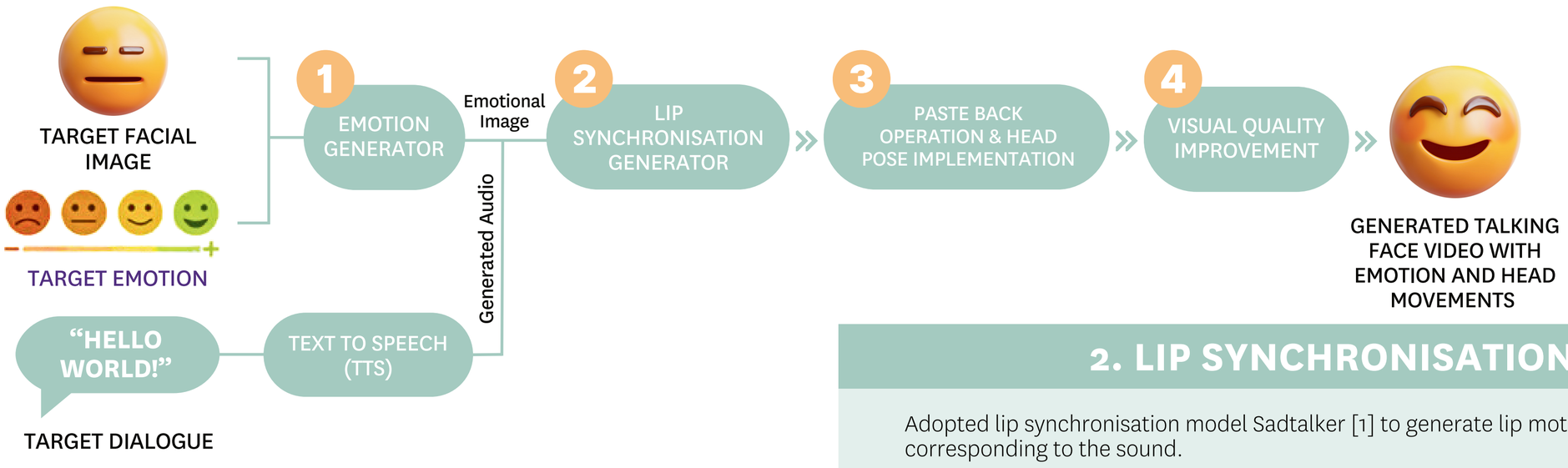
EMOTIONAL TALKING FACE GENERATION SYSTEM

Gayeon Kim & Yugyeong Hong

Supervisor: Ho Seok Ahn & Trevor Gee

HAVE YOU EVER WISHED TO SPEAK TO DEPARTED LOVED ONES AGAIN?

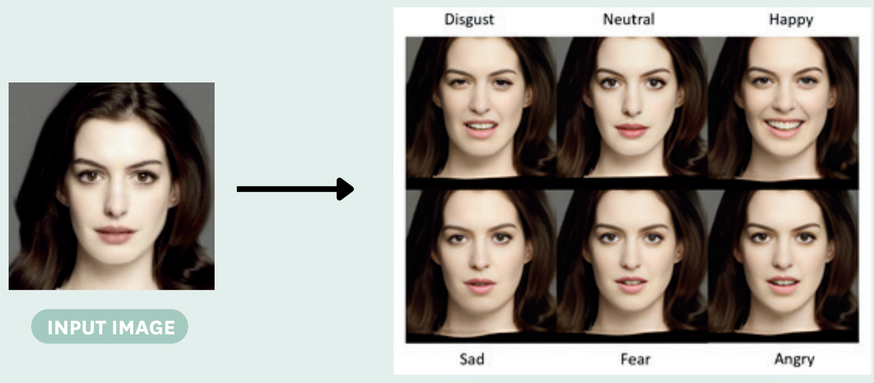
Our project generates a realistic talking face with emotions using just a single facial image and text input.



1. EMOTION

Given a target facial image(2D image portrait), Emotion Generator generate emotional images.

- Adopted emotional discriminator [Eskimez et al., 2020] to generate emotional images, expressning all of Ekman's categorical emotions.
- Incorporating an emotional discriminator and integrating LSTM layers into the generatormodel, thereby learning distinct emotional traits.

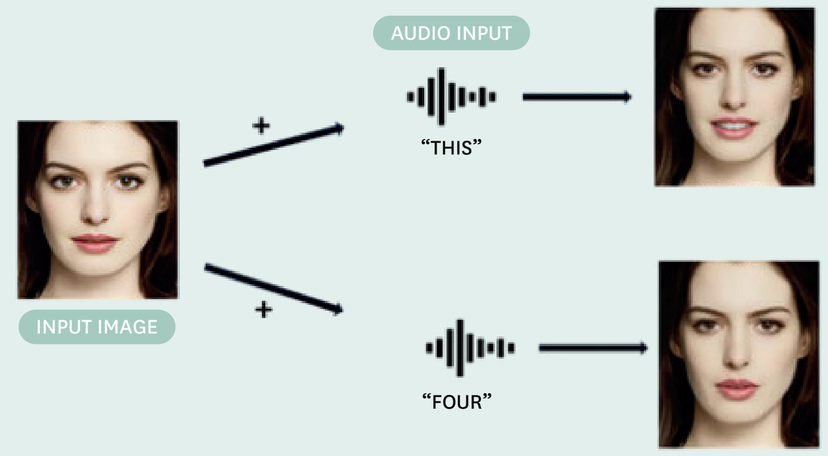


2. LIP SYNCHRONISATION

Adopted lip synchronisation model Sadtalker [1] to generate lip motions corresponding to the sound.

