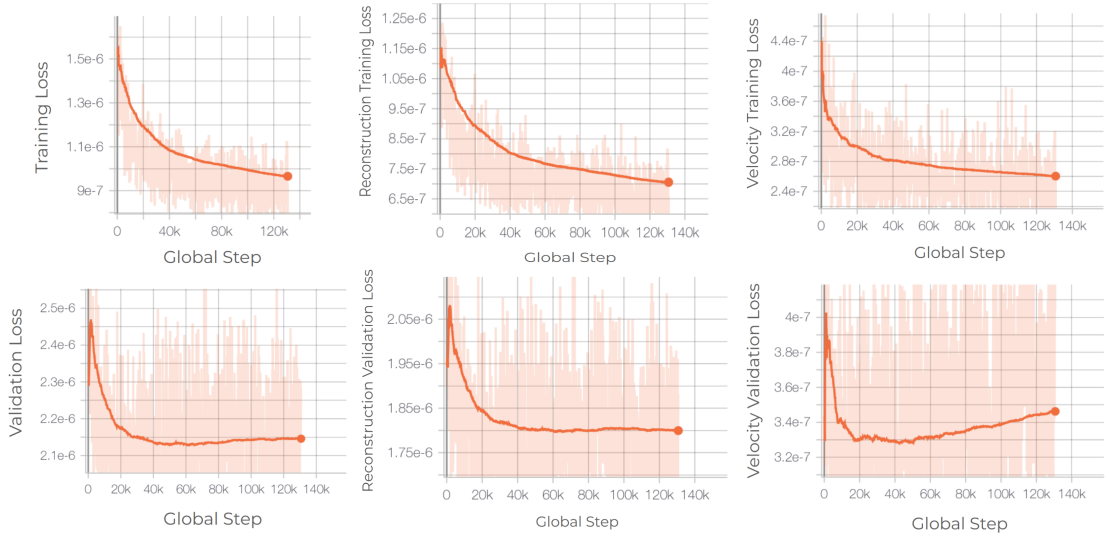


Figure 5.3.: Training and validation (velocity and reconstruction) losses for the DeepSpeech + VOCASET trained model on the VOCA architecture.

(sum of reconstruction and velocity) losses obtained using HuBERT (compared to DeepSpeech) in the VOCA model architecture can be seen in Figures 5.4 and 5.3 respectively. It can be observed that the overall validation loss reaches a lower point, however at a later point in training (approximately the 80k global step mark instead of 60k mark, reaching 2.1e-6 loss instead of 2.13e-6, for HuBERT and DeepSpeech respectively). The velocity loss still seems to show signs of over-fitting using HuBERT, levelling off at 40k and beginning to increase after 80k, however this trend seems a lot less severe compared to the original model using DeepSpeech.

Overall, it can be seen that replacing DeepSpeech with HuBERT not only enhances real-time capabilities, but the fidelity of the output results are also minimally increased.

5.2.3. Rendering Optimisation

Table 5.3 shows the resulting time comparison between the original Pyrender renderer used by VOCA and the new PyTorch3D renderer. As can be observed, the rendering time is dramatically improved, although the overall time is still limited by the time required to process and load the meshes and texture. This increases the attainble FPS from approximately 3 FPS in the original renderer to approximately 14 FPS in the new PyTorch3D renderer, significantly pushing the pipeline towards real-time operation. Overall, the newly implemented renderer is nearly 5 times faster than the original.

Moreover, if the time taken to pre-process the meshes and convert the rendered image sequence to video is neglected, the rendering process is actually optimised by over 24

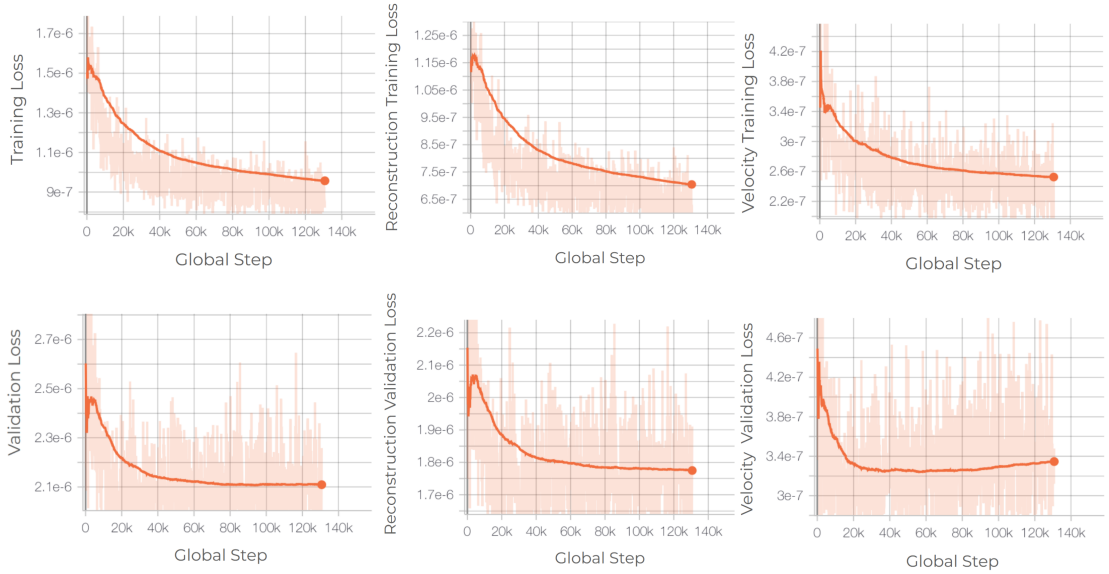


Figure 5.4.: Training and validation (velocity and reconstruction) losses for the HuBERT + VOCASET trained model on the VOCA architecture.

