Talking Face Generation with Multilingual TTS

Hyoung-Kyu Song[*12], Sang Hoon Woo[*1], Junhyeok Lee[1], Seungmin Yang[12],
Hyunjae Cho[13], Youseong Lee[13], Dongho Choi[3], Kang-wook Kim[3]
[1]MINDsLab Inc., South Korea
[2]KAIST, South Korea [3]Seoul National University, South Korea
{hksong35, shwoo, jun3518, myaowng2, chohyunjae, staru}@mindslab.ai
{dongho.choi, full324}@snu.ac.kr

Abstract

Recent studies in talking face generation have focused on building a model that can generalize from any source speech to any target identity. A number of works have already claimed this functionality and have added that their models will also generalize to any language. However, we show, using languages from different language families, that these models do not translate well when the training language and the testing language are sufficiently different. We reduce the scope of the problem to building a language-robust talking face generation system on seen identities, i.e.*, the target identity is the same as the training identity. In this work, we introduce a talking face generation system that generalizes to different languages. We evaluate the efficacy of our system using a multilingual text-to-speech system. We present the joint text-to-speech system and the talking face generation system as a neural dubber system. Our demo is available at https://bit.ly/ml-face-generation-cvpr22-demo. Also, our screencast is uploaded at https://youtu.be/F6h0s0M4vBI.*

1. Introduction

Talking face generation, a task of synthesizing a face video where the lip is synchronized with the input speech, is one of the most popular research topics in neural video generation. When combined with a text-to-speech (TTS) system, the joint system allows users to create a talking video with only a text input and has potential applications in news broadcasting, virtual lectures, and digital concierge. Expanding the task to support multiple languages would significantly reduce the amount of effort required to widen the target audience to the global population.

Recent works in talking face generation claim that their models support input speeches in any language [8, 11].

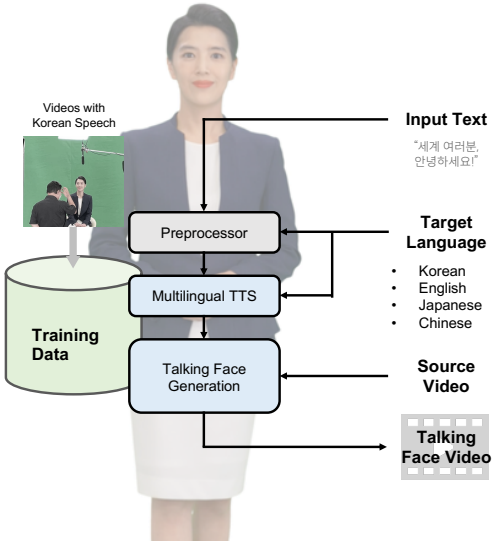


Figure 1. An overview of the demonstration. With the text, the language, and the source video, users can make face videos which include the voice corresponding to text. For training data, we record two hours of footage of the Korean speaker.

However, we observe that such models fail to generalize to certain input speech languages, *e.g.*, Korean. We hypothesize that the robustness of these models depends on the degree of similarity between the training speech language and the input speech language. Thus, we will validate the generalization capabilities of multilingual face generation models using speeches of languages from different language families.

For practical applications of multilingual talking face generation, the speaker's vocal identity should be preserved across different languages. Since multilingual speech datasets for desired speakers are often unavailable, the multilingual talking face generation system requires a multilin-

[*] indicates equal contribution.

gual TTS model capable of cross-lingual speech synthesis. While a number of prior works in multilingual TTS discuss cross-lingual speech synthesis, the selection of languages has been underexplored. Thus, existing works' ability to perform cross-lingual synthesis across languages from different language families is questionable.

In this work, we propose a multilingual talking face generation system shown in Fig. 1. We also describe the two models used in the multilingual TTS module and the talking face generation module: a multilingual adaptation of VITS [6] capable of performing cross-lingual speech synthesis while preserving the speaker's vocal identity, and a talking face generation model capable of generating face videos from synthesized speeches, regardless of the language.

