Faces that Speak: Jointly Synthesising Talking Face and Speech from Text (Supplementary material)

Models	Video Quality		Synchonisation	Diversity
	FID↓	ID-SIM↑	LSE-C↑	DIV↑
1d-conv	19.473	0.849	5.602	0.141
MRF	**18.348**	**0.864**	**5.686**	**0.143**

Table 1. Design choice of audio mapper. These results are obtained on the LRS2 dataset using a one-shot generation setting.

In this supplementary material, we provide additional insights and details that are constrained by space limitations in the main paper. It further offers quantitative and qualitative results to enhance the comprehensive understanding of our framework.

A. Design Choice of Audio Mapper

In our ablation studies on the model architecture of the audio mapper, we compare a conventional 1D convolutional network with the Multi-Receptive Field Fusion (MRF) module. As reported in Table 1, the MRF-based audio mapper enhances the capacity of our framework in every metric. This suggests that incorporating various temporal receptive fields enables our system to effectively convert complex TTS features (linguistic and acoustic) to abundant features, contributing to the generation of more realistic face motions.

B. Ablation on Generation Step of Motion Sampler

We evaluate the TFG performance and time consumption by varying the generation step of our motion sampler. Along with metrics that assess the output quality, we measure the inference speed on a single NVIDIA GeForce RTX 4090 GPU with AMD PRO 3975WX CPU. As indicated in Table 2, a larger step size results in longer inference times. Furthermore, there exists a tendency that the larger step size affects the higher diversity of head pose. Considering this tendency and the latency, we opt to use a step size of 10 for generating talking faces due to its balance between reasonable inference time and performance.

C. User Study on Text-to-Speech

To evaluate perceptual quality of synthesised speech samples, we conduct 5-scale MOS test on two perspectives: naturalness (nMOS) and voice similarity to the target speaker (sMOS). 30 domain-experts evaluated the quality of 30 audio samples while wearing headphones in a controlled environment. The results are shown in Table 3. Above all,

Steps	Video Quality		Synchonisation	Diversity	Latency
	FID↓	ID-SIM↑	LSE-C↑	DIV↑	Speed↑
5	**18.384**	**0.868**	**5.801**	0.132	**1,086**
10	18.348	0.864	5.686	0.143	804
50	18.645	0.857	5.548	**0.151**	131

Table 2. Ablation on synthesis step of motion sampler. "Speed" refers to the number of frames the module can handle per second. *In other words, we measure the time consumed for mapping from audio and prior features to motion features.* These results are obtained on the LRS2 dataset using a one-shot generation setting.

Models	Naturalness	Voice similarity
	nMOS↑	sMOS↑
Ground Truth	4.16±0.18	4.96±0.03
Face-TTS	2.57±0.15	2.18±0.14
Ours (w/ motion)	3.23±0.15	2.45±0.16
Ours (w/o motion)	**3.49±0.15**	**2.94±0.16**

Table 3. MOS results of synthesised speech are presented with 95% confidence interval. nMOS and sMOS represent naturalness and voice similarity, respectively.

our proposed method outperforms Face-TTS in both naturalness and voice similarity. When the motion components are subtracted (*i.e.*, when we use identity features f_{id} rather than f_s), the generation quality and specifically voice similarity are significantly improved. This demonstrates the benefits of using motion-removed features from TFG in synthesising high-quality speech.

D. Failure Cases of Other Baselines

To visually demonstrate the robustness of our framework in the TFG task, we compare it with previous state-of-the-art methods presented in the main paper under challenging conditions for generating realistic talking faces.

As shown in Fig. 1, firstly, Audio2Head struggles to generate natural-looking faces, particularly when the source face is not in a frontal view. Secondly, MakeItTalk fails to generate dynamic lip movements in sync with the audio source. Lastly, SadTalker exhibits artifacts due to its cropping process, resulting in unnatural faces and restricted generatable regions (indicated by yellow boxes). In contrast, our proposed framework consistently produces satisfactory outcomes without the necessity for extracting keypoints or cropping specific regions.

