MSDTRON: A HIGH-CAPABILITY MULTI-SPEAKER SPEECH SYNTHESIS SYSTEM FOR DIVERSE DATA USING CHARACTERISTIC INFORMATION

Qinghua Wu *Quanbo Shen* *Jian Luan* *Yujun Wang*

Xiaomi Corporation, Beijing, China

ABSTRACT

In multi-speaker speech synthesis, data from a number of speakers usually tend to have great diversity due to the fact that the speakers may differ largely in ages, speaking styles, emotions, and so on. It is important but challenging to improve the modeling capabilities for multi-speaker speech synthesis. To address the issue, this paper proposes a high-capability speech synthesis system, called Msdtron, in which 1) a representation of the harmonic structure of speech, called excitation spectrogram, is designed to directly guide the learning of harmonics in mel-spectrogram. 2) conditional gated LSTM (CGLSTM) is proposed to control the flow of text content information through the network by re-weighting the gates of LSTM using speaker information. The experiments show a significant reduction in reconstruction error of mel-spectrogram in the training of the multi-speaker model, and a great improvement is observed in the subjective evaluation of speaker adapted model.

Index Terms— multi-speaker speech synthesis, voice cloning, speaker adaptation, few-shot speech synthesis

1. INTRODUCTION

Neural text-to-speech is very popular in recent years [1, 2], and it can already produce speech that's almost as natural as speech from a real person with high voice quality. However, data collection is still a big challenge. We often need to collect a large amount of data in a high-fidelity recording studio with professional guidance to obtain high voice quality and high consistency of recordings. It is very costly, time-consuming, or even impossible, e.g. in cases of custom speech and Lombard speech [3]. Meanwhile, noisy and diverse data is usually easier to be collected. Thereby multi-speaker speech synthesis is proposed, which uses diverse data from lots of speakers to train a robust multi-speaker generative model. It can be further adapted in different tasks such as speaker adaptation [4], cross-lingual text-to-speech synthesis [5], and style conversion [6].

The state-of-art systems have an encoder-decoder structure network with speaker embedding as additional inputs [6, 7, 8, 9, 10]. Some works investigated the effective representations of speakers, e.g. [8, 9] studied the effects of different speaker embeddings such as x-vector [11], LDE-based speaker encoding [9]. [7] proposed an attention-based variable-length embedding. [12] measured the speaker similarity between the predicted mel-spectrogram and the reference. Some works focused on solving the problem of noisy data [4, 6], e.g. [6] did research into the methods of transfer learning for noisy samples. [13] aimed to disentangle speaker embedding and noise by data augmentation and a conditional generative model. And some works were interested in the controllability of systems in the manner of zero-shot. [7, 9] tried to obtain target voice by feeding target speaker embedding without speaker adaptation. [14] introduced latent variables to control the speaking style.

The previous studies rarely gave insights into what role the characteristic information such as timbre and style played. The characteristic information is usually represented by a fixed-or-variable-length embedding which may not be effective enough, e.g. the pitch embedding is relevant to the harmonics of speech but it's not an effective representation of the harmonic structure. Besides, the embedding of characteristic information is typically concatenated or added to the text content representation or is simply used to perform an affine transformation on it [15]. In this way, the characteristic information is playing a similar role as the text content information in the network, which is not what we expected.

In this paper, we propose an encoder-decoder structured speech synthesis system (Msdtron), investigating the effective use of characteristic information. The major contributions are: 1) Excitation spectrogram is designed to explicitly characterize the harmonic structure of speech. It acts as a skeleton of the target mel-spectrogram to optimize the learning process. 2) Conditional gated LSTM (CGLSTM) is proposed whose input/output/forget gates are re-weighted by speaker embedding while the cell/hidden states are dependent on the text content. That's to say, the speaker embedding controls the flow of text content information by gates without affecting cell state and hidden state directly.

The rest article is organized as follows: Section 2 describes the framework of the proposed system (Section 2.1), excitation spectrogram generator (Section 2.2), and CGLSTM in the decoder (Section 2.3). Section 3 is about the detailed settings and results of experiments. Finally, a conclusion is drawn in Section 4.