times compared to the VOCA's original renderer, achieving an FPS surpassing 100 FPS (compared to 4 FPS on the original renderer), which is more than sufficient to achieve real-time capabilities.

The two renderers are compared visually in Figure 5.5, to show that the optimisation in time of the new renderer has not in any regards reduced the final output resolution. It can be noted that camera and lighting perspectives may differ slightly between the rendered outputs, which can be easily altered, albeit difficult to match exactly to the original, without affecting rendering times.

5.3. Emotional 3D Dataset

Out of the 214 clips generated for the 3D dataset from 2D MEAD, 81% of clips were accurately converted (without jitters), establishing the method as a reasonably reliable method to convert any 2D dataset into a corresponding 3D dataset.

Additionally, the resulting validation losses (which is the sum of reconstruction and velocity loss - see Section 4.4) for training using the newly generated dataset, with methods 1, 2 and 3 of generating the 'init_expression_basis' file can be observed on Figures 5.6, 5.7 and 5.8 respectively. This can be compared against the model trained on the VOCASET dataset in Figure 5.3.

All three methods of training using the new dataset achieve the same minimum loss

	Processing Meshes [s]	Rendering [s]	Converting to Video [s]	Total [s]	FPS
Original Pyrender Renderer	30.2	69.8	0.6	100.6	3
PyTorch3D Renderer	17.2	2.9	1.0	21.1	14

Table 5.3.: Rendering time comparison (on a Nvidia RTX 2080) between the original Pyrender renderer and the new PyTorch3D renderer for processing of a 5 second (approximately 300 frame) textured sequence. The attainable FPS is calculated as 300 / total render time.

Figure 5.5.: Visual demonstration of a renderered frame in an animation sequence using the old Pyrender renderer (left), and the new PyTorch3D renderer (right).

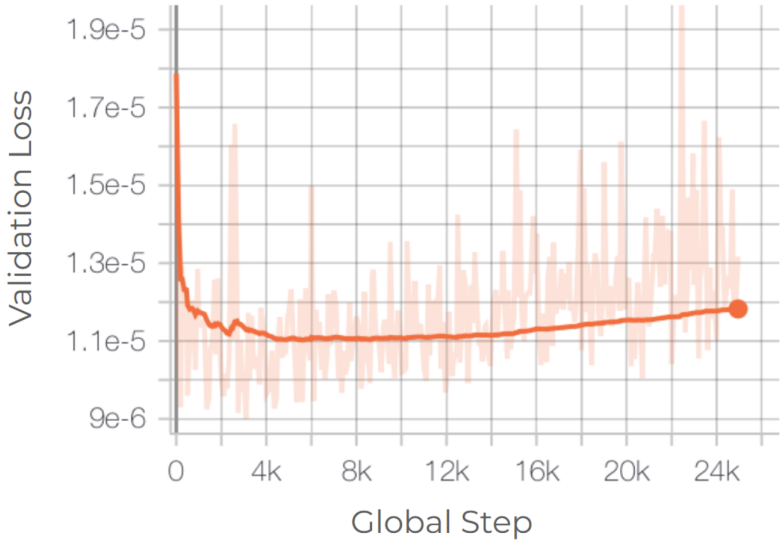


Figure 5.6.: Training using the 3D MEAD dataset on the initial expression basis weights provided by VOCASET.

of 1.1e-5 (compared to 2.1e-6 using the original VOCASET dataset), as shown. This is likely due to the addition of expression and pose in the dataset (where VOCASET had zero-posed expression and poses), which is necessary for the training of an emotional model, however in the original VOCA model architecture (which does not take emotion-based parameters into consideration), could result in higher validation losses. There is also an increase of overfitting, although this is likely due to the dataset being significantly shorter (80 clips, as only a third of the dataset is utilised, compared to 400 clips in VOCASET).

Importantly to note, the results generated using the new dataset show the same trends as the original VOCASET dataset, which is promising, and shows it is likely that with a much larger converted 3D dataset, as well as a modification of VOCA's model architecture to incorporate emotion-based parameter inputs, much better results can be obtained.

As the talking-head synthesised from a model trained on the new dataset in each of the three methods display similar results, and also achieves similar methods, it can be concluded that using each of the three methods is equally valid, and thus the zero weight method can be chosen for simplicity.