Our contributions in this work are the following:

- We introduce a system that can synthesize talking face videos in four languages (Korean, English, Japanese, and Chinese) for a monolingual speaker.
- We build a talking face generation model that is robust to different input speech languages.
- Our demonstration can generate 512×512 facial image sequences faster than 25 fps.

2. Related Work

2.1. Text-To-Speech (TTS)

Traditionally, text-to-speech systems have employed a two-stage pipeline, with the models for each stage developed independently from each other. The first stage uses an acoustic model [13, 14] to generate an intermediate speech representation, namely mel-spectrogram, based on the input text. In the second stage, a vocoder model [7, 12] converts the speech representation to a raw waveform. There have been a number of attempts [1, 13] to reduce text-to-speech systems to a fully end-to-end process, but they required complex input conditions [13] or were slow [1]. Recently, Kim *et al*. [6] proposed VITS, a non-autoregressive end-to-end architecture that surpasses state-of-the-art two-stage models.

Among speech synthesis related studies, some works have focused on multilingual text-to-speech models with cross-lingual capabilities. Zhang *et al*. [21] first proposed the use of domain adversarial training in multilingual text-to-speech training to mitigate the speaker dependency in text representations. However, the quality of the generated speech was highly dependent on the source speaker's language and the target speech language. Maiti *et al*. [9] utilized bilingual speaker data to compute the modification between different languages in the speaker embedding space. No prior works explored cross-lingual speech synthesis with languages from different language families.

2.2. Talking Face Generation

Recent works in talking face generation have focused on generalizing their models to any vocal and visual identity *i.e*. from any input speech to any target face. Among such works, some have successfully built models that generate videos with realistic faces, using a GAN-based approach [11, 18]. Vougioukas *et al*. [18] proposed temporal GANs that include recurrent layers in the architecture for temporal consistency. Prajwal *et al*. [11] suggested the use of modified SyncNet [2] to determine whether the generated images and the corresponding audio are in sync. Both networks described above have low resolution outputs; Vougioukas *et al*. [18] outputs 96×96 images and Prajwal *et al*. [11] outputs 96×128 images. The low resolution of the output images also restricts the maximum resolution of the final output video.

Prior talking face generation studies claim that their systems are training language agnostic, *i.e*., their models can generate videos with speech from any language, regardless of the language of the data the models were trained on. While such claims do hold for some languages, the quality of the outputs tends to degrade significantly when tested on a language from a different family tree than the training language. This phenomenon can be observed in the official interactive demo of [11]. Such cases often require re-training of the model with data in the desired language.

