FCL-TACO2: TOWARDS FAST, CONTROLLABLE AND LIGHTWEIGHT TEXT-TO-SPEECH SYNTHESIS

Disong Wang[1†], *Liqun Deng*[2], *Yang Zhang*[2], *Nianzu Zheng*[2], *Yu Ting Yeung*[2], *Xiao Chen*[2], *Xunying Liu*[1], *Helen Meng*[1]

[1]Human-Computer Communications Laboratory,
The Chinese University of Hong Kong, Hong Kong SAR, China
[2]Huawei Noah's Ark Lab

{dswang, xyliu, hmmeng}@se.cuhk.edu.hk {dengliqun.deng, zhangyang86, zhengnianzu, yeung.yu.ting, chen.xiao2}@huawei.com

ABSTRACT

Sequence-to-sequence (seq2seq) learning has greatly improved text-to-speech (TTS) synthesis performance, but effective implementation on resource-restricted devices remains challenging as seq2seq models are usually computationally expensive and memory intensive. To achieve fast inference speed and small model size while maintain high-quality speech, we propose FCL-taco2, a Fast, Controllable and Lightweight (FCL) TTS model based on Tacotron2. FCL-taco2 adopts a novel semi-autoregressive (SAR) mode for phoneme level based parallel mel-spectrograms generation conditioned on prosody features, leading to faster inference speed and higher prosody controllability than Tacotron2. Besides, knowledge distillation (KD) is leveraged to compress a relatively large FCL-taco2 model to its small version with minor loss of speech quality. Experimental results on English (EN) and Chinese (CN) datasets show that the small version of FCL-taco2 achieves comparable performance with Tacotron2 in terms of speech quality, while it has a **4.8× smaller** footprint with **17.7×** and **18.5× faster** inference speeds on average for EN and CN experiments respectively. Besides, execution on mobile devices shows that the proposed model can achieve faster than real-time speech synthesis. Our code and audio samples are released[1].

Index Terms— Text-to-speech, controllable and efficient, semi-autoregressive, prosody modelling, knowledge distillation

1. INTRODUCTION

As a key component of speech communication technologies, TTS has been widely applied in human-computer interaction including virtual assistants and chatbots [1]. Complex TTS systems are usually deployed in the cloud, as they have high computational requirements, and the generated speech is sent to users. However, to preserve privacy and reduce latency, it is necessary to move TTS models to edge devices, e.g., mobile phones, which require that the TTS models have low latencies and small footprints, but still maintain a high quality in the generated speech outputs. Current state-of-the-art TTS systems based on neural frameworks typically include two important parts [2], i.e., an acoustic model that converts the text inputs (e.g., phonemes) to acoustic features (e.g., mel-spectrograms), and a vocoder that synthesizes the waveform from acoustic features. Previous work mostly focused on the optimization of vocoder for edge deployment [3-5], while less attention has been paid to the acoustic model [6].

To obtain an efficient acoustic model, we propose FCL-taco2 that is a semi-autoregressive (SAR) version of Tacotron2 [7] with explicit prosody modelling. FCL-taco2 improves Tacotron2 in three perspectives. First, FCL-taco2 predicts mel-spectrograms in SAR mode, i.e., mel-spectrograms are generated in autoregressive (AR) mode for individual phoneme and non-autoregressive (NAR) mode for different phonemes. On one hand, AR mode within each phoneme is beneficial for the model to utilize previous speech information for stable and accurate mel-spectrograms generation. One the other hand, NAR mode enables the parallel generation at the phoneme level, which greatly accelerates inference speed. Second, a prosody injector is proposed to infer phoneme duration, pitch and energy features as inputs of FCL-taco2 to achieve flexible prosody control. Third, knowledge distillation (KD) [8] is exploited to transfer the knowledge from a teacher FCL-taco2 to the student model with smaller size. In this way, inference speed is further accelerated, while high-quality speech is maintained.

The main contributions of this paper are: (1) The proposed novel SAR approach can achieve good balance between generated speech quality and inference speed; (2) The incorporation of duration, pitch and energy modelling brings more flexible prosody control than Tacotron2; (3) Three effective KD strategies are proposed for model compression with minor loss of speech quality.