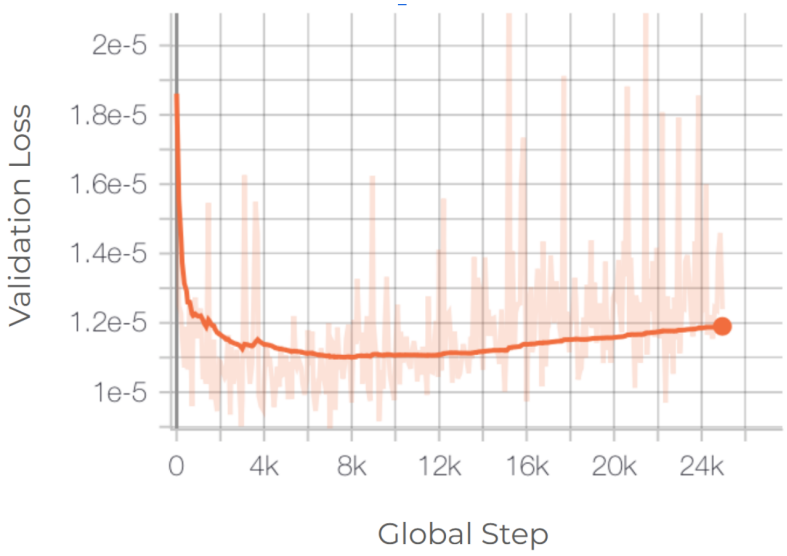


Figure 5.7.: Training using the 3D MEAD dataset on an array of initial expression basis weights initialised to zero.

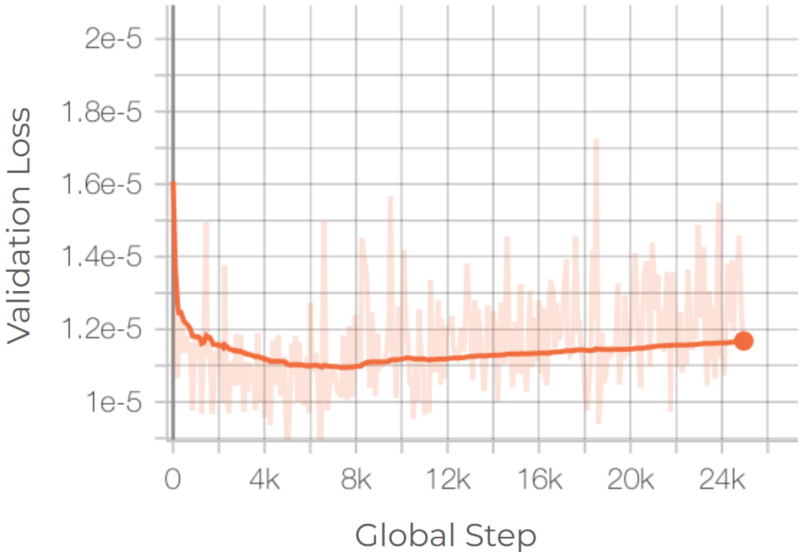


Figure 5.8.: Training using the 3D MEAD dataset on the array of initial expression basis weights saved from training on the zeroed-out weights.

Chapter 6

Conclusion

Following this thesis, a basic pipeline is established that allows a purely audio-driven high-fidelity 3D talking-head to be generated from a single in-the-wild 2D image, generalizing to different subjects and speaker styles. This hugely reduces the complexity of the input modality (3D mesh template) that would otherwise need to be crafted by an artist, or generated, as in this work, with the mesh topology aligned, which is a non-trivial process.

The real-time capabilities of the established pipeline (originally not feasible), is also significantly improved, with up to 5x improvement in speed achieved in the rendering stage, and over 1.6x improvement in speed achieved in the audio processing stage. This enables a processing frame rate of 14 FPS, where previously only 3 FPS was obtainable (bottleneck due to the rendering stage). Neglecting the rendering stage, the work attains real-time capabilities, of up to 28 FPS.

In terms of emotion enhancement, this thesis contributes one of the only 3D emotional datasets, featuring 8 emotions and a large, balanced variety of sentences. This dataset is generated using the recently released (2D) MEAD dataset, with a proposed method to convert any 2D dataset to a corresponding 3D dataset, which could have massive benefits to future emotional 3D talking-head models.

In summary, this thesis has combined and improved all desired properties of an audio-driven talking head model, where previously only select features were optimised. The final pipeline allows audio-driven 3D talking heads to be synthesized, in near real-time, from a single in-the-wild 2D image.

Appendix A

Declaration of Originality

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

TOWARDS REAL-TIME AUDIO-DRIVEN EMOTIONAL TALKING HEADS

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
GAN	TIANHONG

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Zurich, March 2023	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Appendix B

3D Emotional Dataset

B.1. Dataset Split

The following Tables B.1, B.2 and B.3 show the chosen subset of the MEAD dataset converted to 3D format, and the corresponding dataset splits.

B.2. MEAD Transcript

The following Figures B.1 and B.2 show the transcript of the sentences present in the MEAD dataset.