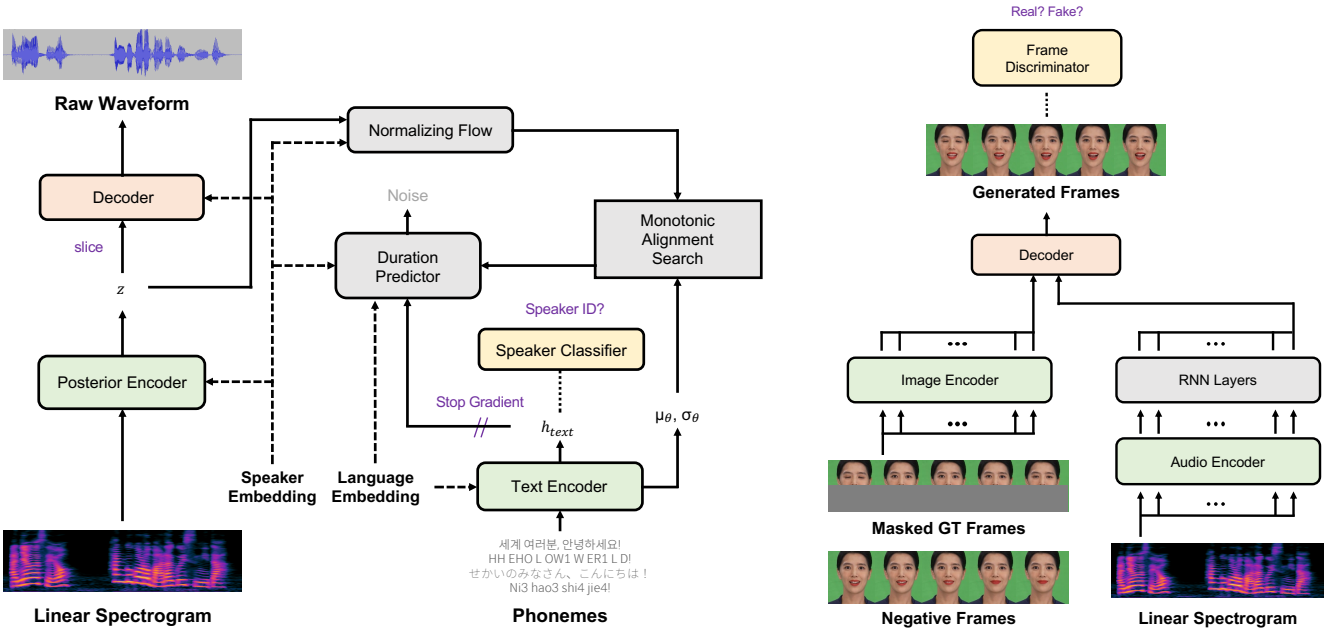
3. Face Generation with Multilingual TTS

3.1. Overall Architecture

The proposed system consists of three main modules: a preprocessor module, a multilingual TTS module, and a talking face generation module. First, the preprocessor module transforms the input text to a phoneme sequence. The multilingual TTS module, in turn, generates a raw speech waveform based on the phoneme sequence with the designated speaker identity. Subsequently, the talking face generation module synthesizes the final output video with the lip movement synchronized to the input speech.

3.2. Preprocessor

While the details of the preprocessing step differ for each language, the common objective is to convert raw text into a sequence of phonemes. The preprocessor first cleans the text by removing any character or symbol that does not belong to the specified language. Then the text goes through a normalization procedure, converting non-verbal texts, *e.g*., digits, dates, and short forms of words, to their verbalized forms. Finally, the preprocessor maps all graphemes in the text to phonemes. We use different phoneme sets for each language in our setup; the phoneme sets for Korean, English, Japanese, and Chinese are Hangul, ARPAbet, Hiragana, and Pinyin, respectively. For Korean and



(a) Training architecture of multilingual TTS

(b) Training architecture of talking face generation

Figure 2. Training pipeline of the multilingual TTS model and the talking face generation. In (a), the multilingual TTS model is trained to synthesize a raw waveform conditioned from language embedding. In (b), the talking face generation model is trained with facial videos with Korean speech from target speaker. For the inference, the TTS system can generate audio with the phoneme sequence from a number of languages.

English, we utilize our in-house grapheme-to-phoneme algorithms. For Japanese and Chinese grapheme-to-phoneme conversion, we use open source libraries SudachiPy[1] and pypinyin[2]. The preprocessor may employ an optional language translation system for downstream applications like a neural dubber.

3.3. Multilingual TTS

We use multi-speaker VITS [6] as the base model for the multilingual TTS module. To enable multilingual speech synthesis for VITS, we add embeddings for each language and input them to submodules. Our preliminary experiments showed that injecting the language embedding to the text encoder and the duration predictor yields the best result. We will refer to this setting as the multilingual VITS.