2. RELATED WORK

Advances in deep learning have enabled TTS to achieve remarkable progress via sequence-to-sequence (seq2seq) acoustic models [7, 9-11], where the attention mechanism is used to align the text and speech automatically to achieve satisfactory synthesis results. However, due to the AR generation mode, the seq2seq based models suffer from slow inference speed, which limits their applications in real-time services. On the contrary, NAR models [12-21] achieve extremely fast inference speed, where a length regulator is used to replace the attention module in AR models to predict phoneme durations and expand the phoneme embeddings for parallel generation. Nonetheless, to achieve comparable speech quality with AR counterparts, NAR models usually require much more parameters [14], which may not be supported by edge

† Work done during internship at Huawei Noah's Ark Lab
[1] Code and audio samples: https://github.com/wendison/FCL-taco2

devices with limited storage. To tackle this problem, [21] proposed a lightweight NAR model based on the convolutional network and adopts the same training scheme as FastSpeech, which suffers from inaccurate duration extraction and spectral information loss. FastSpeech2 [13] alleviates these issues by using forced alignment [22] based accurate phoneme durations and pitch/energy features as conditions to bridge the gap between the text and speech. This work also utilizes a similar component, i.e., prosody injector, to achieve this goal. [23] proposed a phrase-based parallel Tacotron2, which requires a complicated text processor to obtain phrases. In contrast, the proposed FCL-taco2 requires no syntactic analysis and achieves a much faster inference speed, as generation runs at the level of the phoneme, a much smaller unit than the phrase.

2. TACOTRON2

Since FCL-taco2 is based on the architecture of Tacotron2, we first present an overview of Tacotron2. Tacotron2 contains an encoder and a decoder with location-sensitive attention [24]. The encoder consists of an embedding layer, a stack of three convolutional (Conv) layers [25] and a bi-directional LSTM (BLSTM) layer [26]. The decoder contains two-layer Pre-net, two uni-directional LSTM layers, two fully connected (FC) layers for mel-spectrograms and a stop token prediction. In addition, a Post-net with 5 Conv layers is used for mel-spectrograms refinement. During inference, the phoneme sequence is first converted by the encoder into robust hidden representations, which are then consumed by the decoder to predict mel-spectrograms frame by frame in AR mode, which leads to slow inference speed.

3. PROPOSED METHOD

In this section, we first elaborate on the architecture of the proposed FCL-taco2, then present a two-stage learning scheme for obtaining a smaller and faster version of the teacher FCL-taco2 that has a similar footprint as Tacotron2.

3.1. Model architecture

As shown in Fig. 1, FCL-taco2 includes three key components: the encoder, prosody injector and decoder.

The encoder of FCL-taco2 has a structure similar to that of Tacotron2. It takes in the phoneme sequence $\mathbf{X}=\{x_n\}_{1\le n\le N}$ with length N to produce a sequence of hidden representations:

$$\mathbf{H} = \text{Encoder}(\mathbf{X}) \tag{1}$$

$\mathbf{H}=\{\mathbf{h}_n\}_{1\le n\le N}$ contains rich context information of phonemes in the given text and is beneficial for inferring prosody features.

The prosody injector contains three prosody predictors, i.e., duration, pitch and energy predictors. They all have the same structure (2 Conv + 1 FC Layer) and take $\mathbf{H}$ as inputs to predict phoneme-level durations $\mathbf{D}=\{d_n\}_{1\le n\le N}$, pitch values (described by fundamental frequency F_0) $\mathbf{F}=\{f_n\}_{1\le n\le N}$ and energy values $\mathbf{E}=\{e_n\}_{1\le n\le N}$ as:

$$\mathbf{D} = \text{Duration-Predictor}(\mathbf{H}) \tag{2}$$

$$\mathbf{F} = \text{Pitch-Predictor}(\mathbf{H}) \tag{3}$$

$$\mathbf{E} = \text{Energy-Predictor}(\mathbf{H}) \tag{4}$$

Then F_0 and energy values are projected to match the dimension of $\mathbf{H}$ and injected into $\mathbf{H}$ as follows:

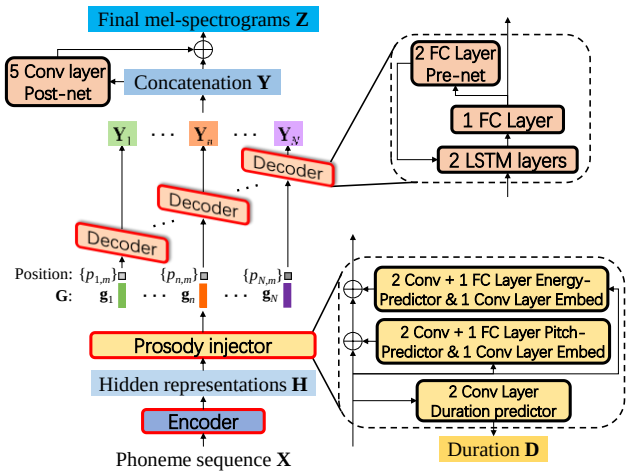

Fig. 1. Block diagram of the proposed FCL-taco2, where the decoder generates mel-spectrograms in AR mode within each phoneme and is shared for all phonemes.

$$\mathbf{G} = \mathbf{H} + \text{Pitch-Embed}(\mathbf{F}) + \text{Energy-Embed}(\mathbf{E}) \tag{5}$$

The decoder takes $\mathbf{G}=\{\mathbf{g}_n\}_{1\le n\le N}$ to generate mel-spectrograms in SAR mode. Specifically, the mel-spectrograms $\mathbf{Y}_n=\{\mathbf{y}_{n,m}\}_{1\le m\le d_n}$ corresponding to n^{th} phoneme are generated in AR mode:

$$\mathbf{y}_{n,m} = \text{Decoder}(\mathbf{g}_n, p_{n,m}, \mathbf{y}_{n,m-1}) \tag{6}$$

i.e., the m^{th} frame of $\mathbf{Y}_n$ is conditioned on $\mathbf{g}_n$, $p_{n,m}$ and the previous frame $\mathbf{y}_{n,m-1}$, where $p_{n,m}$ is a value (0~1) that denotes the relative position of every frame inside each phoneme and is appended to $\mathbf{g}_n$, which facilitates the robust generation. The generation length of $\mathbf{Y}_n$ is determined by the predicted duration d_n of n^{th} phoneme. We note that the decoder is shared for all phonemes, as a result, the mel-spectrograms of different phonemes can be generated in parallel, which greatly speeds up the generation. The mel-spectrograms of different phonemes are concatenated as $\mathbf{Y}=\{\mathbf{Y}_n\}_{1\le n\le N}$, which can be further refined by Post-net to obtain the final output mel-spectrograms $\mathbf{Z}$ as:

$$\mathbf{Z} = \mathbf{Y} + \text{Post-net}(\mathbf{Y}) \tag{7}$$

3.2. Two-stage learning

The two-stage learning scheme includes: (1) First training a teacher FCL-taco2 (FCL-taco2-T) of similar size as Tacotron2; and then (2) KD is performed to compress FCL-taco2-T to obtain the student model with significantly fewer parameters.

3.2.1. FCL-taco2-T training

FCL-taco2-T has the identical encoder, decoder and Post-net design as Tacotron2 without the attention module and stop token prediction. During training, the prosody predictors are jointly trained with the remaining modules of FCL-tcao2-T. Therefore, the total loss for training the FCL-taco2-T is:

$$L^{GT} = \lambda_1 L_m^{GT} + \lambda_2 L_d^{GT} + \lambda_3 L_f^{GT} + \lambda_4 L_e^{GT} \tag{8}$$

where L_m^{GT} is a combination of L1- and L2-norm between the ground-truth (GT) and predicted mel-spectrogram, L_d^{GT}, L_f^{GT} and L_e^{GT} are L2-norms between the GT and predicted values for duration, F_0 and energy, respectively. λ_j ($1\le j\le 4$) are constant weights that are all set to 1 empirically. The GT phoneme durations are extracted

from pairs of text and speech by Montreal Forced Aligner (MFA) [22]. F_0 and energy are also phoneme-level values that are averaged over each phoneme using the extracted durations.

3.2.2. Knowledge distillation

KD was proposed in the teacher-student framework [22] that aims to transfer the knowledge from a large teacher model to a small student model imitating the behaviors of the teacher. As FCL-taco2-T is still relatively large, KD can be used to compress it to remove the redundancy and fit into edge devices. In this paper, the student FCL-taco2 (FCL-taco2-S) has the same number of layers as the teacher, but each layer has a smaller size. To maintain high-quality speech, we propose three effective distillation strategies, i.e., mel-spectrogram distillation (MSD), hidden representation distillation (HRD) and prosody distillation (PD).