	Common	Unique	Neutral	Total
1st male (M003)	1*0: 28 1*1: 28 1*2: 28 1*3: 28 1*4: 28 1*5: 28 1*6: 28 1*7: 28	3*1: 14, 15, 16 2*2: 5, 6 2*3: 6, 7 2*4: 4, 5 2*5: 4, 5 2*6: 4, 5 3*7: 6, 7, 8	1*0: 4	25
2nd male (M005)	1*0: 27 1*1: 27 1*2: 27 1*3: 27 1*4: 27 1*5: 27 1*6: 27 1*7: 27	2*1: 6, 7 3*2: 7, 8, 9 2*3: 4, 5 2*4: 6, 7 2*5: 6, 7 2*6: 6, 7 2*7: 9, 10	2*0: 5, 6	25
3rd male (M009)	1*0: 30 1*1: 30 1*2: 30 1*3: 30 1*4: 30 1*5: 30 1*6: 30 1*7: 30	2*1: 8, 9 2*2: 10, 11 3*3: 8, 9, 10 2*4: 8, 9 2*5: 8, 9 2*6: 8, 9 2*7: 4, 5	2*0: 7, 8	25
1st female (W011)	1*0: 25 1*1: 25 1*2: 25 1*3: 25 1*4: 25 1*5: 25 1*6: 25 1*7: 25	2*1: 10, 11 2*2: 12, 13 2*3: 11, 12 3*4: 10, 11, 12 2*5: 19, 20 2*6: 10, 11 2*7: 11, 12	2*0: 9, 10	25
2nd female (W015)	1*0: 23 1*1: 23 1*2: 23 1*3: 23 1*4: 23 1*5: 23 1*6: 23 1*7: 23	2*1: 12, 13 2*2: 14, 15 2*3: 13, 14 2*4: 13, 14 3*5: 10, 11, 12 2*6: 12, 13 2*7: 13, 14	2*0: 11, 12	25
3rd female (W016)	1*0: 21 1*1: 21 1*2: 21 1*3: 21 1*4: 21 1*5: 21 1*6: 21 1*7: 21	2*1: 17, 18 2*2: 16, 17 2*3: 15, 16 2*4: 15, 16 2*5: 13, 14 3*6: 14, 15, 16 2*7: 15, 16	2*0: 13, 14	25
Total	6*0, 6*1, 6*2, 6*3, 6*4, 6*5, 6*6, 6*7,	13*1, 13*2, 13*3, 13*4, 13*5, 13*6, 13*7	11*0	150

Table B.1.: Selected dataset subset, training split. The notation displayed in the table is a*b: x, where a is the number of clips selected, b is the emotion selected (0 = neutral, 1 = angry, 2 = contempt, 3 = disgusted, 4 = fear, 5 = happy, 6 = sad, 7 = surprised), and x is the sentence number for that emotion (see Appendix B.2 for sentence number).

	Common	Unique	Neutral	Total
4th female (W014)	1*0: 24 1*1: 24 1*2: 24 1*3: 24 1*4: 24 1*5: 24 1*6: 24 1*7: 24	1*1: 19 1*2: 18 1*3: 17 1*4: 19 1*5: 15 1*6: 17 1*7: 17	1*0: 15	16
4th male (M012)	1*0: 26 1*1: 26 1*2: 26 1*3: 26 1*4: 26 1*5: 26 1*6: 26 1*7: 26	1*1: 20 1*2: 19 1*3: 18 1*4: 17 1*5: 16 1*6: 18 1*7: 18	1*0: 16	16
Total	2*0, 2*1, 2*2, 2*3, 2*4, 2*5, 2*6, 2*7	2*1, 2*2, 2*3, 2*4, 2*5, 2*6, 2*7	2*0	32

Table B.2.: Selected dataset subset, validation split. The notation displayed in the table is a*b: x, where a is the number of clips selected, b is the emotion selected (0 = neutral, 1 = angry, 2 = contempt, 3 = disgusted, 4 = fear, 5 = happy, 6 = sad, 7 = surprised), and x is the sentence number for that emotion (see Appendix B.2 for sentence number).

	common	unique	neutral	total
5th female (W009)	1*0: 29 1*1: 29 1*2: 29 1*3: 29 1*4: 29 1*5: 29 1*6: 29 1*7: 29	1*1: 4 1*2: 20 1*3: 19 1*4: 18 1*5: 17 1*6: 19 1*7: 19	1*0: 17	16
5th male (M007)	1*0: 2 1*1: 2 1*2: 2 1*3: 2 1*4: 2 1*5: 2 1*6: 2 1*7: 2	1*1: 1 1*2: 4 1*3: 20 1*4: 20 1*5: 18 1*6: 20 1*7: 20	1*0: 18	16
total	2*0, 2*1, 2*2, 2*3, 2*4, 2*5, 2*6, 2*7	2*1, 2*2, 2*3, 2*4, 2*5, 2*6, 2*7	2*0	32

Table B.3.: Selected dataset subset, testing split. The notation displayed in the table is a*b: x, where a is the number of clips selected, b is the emotion selected (0 = neutral, 1 = angry, 2 = contempt, 3 = disgusted, 4 = fear, 5 = happy, 6 = sad, 7 = surprised), and x is the sentence number for that emotion (see Appendix B.2 for sentence number).