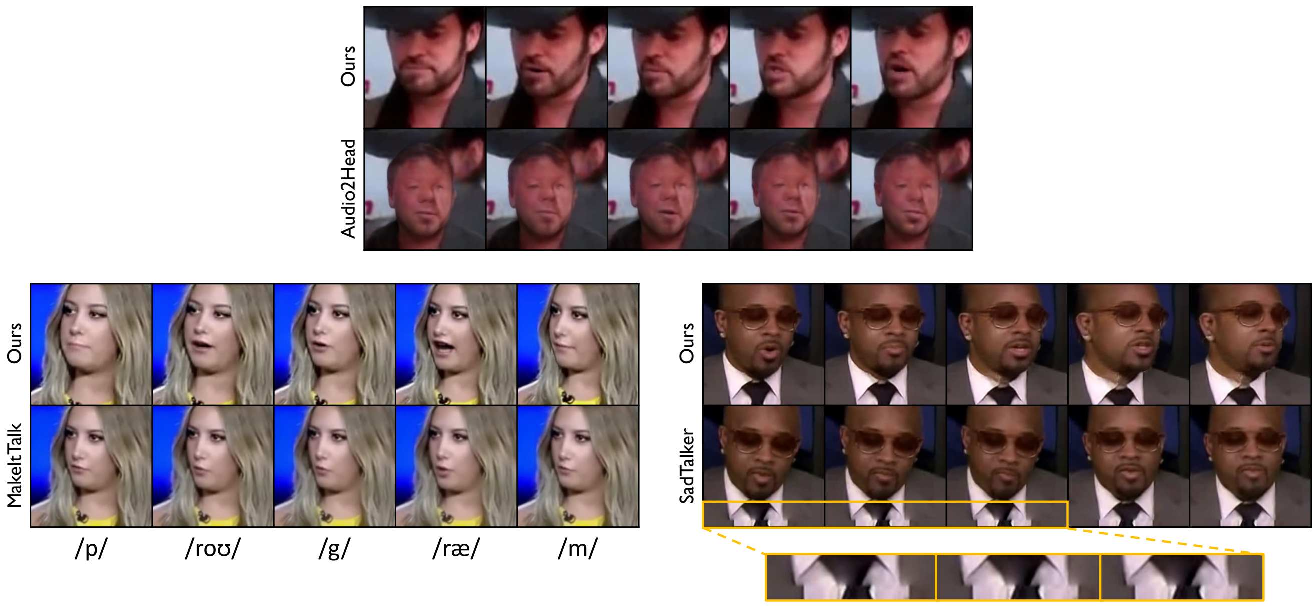

Figure 1. Failure cases of other baselines. We present instances where other baselines fall in generating natural-looking talking faces and compare them with our framework, which consistently exhibits higher-quality results.

E. Generated Image Samples from Identity Features

To validate the effectiveness of identity features acquired by our TFG system, we visualise face images generated from these features. In Fig. 2 (a), we present results from mapping various source images to the reference space, where each face shares the same facial motion but has different identities. The effectiveness of our approach in preserving identity is evident through the distinct and recognisable facial features of each individual. Additionally, as shown in Fig. 2 (b), our approach consistently produces similar images for input images with the same identity, emphasising the method's ability to capture an individual's distinct facial identity despite differences in input images. These results underscore the effectiveness of our approach in finding identity features crucial for generating a consistent style of speeches, even when facial motions differ but the identity remains the same.

F. Additional Qualitative Results

To further support our framework's capacity to generate natural talking faces, we visualise additional qualitative results on LRS2 and VoxCeleb2 datasets under the one-shot generation setting.

As illustrated in Figures 3 and 4, our framework is capable of generating diverse facial motions and natural lip motions that reflect acoustic energy. Our model generates actively moving lip motions aligned to the synthesised speeches (refer to the yellow arrows). Importantly, the utilisation of both linguistic and acoustic features obtained from our TTS system contributes to enhancing the naturalness of the generated talking faces.

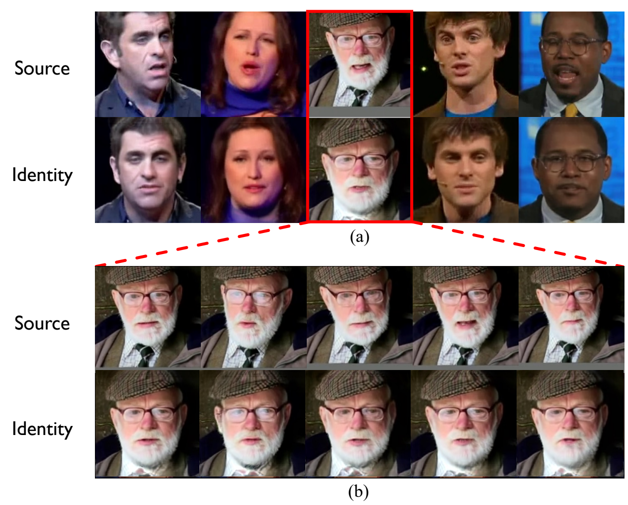

Figure 2. Samples generated with the identity features. We demonstrate how well our model preserves identity by mapping various identities. In Fig. (a), we generate diverse identity image samples having different identities by feeding each identity feature to our generator. In Fig. (b), we further visualise every identity image from a single video. These results prove that our model is robust to maintain the source identities and well-generalised to various identities.