3.4. Talking Face Generation

For talking face generation, we use our internal talking face generation model with output image resolution of 512×512. Our face generation model comprises three components: an image encoder, an audio encoder, and a decoder. The image encoder and the audio encoder extract features from the input image and the audio, respectively. We concatenate the extracted features from both encoders and feed them to the decoder, which then generates lip-synchronized face images. All three components of our module are primarily CNN-based models. The audio encoder that employs additional RNN layers to maintain temporal consistency. Thus, our module can generate image sequences with minimal autoregressive computation, reducing the network latency.

3.5. Dataset Collection

For the training data of our target identity, we record two hours of footage of the target speaker speaking in Korean. The recorded video and audio data are resampled to 25 fps and 22050 Hz, respectively.

We only use the facial regions from the image frames as the input to the talking face generation model. To extract the facial region from a image region, we first estimate facial landmarks with a pre-trained model provided by [4]. A prior study [18] cropped the facial regions tightly around the detected faces. However, we find that if the entire head is not included in the cropped region, the final output image contains borderline discontinuity. To mitigate this issue, we expand the facial regions to include the entire head with extra margins.

For the multilingual VITS training data, we use a mix of our internal datasets and open source datasets, in addition

[1] https://github.com/WorksApplications/SudachiPy
[2] https://github.com/mozillazg/python-pinyin

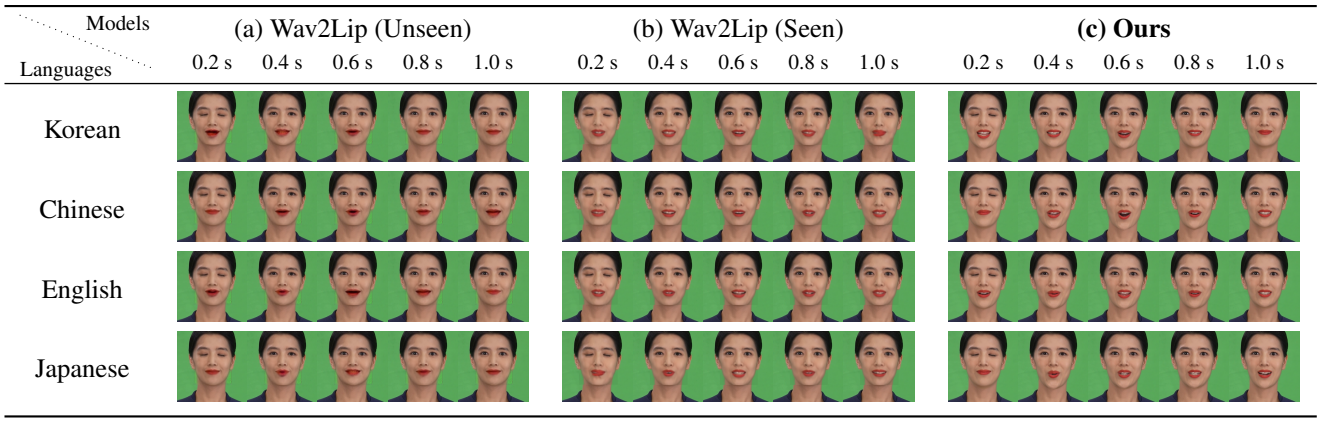


Figure 3. Comparison of the output image sequences. For (a), we use the official interactive demo of Wav2Lip. In (b), the model was trained with the same target identity dataset as (c). All output images were cropped using the same method as our system's facial region cropping. The image sequences for each model-language pair were taken every 0.2 seconds over 1.0 second. Note that all models used the same input speech and the source video.

to the target speech data. Our internal datasets comprise 28 hours of Korean speech from 25 speakers and 13 hours of English speech from 13 speakers. We also include several open source text-to-speech datasets as a part of our train set: the Korean Single Speaker Speech dataset [10] for Korean, the LJ Speech dataset [5] and the LibriTTS dataset [20] for English, the Voice Actress corpus[3], the JSUT dataset [16], and the JVS dataset [17] for Japanese, and the AISHELL-3 dataset [15] for Chinese. Overall, our train set includes speech from 472 speakers totaling 206 hours.