Angry:

1. She had your dark suit and greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Will you tell me why?
4. Who authorised the unlimited expense account?
5. Destroy every file related to my audits.
6. The cat's meow always hurts my eyes.
7. Why else would Danny allow others to go?
8. Why do we need bigger and better bombs?
9. Nuclear rockets can destroy airfields with ease.
10. You're so preoccupied you've let your faith grow dim.
11. Cory and Trish played tag with beach balls for hours.
12. He will allow a rare lie.
13. Withdraw all phoney accusations at once.
14. Right now may not be the best time for business mergers.
15. Kindergarten children decorate their classroom for all holidays.
16. A few years later the dome fell in.
17. But in this one section, we welcome the auditors.
18. Lots of people are roaming the streets with costumes and masks and having a ball.
19. In many of his poems, death comes by train - a strong evocative visual image.
20. Then he would realise there are things only he himself would think.
21. Tot placed top priority on getting his bike fixed.
22. One even gave my little dog a biscuit.
23. I'll have a scoop of that exotic purple and turquoise sorbet.
24. Land based radar will help with his task.
25. The plaintiff in school does segregation cases.
26. His superiors had also preached this, saying it was a way for eternal honour.
27. It was not whatever tale was told by tales.
28. No, the man was not drunk. He wondered how he had got tied up with this stranger.
29. No price is too high when true love is at stake.
30. The revolution is now on the way and material handling makes it much easier.

Contempt:

1. She had your dark suit and greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Will you tell me why?
4. Are your grades higher or lower than Nancy's?
5. This was easy for us.
6. Only lawyers love millionaires.
7. It's illegal to postdate a check.
8. He stole a dime from a beggar.
9. His failure to open the store by 8 cost him his job.
10. Let us differentiate a few of these ideas.
11. The big dog loved to chew on the old rag doll.
12. Family loyalties in corporate work have been unbroken for generations.
13. Withdraw only as much money as you need.
14. The ways to rent a chauffeur driven car.
15. No one material is best for all situations.
16. Mosquitoes exist in warm humid environments.
17. We of the liberal led world, got all set for peace and rehabilitation.
18. Can your insurance company aid you in reducing administrative costs?
19. She sprang up and went swiftly to the bedroom.
20. He ate 4 extra eggs for breakfast.
21. Tot placed top priority in getting his bike fixed.
22. One even gave my little dog a biscuit.
23. I'll have a scoop of that exotic purple and turquoise sorbet.
24. Land based radar will help with his task.
25. The plaintiff in school does segregation cases.
26. His superiors had also preached this, saying it was a way for eternal honour.
27. It was not whatever tale was told by tales.
28. No, the man was not drunk. He wondered how he had got tied up with this stranger.
29. No price is too high when true love is at stake.
30. The revolution is now on the way and material handling makes it much easier.

Disgusted:

1. She had your dark suit and greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Will you tell me why?
4. Please take this dirty table cloth to the cleaners for me.
5. The small boy put the worm on the hook.
6. You're not leaving up to your own principles, she told my discouraged people.
7. Don't do Charlie's dirty dishes.
8. Will Robin wear a yellow lily?
9. Young children should avoid exposure to contagious diseases.
10. Military personnel are expected to obey government orders.
11. Basketball can be an entertaining sport.
12. How good is your endurance?
13. Bart burned paper and leaves in a big bonfire.
14. December and January are nice months to spend in Miami.
15. If people were more generous, there would be no need for welfare.
16. If the farmers rent it, the rent must be paid.
17. Laboratory asks for physics.
18. Pretty soon a lady came along carrying a folded umbrella as a walking stick.
19. How much and how many profits could the majority make out of the losses of a few.
20. Does society really exist as an anti and over conglomeration over men? (?)
21. Tot placed top priority on getting his bike fixed.
22. One even gave my little dog a biscuit.
23. I'll have a scoop of that exotic purple and turquoise sorbet.
24. Land based radar will help with his task.
25. The plaintiff in school does segregation cases.
26. His superiors had also preached this, saying it was a way for eternal honour.
27. It was not whatever tale was told by tales.
28. No, the man was not drunk. He wondered how he had got tied up with this stranger.
29. No price is too high when true love is at stake.
30. The revolution is now on the way and material handling makes it much easier.