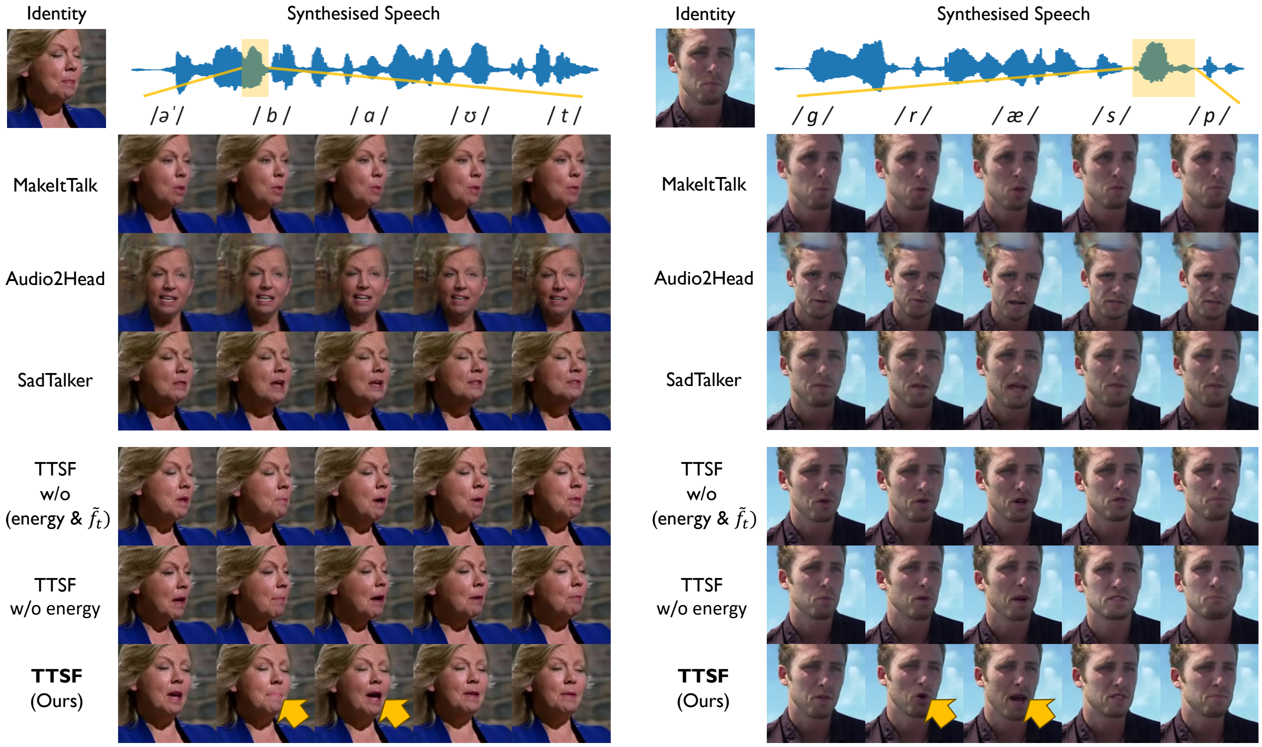

Figure 3. Qualitative results on LRS2 dataset. Our approach outperforms all the baselines in terms of generating natural facial motions, encompassing lip shape and head pose.