Sadtalker

- Uses Wav2Lip [2] at its initial stage, which employs a lip discriminator to generate accurate lip motions corresponding to the sound.
- Then 3DMM reconstructs 3D face structure to capture and render various facial expressions, including lip movements.
- Only the lip movements employed in our project aiming for improved accuracy in lip synchronisation.



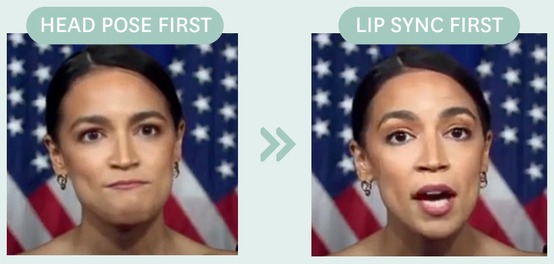
[1] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, Fei Wang, "SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation", 2023
[2] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri and C. V. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation in the Wild", 2020

3. PASTE BACK OPERATION & HEAD MOVEMENT

Restored the image to its uncropped state first to ensure accurate head pose implementation by providing spatial references around the face.

Then, leveraged DPE [3] to implement a head pose from the selected video source

- More accurate head pose as it does not rely on paired data or predefined 3DMMs.
- Focuses on disentanglement of pose and expression in the latent space using the Motion Editing module.
- Order is essential here since implementing a head pose before the lip synchronisation does not work.

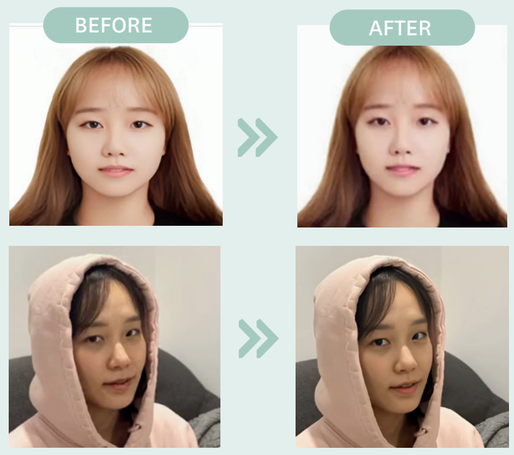


[3] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, Dong-ming Yan, "DPE: Disentanglement of Pose and Expression for General Video Portrait Editing", 2023

4. IMPROVING VISUAL QUALITY

Integrated the Codeformer [4] to enhance the visual quality.

- Maps low-quality inputs to high-quality outputs then predicts the corrupted or missing parts
- Optimises the results across varying levels of degradation.



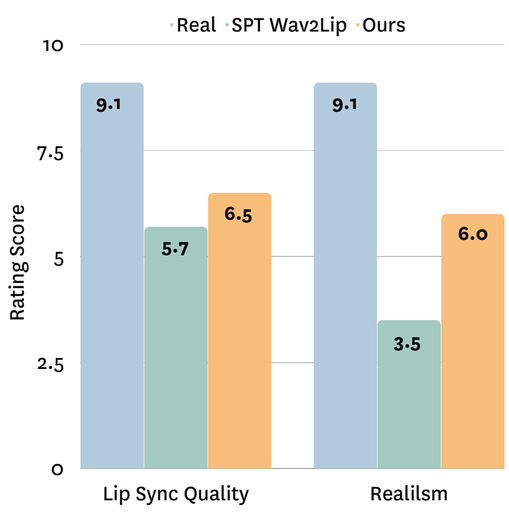
[4] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, Chen Change Loy, "Towards Robust Blind Face Restoration with Codebook Lookup Transformer", 2022

RESULTS

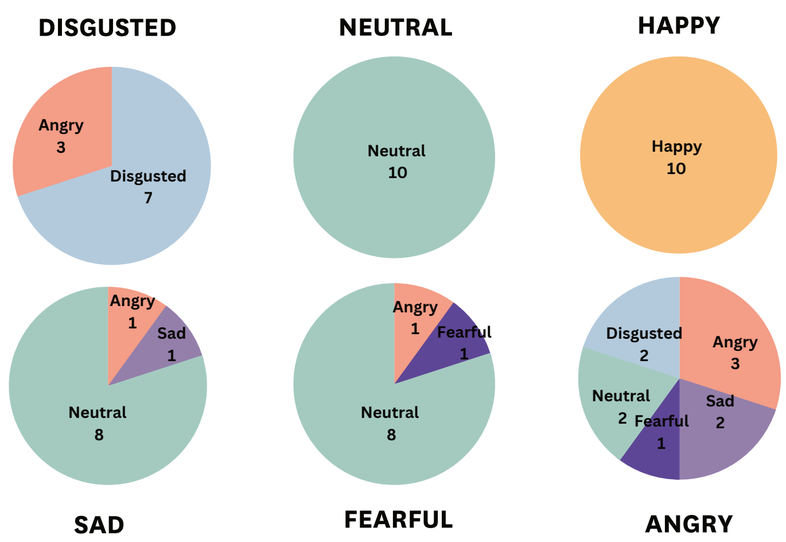
- Our model improved in lip synchronisation quality outperforming SPT Wav2Lip which was developed by last year's students.
- There was a notable improvement in realism, which was determined by both lip synchronization and head pose implementation.
- Our model successfully generated enhanced-quality emotional images expressing all of Ekman's categorical emotion.

Data collected from individuals ranging in age from 18 to 50 to ensure diverse feedback. Sample size of 10

LIP SYNC & REALISM SCORE
(out of 10, 10 being the best)



EMOTION RECOGNISATION SCORE



UNIVERSITY OF AUCKLAND
Waipapa Taumata Rau
NEW ZEALAND

ENGINEERING
DEPARTMENT OF ELECTRICAL, COMPUTER, & SOFTWARE ENGINEERING

PROJECT 64