Fear:

1. She had your dark suit and greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Will you tell me why?
4. Call an ambulance for medical assistance!
5. Tornadoes often destroy acres of farmlands.
6. Destroy every file related to my audits.
7. Would you allow acts of violence?
8. The high security prison was surrounded by barbed wire.
9. His shoulders felt as if they were broken.
10. The fish began to leap frantically on the surface of the small lake.
11. Straw hats are out of fashion this year.
12. That diagram makes sense only after much study.
13. Special task forces rescue hostages from kidnappers.
14. The tooth fairy forgot to come when Roger's tooth fell out.
15. Will Robin wear a yellow lily?
16. Their props are two step-ladders, a chair and a pom van.
17. This is a problem that goes considerably beyond questions of salary and engineer.
18. The pulsing glow of a cigarette.
19. One looked down on a sea of leaves, a break wave of flowers.
20. You will achieve a more vivid sense of what it is, by realising what it is not.
21. Tot placed top priority on getting his bike fixed.
22. One even gave my little dog a biscuit.
23. I'll have a scoop of that exotic purple and turquoise sorbet.
24. Land based radar will help with his task.
25. The plaintiff in school does segregation cases.
26. His superiors had also preached this, saying it was a way for eternal honour.
27. It was not whatever tale was told by tales.
28. No, the man was not drunk. He wondered how he had got tied up with this stranger.
29. No price is too high when true love is at stake.
30. The revolution is now on the way and material handling makes it much easier.

Figure B.1.: Transcript of the MEAD [32] dataset (part 1). Sentences highlighted in green and blue are common between all emotions. Sentences highlighted in orange or red and shared between the neutral emotion and one or two other emotions. Sentences in black are unique sentences for that emotion.

Happy:

1. She had your dark suit and greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Will you tell me why?
4. Those musicians harmonise marvellously.
5. The eastern coast is a place for pure pleasure and excitement.
6. Tim takes Sheila to see movies twice a week.
7. They used an aggressive policeman to flag thoughtless motorists.
8. When you're less fatigued, things just naturally look brighter.
9. By that time, perhaps something better can be done.
10. She found herself able to sing any role and any song that struck her fancy.
11. That noise problem grows more annoying each day.
12. Project development was progressing too slowly.
13. The oasis was a mirage.
14. Are your grades higher or lower than Nancy's?
15. Serve the coleslaw, after add the oil.
16. By that, one feels that magnetic forces are as general as electric forces.
17. His artistic accomplishments guaranteed him access into any social gathering.
18. He would not carry a briefcase.
19. Obviously, the bridal pair has many adjustments to make to their new situation.
20. Both the conditions and the complexity are documented in considerable detail.
21. Tot placed top priority on getting his bike fixed.
22. One even gave my little dog a biscuit.
23. I'll have a scoop of that exotic purple and turquoise sorbet.
24. Land based radar will help with his task.
25. The plaintiff in school does segregation cases.
26. His superiors had also preached this, saying it was a way for eternal honour.
27. It was not whatever tale was told by tales.
28. No, the man was not drunk. He wondered how he had got tied up with this stranger.
29. No price is too high when true love is at stake.
30. The revolution is now on the way and material handling makes it much easier.

Sad:

1. She had your dark suit and greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Will you tell me why?
4. The prospect of cutting back spending is an unpleasant one for any governor.
5. The diagnosis was discouraging, however he was not overly worried.
6. We can die too, we can die like real people. People never live forever.
7. He didn't forget her at all, and if he found other women, it'd be bad.
8. There would still be plenty of moments of regret and sadness, and guilty relief.
9. She drank greedily, and murmured thank you, as she lowered her head.
10. There's no chance now, of all of us getting away.
11. Before Thursday's exam, review every formula.
12. They enjoy it while I audition.
13. John cleaned shellfish for a living.
14. He stole a dime from a beggar.
15. Jeff thought you argued in favour of a -__ purchase. (intensity?)
16. However, the letter remained, augmented by several lunchroom stoppers.
17. American newspaper reviewers like to call this place nihilistic.
18. But the ships are very slow now, and we don't get so many sailors anymore.
19. It is one of the rare public ventures here, on which everyone is agreed.
20. No manufacturer has taken the initiative in pointing out the costs involved.
21. Tot placed top priority on getting his bike fixed.
22. One even gave my little dog a biscuit.
23. I'll have a scoop of that exotic purple and turquoise sorbet.
24. Land based radar will help with his task.
25. The plaintiff in school does desegregation cases.
26. His superiors had also preached this, saying it was a way for eternal honour.
27. It was not whatever tale was told by tales.
28. No, the man was not drunk. He wondered how he had got tied up with this stranger.
29. No price is too high when true love is at stake.
30. The revolution is now on the way and material handling makes it much easier.

Surprised:

1. She had your dark suit and greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Will you tell me why?
4. The carpet cleaners shampooed our oriental rug.
5. His shoulders felt as if it was broken.
6. The patient and the surgeon and both recuperating from a lengthy operation.
7. He ate four eggs for breakfast.
8. While waiting for Chipper, she crisscrossed the square many times.
9. I just saw Jim near the new archeological museum.
10. I took her word for it, but is she really going with you?
11. The viewpoint overlooked the ocean.
12. I'd ride the subway, but I haven't enough change.
13. The clumsy customer spilled some expensive perfume.
14. Please take my potatoes out of defrost.
15. Grandmother outgrew her upbringings and petticoats.
16. Salvation reconsidered.
17. Properly used present books are an excellent instrument of enlightenment.
18. Lighted windows glow duo bright through the downboard?
19. But this doesn't detract from its merit as an interesting if not great film.
20. He further proposed grants of an unspecified sum for experimental hospitals.
21. Tot placed top priority on getting his bike fixed.
22. One even gave my little dog a biscuit.
23. I'll have a scoop of that exotic purple and turquoise sorbet.
24. Land based radar will help with his task.
25. The plaintiff in school does segregation cases.
26. His superiors had also preached this, saying it was a way for eternal honour.
27. It was not whatever tale was told by tales.
28. No, the man was not drunk. He wondered how he had got tied up with this stranger.
29. No price is too high when true love is at stake.
30. The revolution is now on the way and material handling makes it much easier.