Mel-spectrogram distillation: MSD is the sequence-level KD [12] that forces the student to learn to output the similar mel-spectrograms as those generated by the teacher. Therefore, the MSD loss L_{MSD} is the difference (i.e., L1- and L2-norm) between the predicted mel-spectrograms of the teacher and student.

Hidden representation distillation: HRD aims to distill the hidden representations of the teacher to the student, which has been proved to be effective in image and natural language processing tasks [27, 28]. In the proposed FCL-taco2 model, the hidden representations include the layers' outputs of the encoder (i.e., Embedding layer, Conv layers and hidden states of BLSTM layer), decoder (i.e., FC layers in Pre-net and hidden states of LSTM) and Post-net layers, then HRD loss is defined as:

$$L_{HRD} = \sum_{i \in I} \left\| \mathbf{K}_S^i \mathbf{W}^i - \mathbf{K}_T^i \right\|_2 \quad (9)$$

where $\| \cdot \|_2$ is L2-norm, I is the set containing the indices of hidden representations, $\mathbf{K}_S^i \in \mathbb{R}^{l \times p}$ and $\mathbf{K}_T^i \in \mathbb{R}^{l \times q}$ are the i^{th} hidden representations of student and teacher respectively, p and q denote the layer dimension and $p \le q$, $\mathbf{W}^i \in \mathbb{R}^{p \times q}$ is a learnable matrix to project the hidden layer of student to match the dimension of teacher.

Prosody distillation: PD is adopted to distill the prosody predictions and embeddings of the teacher to the student, and the PD loss is defined as:

$$L_{PD} = L_d^{ST} + L_f^{ST} + L_e^{ST} + \left\| \mathbf{K}_S^f \mathbf{W}^f - \mathbf{K}_T^f \right\|_2 + \left\| \mathbf{K}_S^e \mathbf{W}^e - \mathbf{K}_T^e \right\|_2 \quad (10)$$

where L_d^{ST}, L_f^{ST} and L_e^{ST} are the L2-norms between the predicted values of student and teacher for duration, F_0 and energy, respectively. $\mathbf{K}_*^f$ and $\mathbf{K}_*^e$ are the F_0 and energy embeddings of student or teacher, $\mathbf{W}^f$ and $\mathbf{W}^e$ are learnable projection matrices.

During the training of student model, both GT acoustic features and knowledge distilled from the teacher are used to guide the student learning. Therefore, the training loss is:

$$L_{student} = \alpha_1 L^{GT} + \alpha_2 L_{MSD} + \alpha_3 L_{HRD} + \alpha_4 L_{PD} \quad (11)$$

where α_j ($1 \le j \le 4$) are all set to 1 in our experiments.

4. EXPERIMENTS AND ANALYSIS

4.1. Experimental settings

To verify the effectiveness of proposed methods, we conduct experiments on English (EN) and Chinese (CN) datasets. The EN dataset is the publicly available LJSpeech [29] with around 24 hours of a single female speech recorded at 22.05kHz. The CN dataset is internal with a child voice recorded for 15 hours at 16kHz. Each dataset is randomly split into training (95%), validation (5%) and testing (5%) data. The text is converted by MFA tool [22] into phoneme and pinyin sequences with tones, which are adopted as the inputs of TTS models for EN and CN experiments, respectively. The target mel-spectrogram is set to 80-band and calculated with FFT size of 1024 and hop length of 256.

Table 1. Hyperparameters of the TTS models, where a single value refers to the dimension of the hidden layer, $a \times b$ means a layers with b units per layer, (a, c, d) means a 1-D convolutional layers with c filters and d kernel size per layer.

Models		Tacotron2	FCL-taco2-T	FCL-taco2-S
Encoder	Embed	512	512	256
	Conv	(3,512,5)	(3,512,5)	(3,256,5)
	BLSTM	1×512	1×512	1×256
Decoder	Attention	128	/	/
	Pre-net	2×256	2×256	2×256
	LSTM	2×1024	2×1024	2×256
Post-net		(5,512,5)	(5,512,5)	(5,128,5)
Prosody predictor		/	(2,384,3)+384	(2,384,3)+384
Pitch/Energy-Embed		/	(1,512,9)	(1,256,9)

All acoustic models are trained by Espnet toolkit [30], and four acoustic models are compared, i.e., Tacotron2, FCL-taco2-T, FCL-taco2-S and FastSpeech2, hyperparameters of the former three models are listed in Table 1. FastSpeech2 is an improved version of FastSpeech, and it follows architecture settings of Espnet scripts. Compared with Tacotron2, FCL-taco2-T removes the attention module and has additional prosody predictors plus Pitch/Energy-Embed layers. We adopt Parallel WaveGAN (PWG) [31] as the vocoder to synthesize the waveform from the predicted mel-spectrograms.