3.6. Training Details

The training procedure for the multilingual VITS remains largely the same as the original VITS training [6]. We employ a couple of techniques on top of the original VITS training to improve the quality of multilingual speech synthesis. First, we apply domain adversarial training [3] to minimize the speaker information leakage to the encoded text representations. Following Zhang *et al*. [21], we add a speaker classifier with a gradient reversal layer after the text encoder. For the classification loss scale factor λ, we follow the schedule from [3].

During the initial experiments, we observed that feeding the speaker embedding directly to the duration predictor degrades the quality of the output speech. We hypothesize that the unseen combination of language embedding and speaker embedding introduces instability. To resolve this issue, we add a regularization term to the loss, such that the mean of all speaker embeddings are pushed towards the zero vector. During inference, if the speaker's original language does not match the input language, we use the zero vector in place of the speaker embedding, similarly to the use of the zero vector as the residual encoding in Zhang *et al*. [21].

[3] http://voice-statistics.github.io/

In training our talking face generation model, we follow Prajwal *et al*. [11] and use masked ground truth frames and negative frames, *i.e*., frames from different part of the video. In addition, we apply various augmentations, *e.g*., translation, rotation, zoom in/out to the facial region, instead of normalizing the facial regions to minimize the spatial variance. This drives the generator to be more robust to the rotation of the neck or the size ratio of the face to the image when generating the face. We also adopt adversarial training with multi-scale discriminator [19] as the auxiliary training to improve the perceptual visual quality.

4. Experiments

In Fig. 3, we compare the outputs from our model to the outputs from two versions of the Wav2Lip model. For both versions of Wav2Lip, we observe a number of artifacts in the output images that degrade the overall image quality. First, the borderlines of the bounding box are clearly visible near the chin in the output images. Also, the generated images from both models are fuzzy and lack details compared to the outputs from our model. In Wav2Lip (Unseen) outputs, we observe that the inside of the mouth is filled with black in most frames, erasing out details including the teeth. Wav2Lip (Seen) outputs display different artifacts; when the lips are closed, the border between the upper and the lower lips disappears. We believe that the use of adversarial training forces our model to generate more fine-grained details, resulting in higher quality image overall.

We also report the characteristics of the model outputs for out-of-training-language speeches. The outputs from Wav2Lip (Unseen) show abrupt lip movements for non-English speeches, characterized by sudden closing of the mouth. On the other hand, the outputs from Wav2Lip (Seen) exhibit a very different behavior for out-of-training-

Stage			Time (s)	fps
TTS	(a)	Total	0.0655	-
Talking Face Generation	(b.1)	Face Generation	0.3964	63.1
	(b.2)	Video Encoding	0.2981	-
	(b)	Total	0.6945	36.0
	(a) + (b.1)		**0.4619**	**54.1**
	(a) + (b)		**0.7600**	**32.9**
Overall System	(c)	Total	**0.8251**	**30.3**

Table 1. Throughputs for each stage in overall system. Time values in the second column are measured for generating one second audio or video. Note that the original video is rendered with 25 fps.

language speeches; in Wav2Lip (Seen) output images, the lip movements are minimal compared to the outputs from the other two models. Our model, on the contrary, produces a more gradual and diverse lip movements regardless of the input speech language. This shows that our training method builds a model more robust to different input speeches.

After generating the facial image sequences, the images are merged into the source FHD video before the end users receive them. Specifically, we replace the facial regions of the source FHD video frames with the generated face image sequences. We measure the end-to-end latency of the system as the duration between the user request and the service response containing the final video. We deploy our system on a desktop with AMD Ryzen 7 3800XT and Nvidia GeForce RTX 3080 and measure the speed of our system. As shown in Table 1, the speed of the entire system is faster than real-time. Note that the video encoding time is measured for reference videos that are loaded into the system beforehand; if the system were to support custom reference videos, *i.e.*, user-inputted reference videos, the system will require additional preprocessing time.