Neutral:

1. She had your dark suit and greasy wash water all year.
2. Don't ask me to carry an oily rag like that.
3. Will you tell me why?
4. Bridges, tunnels and ferries are the most common methods of river crossings.
5. The moment of truth is the moment of crisis.
6. The best way to learn is to solve extra problems.
7. Therpon followed the demonstration that tyranny knew no ideological confines.
8. Calcium makes bones and teeth strong.
9. Catastrophic economic cutbacks neglect the poor.
10. Allow leeway here, but rationalise all errors.
11. Greg buys fresh milk each weekday morning.
12. Agriculture products are unevenly distributed.
13. The nearest synagogue may not be within walking distance.
14. As such, it was beyond politics and had no need of justification by message.
15. He always seemed to have money in his pocket.
16. No return address whatsoever.
17. Keep your seats boys, I just want to put some finishing touches on this thing.
18. He ripped down the cellophane carefully, and laid three dogs down on the tinfoil.
19. Who authorised the unlimited expense account?
20. Destroy every file related to my audits.
21. Please take this dirty tablecloth to the cleaners for me.
22. The small boy put the worm on the hook.
23. Call an ambulance for medical assistance.
24. Tornadoes often destroy acres of farmland.
25. The carpet cleaner shampooed our oriental rug.
26. His shoulders felt as if it were broken.
27. The prospect of cutting back spending is an unpleasant one for any governor.
28. The diagnosis was discouraging, however he was not overly worried.
29. Those musicians harmonise marvellously.
30. The eastern coast is a place for pure pleasure and excitement.
31. Tot placed top priority on getting his bike fixed.
32. One even gave my little dog a biscuit.
33. I'll have a scoop of that exotic purple and turquoise sorbet.
34. Land based radar will help with his task.
35. The plaintiff in school does segregation cases.
36. His superiors had also preached this, saying it was a way for eternal honour.
37. It was not whatever tale was told by tales.
38. No, the man was not drunk. He wondered how he had got tied up with this stranger.
39. No price is too high when true love is at stake.
40. The revolution is now on the way and material handling makes it much easier.

Figure B.2.: Transcript of the MEAD [32] dataset (part 2). Sentences highlighted in green and blue are common between all emotions. Sentences highlighted in orange or red and shared between the neutral emotion and one or two other emotions. Sentences in black are unique sentences for that emotion.

List of Figures

List of Tables

Bibliography

[1] Yang Zhou et al. "MakeItTalk: Speaker-Aware Talking-Head Animation". In: *ACM Transactions on Graphics* 39.6 (2020).

[2] Suzhen Wang et al. "Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion". In: *CoRR* abs/2107.09293 (2021). arXiv: 2107.09293. URL: https://arxiv.org/abs/2107.09293.

[3] Suzhen Wang et al. "One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning". In: *CoRR* abs/2112.02749 (2021). arXiv: 2112.02749. URL: https://arxiv.org/abs/2112.02749.

[4] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. "Realistic Speech-Driven Facial Animation with GANs". In: *CoRR* abs/1906.06337 (2019). arXiv: 1906.06337. URL: http://arxiv.org/abs/1906.06337.

[5] Yuanxun Lu, Jinxiang Chai, and Xun Cao. "Live Speech Portraits: Real-Time Photorealistic Talking-Head Animation". In: *CoRR* abs/2109.10595 (2021). arXiv: 2109.10595. URL: https://arxiv.org/abs/2109.10595.

[6] Guo Yudong et al. "AD-NeRF: Audio Driven Neural Radiance Fields for Talking Head Synthesis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 5784–5794.

[7] Lele Chen et al. "Talking-head Generation with Rhythmic Head Motion". In: *CoRR* abs/2007.08547 (2020). arXiv: 2007.08547. URL: https://arxiv.org/abs/2007.08547.

[8] Li et al. "Write-a-speaker: Text-based Emotional and Rhythmic Talking-head Generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.3 (May 2021), pp. 1911–1920. DOI: 10.1609/aaai.v35i3.16286. URL: https://ojs.aaai.org/index.php/AAAI/article/view/16286.

[9] Xinya Ji et al. "EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model". In: *ACM SIGGRAPH 2022 Conference Proceedings*. SIGGRAPH '22. Vancouver, BC, Canada: Association for Computing Machinery, 2022. ISBN: 9781450393379. DOI: 10.1145/3528233.3530745. URL: https://doi.org/10.1145/3528233.3530745.

[10] K R Prajwal et al. *A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild*. 2020. arXiv: 2008.10010 [cs.CV].