Tacotron2, FCL-taco2-T and FCL-taco2-S are trained by Adam optimizer [32] with batch size of 32 and learning rate of 10^{-3} for 200, 100 and 100 epochs, respectively. FastSpeech2 is trained for 200 epochs with default settings of Espnet. To measure the speech quality of different models, two objective metrics, i.e., Mel Cepstral Distortion (MCD) and Root Mean Square Error of F_0 (F_0-RMSE), are adopted and calculated for all samples in the testing data. Besides, we randomly select 20 samples from the testing data for Mean Opinion Score (MOS) rating of speech naturalness, MOS is 5-scale (1-bad, 2-poor, 3-fair, 4-good, 5-excellent) and given by 15 listeners who are proficient in both English and Chinese.

4.2. Experimental results and analysis

4.2.1. Speech quality comparisons

Results of speech quality comparison are shown in Table 2. The evaluation of GT speech and the speech synthesized from GT mel-spectrograms, i.e., GT (Mel+PWG), are also presented, where the latter can be treated as the upper bound of voice quality of different TTS models. We can observe that for both EN and CN experiments, the proposed FCL-taco2-T and FCL-taco2-S achieve lower MCD and F_0-RMSE than Tacotron2, indicating that providing pitch and energy as inputs facilitates more accurate spectral features prediction. Besides, MOS results show that FCL-taco2-T also improves Tacotron2 and achieves similar performance to FastSpeech2. Also, FCL-taco2-S still preserves high-quality speech with minor performance degradation compared with its teacher and is comparable to Tacotron2, which shows the superiority of proposed methods to generate high-quality speech.

4.2.2. Impact of proposed knowledge distillation strategies

An ablation study is conducted to analyze the impact of different KD strategies, the results are illustrated in Table 3 and show that

Table 2. Speech quality comparisons of different TTS methods on EN and CN experiments.

Methods	MCD (dB)		F_0-RMSE (Hz)		MOS	
	EN	CN	EN	CN	EN	CN
GT	0	0	0	0	4.38	4.64
GT(Mel+PWG)	5.28	3.66	31.61	24.06	4.07	4.5
Tacotron2	7.17	5.63	44.01	38.31	3.78	3.86
FastSpeech2	**6.83**	**5.35**	41.68	35.10	**3.84**	**3.92**
FCL-taco2-T	6.90	5.38	**41.39**	**34.57**	3.83	**3.92**
FCL-taco2-S	6.91	5.46	41.66	35.11	3.74	3.82

Table 3. Impact of different KD strategies

Methods	MCD (dB)		F_0-RMSE (Hz)		MOS	
	EN	CN	EN	CN	EN	CN
FCL-taco2-S	**6.91**	**5.46**	**41.66**	**35.11**	**3.74**	**3.82**
w/o MSD	6.95	5.48	**41.66**	35.16	3.67	3.72
w/o HRD	6.96	5.57	41.90	35.45	3.72	3.78
w/o PD	6.92	5.52	41.98	35.75	3.69	3.69
w/o MSD+PD	7.07	5.57	41.87	35.41	3.53	3.61
w/o KD	7.20	5.84	42.36	35.86	3.40	3.53

Table 4. Model size and inference speed (RTF) comparisons on a single-CPU server, smaller RTF denotes faster speed

Models	Number of parameters	Inference speed		Speedup	
		EN	CN	EN	CN
Tacotron2	26.1M	0.408	0.463	/	/
FastSpeech2	53.1M	**0.021**	**0.024**	**19.4×**	**19.3×**
FCL-taco2-T	29.0M	0.077	0.090	5.3×	5.1×
FCL-taco2-S	**5.4M**	0.023	0.025	17.7×	18.5×

all KD strategies facilitate the learning of FCL-taco2-S to reduce MCD and F_0-RMSE. MOS results show that the lack of one or two KD strategies causes noticeable quality degradation, and performance drops significantly when no KD is adopted, which verifies the effectiveness of the proposed KD strategies for transferring the knowledge from teacher to student to maintain high-quality speech.