In this demo, we choose MP4, a universal file format for videos, as the output format so that the output videos can be played on any device. However, the MP4 format does not allow streaming until the entire video is generated, leading to longer latency for the end user. Switching to a more streaming-friendly format, such as MKV, would significantly reduce the end-to-end latency and improve its applicability.

5. Broader Impact

Unlike existing works on talking face generation, the objective of our system is not to support inference on unseen identities. Instead, we focus on generating high resolution talking face videos where the target identity is seen during training. The proposed system can facilitate the production of video-based media such as virtual newscast or online tutoring. Combined with a language translation system, our system allows users to generate four versions of a video, each in a different language, significantly boosting the accessibility of the content.

While the proposed technology does come with benefits, we acknowledge that it can also be used with malicious intents. Since the system can generate video based on any text, an adversarial user may attempt to create deepfakes with harmful contents. Regarding such vulnerabilities, we first note that the data required to train the proposed system would likely be unobtainable without an agreement with the target identity, substantially reducing the entities that can train the system. The system may also employ a content filter, *e.g.*, a hate speech filter to further reduce the risk of generating malicious contents. Also, we strongly believe that with proper accountability management involving tracking the use of every generated video, the system can be safely used with minimal liability.

6. Conclusion

In this work, we present a robust talking face generation system compatible with multilingual speech from a speech synthesis model. We describe a talking face generation model robust to input speech language, as well as techniques to equip a state-of-the-art TTS model with multilingual synthesis capabilities. By combining the face generation model and the TTS model, we build a system that can generate talking face videos in four languages without a multilingual parallel dataset. We demonstrate our system's ability to generalize across languages by evaluating with languages from different language families. We also show that our system is feasible in industrial settings by deploying our demo on a desktop with no external computational resources. We hope that our system can help content creators improve the accessibility of their works past language barriers.

7. Acknowledgement

This work was supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSTI) (No. 2021-0-00062-002, Development of Joint Work Automation Management Software Technology Based on Task Awareness).

References

[1] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. WaveGrad 2: Iterative Refinement for Text-to-Speech Synthesis. In *Proceedings of the Interspeech*, pages 3765–3769, 2021. 2

[2] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian Conference on Computer Vision*, pages 251–263. Springer, 2016. 2

[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(1):2096–2030, 2016. 4

[4] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *European Conference on Computer Vision*, pages 152–168. Springer, 2020. 3

[5] Keith Ito and Linda Johnson. The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/, 2017. 4

[6] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021. 2, 3, 4

[7] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[8] Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. LipSync3D: Data-Efficient Learning of Personalized 3D Talking Faces from Video using Pose and Lighting Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2755–2764, 2021. 1

[9] Soumi Maiti, Erik Marchi, and Alistair Conkie. Generating Multilingual Voices Using Speaker Space Translation Based on Bilingual Speaker Data. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7624–7628, 2020. 2

[10] Kyubyong Park. KSS Dataset: Korean Single speaker Speech Dataset. https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset. 4

[11] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 1, 2, 4

[12] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A Flow-based Generative Network for Speech Synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621, 2019. 2

[13] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *Proceedings of the International Conference on Learning Representations*, 2021. 2

[14] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018. 2

[15] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Proceedings of the Interspeech*, pages 2756–2760, 2021. 4

[16] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017. 4

[17] Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito, Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari. JVS corpus: free Japanese multi-speaker voice corpus. *arXiv preprint arXiv:1908.06248*, 2019. 4

[18] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs. *arXiv preprint arXiv:1805.09313*, 2018. 2, 3

[19] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4

[20] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proceedings of the Interspeech*, 2019. 4

[21] Yu Zhang, Ron J. Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, R.J. Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran. Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. In *Proceedings of the Interspeech*, pages 2080–2084, 2019. 2, 4