[11] Haozhe Wu et al. "Imitating Arbitrary Talking Style for Realistic Audio-Driven Talking Face Synthesis". In: *Proceedings of the 29th ACM International Conference on Multimedia*. ACM, Oct. 2021. DOI: 10.1145/3474085.3475280. URL: https://doi.org/10.1145%5C%2F3474085.3475280.

[12] Alexander Richard et al. "MeshTalk: 3D Face Animation From Speech Using Cross-Modality Disentanglement". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 1173–1182.

[13] Daniel Cudeiro et al. "Capture, Learning, and Synthesis of 3D Speaking Styles". In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 10101–10111. URL: http://voca.is.tue.mpg.de/.

[14] Yu Zixiao, Wang Haohong, and Jian Ren. "RealPRNet: A Real-Time Phoneme-Recognized Network for "Believable" Speech Animation". In: *IEEE Internet of Things Journal* 9.7 (2022), pp. 5357–5367. DOI: 10.1109/JIOT.2021.3110468.

[15] Qianyun Wang, Zhenfeng Fan, and Shihong Xia. "3D-TalkEmo: Learning to Synthesize 3D Emotional Talking Head". In: *CoRR* abs/2104.12051 (2021). arXiv: 2104.12051. URL: https://arxiv.org/abs/2104.12051.

[16] Guanzhong Tian, Yi Yuan, and Yong Liu. "Audio2Face: Generating Speech/Face Animation from Single Audio with Attention-Based Bidirectional LSTM Networks". In: *2019 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. 2019, pp. 366–371. DOI: 10.1109/ICMEW.2019.00069.

[17] Awni Y. Hannun et al. "Deep Speech: Scaling up end-to-end speech recognition". In: *CoRR* abs/1412.5567 (2014). arXiv: 1412.5567. URL: http://arxiv.org/abs/1412.5567.

[18] Tianye Li et al. "Learning a model of facial shape and expression from 4D scans". In: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36.6 (2017), 194:1–194:17. URL: https://doi.org/10.1145/3130800.3130813.

[19] Xinya Ji et al. "Audio-Driven Emotional Video Portraits". In: *CoRR* abs/2104.07452 (2021). arXiv: 2104.07452. URL: https://arxiv.org/abs/2104.07452.

[20] Zhibo Wang et al. "Emotion-Preserving Blendshape Update With Real-Time Face Tracking". In: *IEEE Transactions on Visualization and Computer Graphics* 28.6 (2022), pp. 2364–2375. DOI: 10.1109/TVCG.2020.3033838.

[21] Yao Feng et al. "Learning an Animatable Detailed 3D Face Model from In-The-Wild Images". In: *CoRR* abs/2012.04012 (2020). arXiv: 2012.04012. URL: https://arxiv.org/abs/2012.04012.

[22] Chen Cao et al. "Real-Time 3D Neural Facial Animation from Binocular Video". In: *ACM Trans. Graph.* 40.4 (July 2021). ISSN: 0730-0301. DOI: 10.1145/3450626.3459806. URL: https://doi.org/10.1145/3450626.3459806.

[23] Matthew Loper et al. "SMPL: A Skinned Multi-Person Linear Model". In: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Oct. 2015), 248:1–248:16.

[24] Chen Cao et al. "Facewarehouse: A 3D facial expression database for visual computing". In: *IEEE Transactions on Visualization and Computer Graphics* 20.3 (2014), pp. 413–425. DOI: 10.1109/tvcg.2013.249.

[25] Pascal Paysan et al. "A 3D Face Model for Pose and Illumination Invariant Face Recognition". In: *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*. 2009, pp. 296–301. DOI: 10.1109/AVSS.2009.58.

[26] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[27] Soubhik Sanyal et al. "Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[28] Zhen-Hua Feng et al. *Evaluation of Dense 3D Reconstruction from 2D Face Images in the Wild*. 2018. DOI: 10.48550/ARXIV.1803.05536. URL: https://arxiv.org/abs/1803.05536.

[29] Nikhila Ravi et al. *Accelerating 3D Deep Learning with PyTorch3D*. 2020. DOI: 10.48550/ARXIV.2007.08501. URL: https://arxiv.org/abs/2007.08501.

[30] Wei-Ning Hsu et al. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units". In: *CoRR* abs/2106.07447 (2021). arXiv: 2106.07447. URL: https://arxiv.org/abs/2106.07447.

[31] Atijit Anuchitanukul and Lucia Specia. *Burst2Vec: An Adversarial Multi-Task Approach for Predicting Emotion, Age, and Origin from Vocal Bursts*. 2022. DOI: 10.48550/ARXIV.2206.12469. URL: https://arxiv.org/abs/2206.12469.

[32] Kaisiyuan Wang et al. "MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation". In: *ECCV*. Aug. 2020.

[33] Steven R. Livingstone and Frank A. Russo. *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. Version 1.0.0. Funding Information Natural Sciences and Engineering Research Council of Canada: 2012-341583 Hear the world research chair in music and emotional speech from Phonak. Zenodo, Apr. 2018. DOI: 10.5281/zenodo.1188976. URL: https://doi.org/10.5281/zenodo.1188976.