4.2.3 Model size and inference speed comparisons

Table 4 gives model size and inference speed comparisons on a single-CPU (Intel(R) Xeon(R) Gold 6148 @ 2.4GHz) server for different TTS acoustic models. With the proposed KD strategies, FCL-taco2-S has only 5.4 million (M) parameters and is **4.8×**, **5.4×** and **9.8× smaller** compared with Tacotron2, FCL-taco2-T and FastSpeech2, respectively. The speed is measured by real time factor (RTF) that denotes the time used to generate one second of mel-spectrogram frames. 1000 sentences with an average of 50 phonemes or pinyin units are used for testing. Due to the AR and NAR property, Tacotron2 and FastSpeech2 have the slowest and fastest inference speed, respectively. Our proposed FCL-taco2 is based on SAR mode and achieves the speed trade-off, and we observe that FCL-taco2-S with much smaller size can approximate the speed of FastSpeech2. Besides, we also deploy FCL-taco2-S on mobile phones with 4 different 8-core ARM processors for speed testing, and the results are shown in Table 5. We can observe that FCL-taco2-S performs well on low-end (Kirin 955 & Exynos 8890), middle-end (Kirin 970) and high-end (Kirin 980) mobile phones with RTF that is much less than 1. This indicates that FCL-taco2-S can be efficiently deployed on edge devices for faster than real-time speech synthesis.

Table 5. Inference speed (RTF) of FCL-taco2-S on mobile phones with different 8-core ARM processors

8-core ARM processors	Inference speed	
	EN	CN
Kirin 980 (2×A76@2.6GHz+2×A76@1.92GHz+4×A55@1.8GHz)	0.067	0.043
Kirin 970 (4×A73@2.36GHz+4×A53@1.8GHz)	0.171	0.127
Kirin 955 (4×A72@2.5GHz+4×A53@1.8GHz)	0.180	0.152
Exynos 8890 (4×M1@2.3GHz+4×A53@1.6GHz)	0.192	0.171

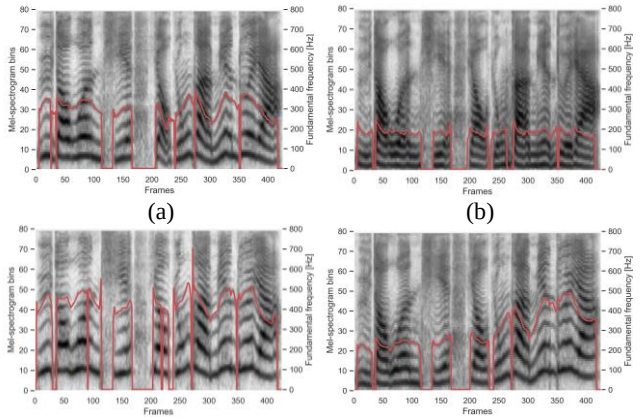

Fig. 2. Mel-spectrograms with the predicted F_0 that is multiplied by a ratio r: (a) r=1; (b) r=0.5; (c) r=1.5; (d) r is linearly increased from 0.5 to 1.5 phoneme by phoneme. Red curve denotes the F_0 contour. The corresponding text is '遇到问题要勇敢面对呀'.

4.2.4. Prosody manipulation

As the duration, pitch and energy are explicitly modelled in the proposed FCL-taco2, flexible prosody manipulation can be performed, e.g., modification of the duration to control the speech rate or F_0 to change the intonations or stress. Fig. 2 gives an example showing that by manually multiplying the predicted F_0 with different ratios for speech synthesis, the generated speech has the desired pitch contour changes, this shows it is simple yet effective to change F_0 of the generated speech with high quality. Readers are encouraged to listen to our demo for more details[1].

5. CONCLUSIONS

This paper presents a novel TTS system, i.e., FCL-taco2, which improves Tacotron2 in terms of three main aspects, namely, being (i) Fast, achieved by a novel SAR approach proposed for phoneme level based parallel mel-spectrograms generation to significantly speed up inference; (ii) Controllable, achieved by adding a prosody injector to offer high controllability of different prosody features; and (iii) Lightweight, achieved by three effective KD strategies to greatly compress a relatively large FCL-taco2 model, while maintain high-quality generated speech. These advantages enable the proposed model to be deployable on edge devices. Future study will focus on exploring more efficient architectures (e.g., Conv) and effective compression techniques (e.g., quantization) for TTS.