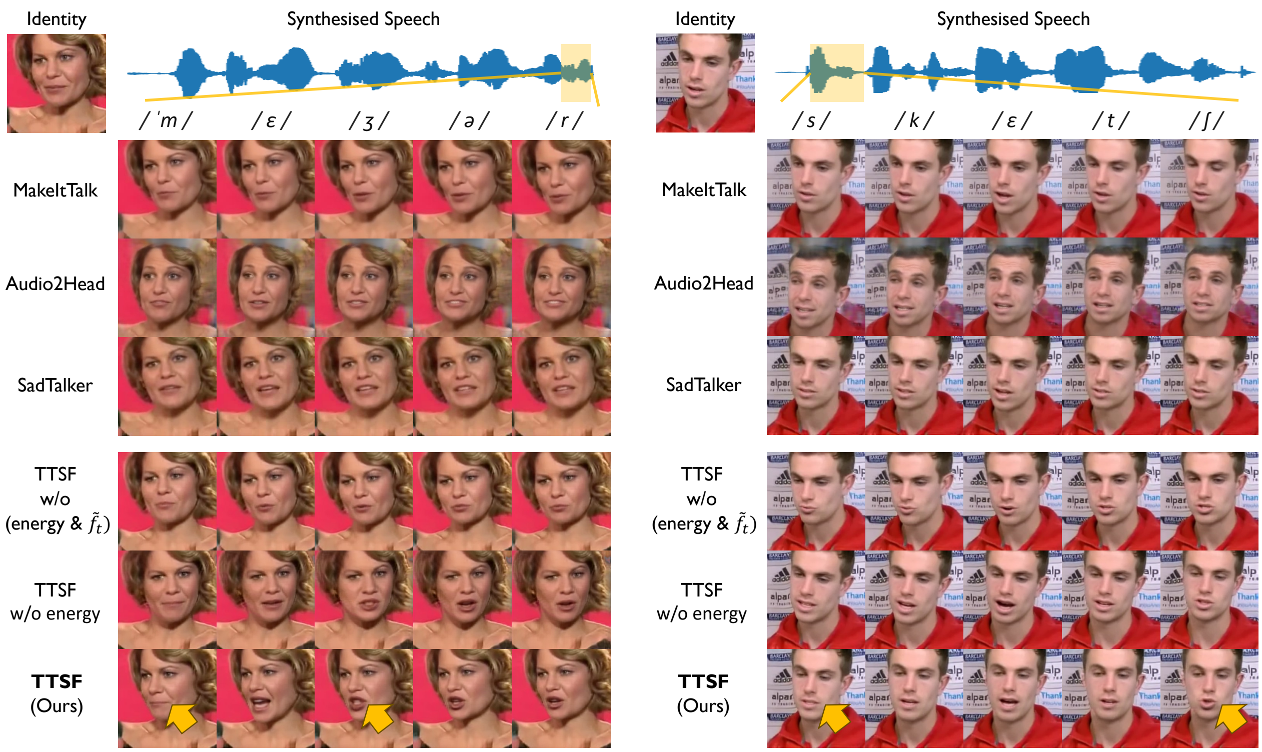

Figure 4. Qualitative results on VoxCeleb2 dataset. Our approach outperforms all the baselines in terms of generating natural facial motions, encompassing lip shape and head pose.

G. Guideline for User Study

For a comprehensive understanding of the process of user study, we offer the guidelines used for user studies.

Text-driven Talking Face Generation. The questions are as follows:

- (Lip Sync Quality) How well synchronised between lip and audio?
- (Head Movement Naturalness) How naturally does the head move?
- (Overall Quality) How realistic is the video?

Participants were guided to assign a rating of 5 to the highest-quality video and a rating of 1 to the lowest-quality video for score calibration.

Possible reasons for the poor scores of the baseline models in the MOS test include:

- As MakeItTalk relies on extracting facial landmarks, it struggles when exact landmarks cannot be discerned, resulting in poor lip synchronisation and head movements.
- Audio2Head exhibits challenges in preserving identity, particularly when the source face is not centered in the region.
- SadTalker exhibits artifacts at the boundary of the cropped image and limited facial movement, hindering the generation of realistic talking faces.

In contrast, our framework maintains the identity of the source image while incorporating fine details, surpassing baseline models in overall quality.

Text-to-Speech. The questions for evaluating TTS systems are as follows:

- (Naturalness) How close is the audio source to real speech in its quality?
- (Voice similarity) How similar is the voice in the audio to the original voice?

Participants were asked to rate naturalness and similarity on a scale from 1 to 5, with 1 representing the lowest quality and 5 representing the highest quality.

We assume that the performance degradation of the baselines is due to the direct utilisation of visual features from the source image, which still retains motion components, leading to inconsistency with the target voice. On the other hand, our framework synthesises high-quality speech robust to diverse facial motions by utilising motion-removed features, namely identity features f_{id}, obtained from the TFG system.

H. Detail Explanations of Dataset Split

We use *trainval* split of LRS3 dataset for training and evaluate our method on VoxCeleb2 and LRS2 datasets. For robust training, we exclude video samples shorter than 1.3 seconds and longer than 7 seconds. Additionally, we remove speakers who have less than 14 seconds of total video. Our training dataset consists of approximately 21 hours of video with 20,337 samples and 1,687 speakers. For test sets, we sample random transcriptions from LRS2, and select 300 random speakers from each of the LRS2 and VoxCeleb2 datasets. Note that there is no overlap among the sampled speakers.

I. Potential Biases in the Generation Process

Our model establishes voice characteristics by leveraging facial features, meaning that when input images share similar facial attributes, our model generates similar voices. For instance, individuals with longer hair, often associated with females, statistically lead our model to produce a voice with a relatively higher pitch. This inductive bias is derived from the training dataset.

J. Ethical Statements

The talking face generation model has also raised concerns about the potential misuse of deepfakes and manipulated media. The misuse could have severe consequences, including spreading false information and causing harm to individuals and communities.

To address these concerns, we will limit the usage of our model and provide access only to trusted communities such as those working on technologies beneficial to society. Additionally, steps must be taken to ensure that the technology is used ethically and responsibly. This includes educating users on the potential risks and providing clear guidelines.