6. ACKNOWLEDGEMENTS

This research is partially supported by a grant from the HKSARG Research Grants Council General Research Fund (Project Reference No. 14208817).

7. REFERENCES

[1] H.-Y. Shum, X.-d. He, and D. Li, "From Eliza to XiaoIce: challenges and opportunities with social chatbots," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, pp. 10-26, 2018.

[2] Z.-H. Ling, S.-Y. Kang, H. Zen, A. Senior, M. Schuster, X.-J. Qian, *et al.*, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Processing Magazine*, vol. 32, pp. 35-52, 2015.

[3] J.-M. Valin and J. Skoglund, "A real-time wideband neural vocoder at 1.6 kb/s using LPCNet," *arXiv preprint arXiv:1903.12087*, 2019.

[4] B. Zhai, T. Gao, F. Xue, D. Rothchild, B. Wu, J. E. Gonzalez, *et al.*, "SqueezeWave: Extremely Lightweight Vocoders for On-device Speech Synthesis," *arXiv preprint arXiv:2001.05685*, 2020.

[5] S. Hussain, M. Javaheripi, P. Neekhara, R. Kastner, and F. Koushanfar, "Fastwave: Accelerating autoregressive convolutional neural networks on fpga," *arXiv preprint arXiv:2002.04971*, 2020.

[6] H. Zen, Y. Agiomyrgiannakis, N. Egberts, F. Henderson, and P. Szczepaniak, "Fast, compact, and high quality LSTM-RNN based statistical parametric speech synthesizers for mobile devices," *arXiv preprint arXiv:1606.06061*, 2016.

[7] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779-4783.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, Z. Y. Jaitly, *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," in *Interspeech*, 2017, pp. 4006-4010..

[10] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, *et al.*, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," *arXiv preprint arXiv:1710.07654*, 2017.

[11] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 6706-6713.

[12] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, *et al.*, "Fastspeech: Fast, robust and controllable text to speech," in *Advances in Neural Information Processing Systems*, 2019, pp. 3171-3180.

[13] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech," *arXiv preprint arXiv:2006.04558*, 2020.

[14] K. Peng, W. Ping, Z. Song, and K. Zhao, "Parallel neural text-to-speech," *arXiv preprint arXiv:1905.08459*, 2019.

[15] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-TTS: A Non-Autoregressive Network for Text to Speech Based on Flow," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7209-7213.

[16] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search," *arXiv preprint arXiv:2005.11129*, 2020.

[17] Z. Zeng, J. Wang, N. Cheng, T. Xia, and J. Xiao, "Alignntts: Efficient feed-forward text-to-speech system without explicit alignment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6714-6718.

[18] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "MoBoAligner: a Neural Alignment Model for Non-autoregressive TTS with Monotonic Boundary Search," *arXiv preprint arXiv:2005.08528*, 2020.

[19] S. Beliaev, Y. Rebryk, and B. Ginsburg, "TalkNet: Fully-Convolutional Non-Autoregressive Speech Synthesis Model," *arXiv preprint arXiv:2005.05514*, 2020.

[20] A. Łańcucki, "FastPitch: Parallel Text-to-speech with Pitch Prediction," *arXiv preprint arXiv:2006.06873*, 2020.

[21] J. Vainer and O. Dušek, "SpeedySpeech: Efficient Neural Speech Synthesis," *arXiv preprint arXiv:2008.03802*, 2020.

[22] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Interspeech*, 2017, pp. 498-502.

[23] Y. Cong, R. Zhang, and J. Luan, "PPSpeech: Phrase based Parallel End-to-End TTS System," *arXiv preprint arXiv:2008.02490*, 2020.

[24] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577-585.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735-1780, 1997.

[27] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133-4141.

[28] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, *et al.*, "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019.

[29] K. Ito, "The lj speech dataset," 2017, Available at https://keithito.com/LJ-Speech-Dataset (visited Sep. 2020)

[30] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, *et al.*, "Espnet: End-to-end speech processing toolkit," in *Interspeech*, 2018, pp. 2207-2211.

[31] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199-6203.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.