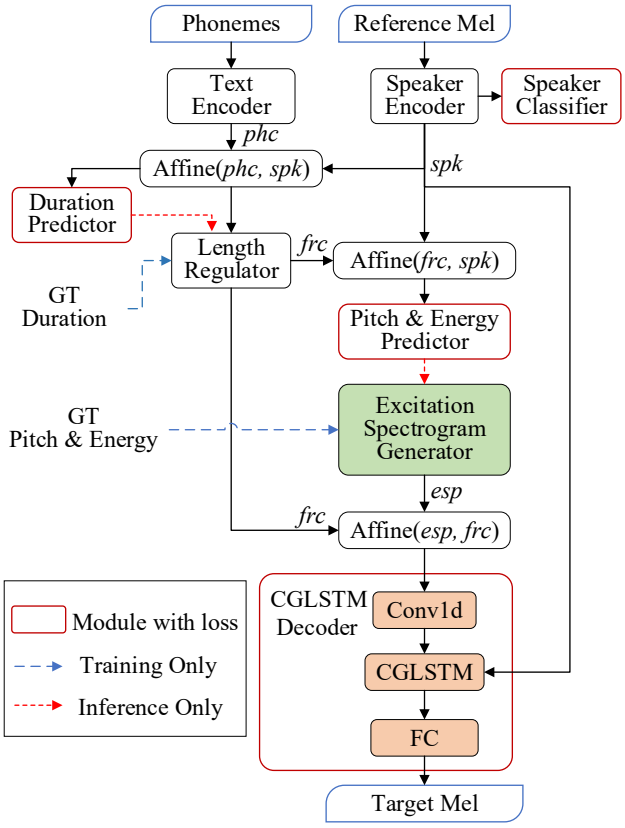

Fig. 1. The overall framework of Msdtron. (p.s. GT means ground-truth)

2. MSDTRON

2.1. Framework

The framework of Msdtron is illustrated as Figure 1, a state-of-art encoder-decoder structure. The text-encoder is a standard Tacotron2 encoder [2] which has a stacked Conv1d followed by BLSTM. It takes phoneme sequences as inputs and learns an embedding of the text content. Besides, a sentence-level speaker embedding is extracted from the reference mel-spectrogram by a GST-like [16] speaker-encoder, with a stacked Conv2d+BatchNorm reference encoder, and a multi-head attention [17]. Instead of introducing Gradient Reversal Layer (GRL) [18] to remove the text content information from the speaker embedding, the reference is randomly chosen from the utterances of the same speaker with the target [10].

Similar to [1], a length regulator is used for the alignment between phoneme sequences and target mel-spectrogram sequences. The duration predictor is simply a layer of BLSTM with pre-dense and post-dense layers.

Pitch and energy are predicted separately with the same network structure, a stacked Conv1d with post-dense. Then excitation spectrogram is determined by the pitch and energy (see Section 2.2), which explicitly represents the harmonic structure of speech.

The decoder is to predict the target mel-spectrogram with the proposed CGLSTM (see Section 2.3) in it.

Finally, in the flow of information, affine transformations are carried out. It is defined as follows:

$$affine(x_i, x_c) = x_i \odot proj^1(x_c) + proj^2(x_c) \quad (1)$$

$$proj(x) = w \otimes x + b \quad (2)$$

where $\odot$ means element-wise multiplication, and $\otimes$ means matrix multiplication.

2.2. Excitation Spectrogram Generator

In the source-filter analysis [19], speech is produced when an excitation signal passes through a filter representing the vocal track characteristics. It's composed of harmonic and noise components, as a result of periodic or aperiodic signals modulated by the vocal track filter. Existing studies usually use pitch information to model the periodicity of speech. Unfortunately, the pitch has a poor ability to express the harmonic characteristics of speech. Thus an explicit representation of harmonic structure in the spectrogram is proposed, called excitation spectrogram, in this session.

To make things easy, it is assumed that speech signals consist of harmonic components only at voiced segments, and noise components only at unvoiced segments. The production of voiced speech is modeled as a time-varying periodic impulse passing through the vocal track filter and generating harmonics at multiples of the periodic frequency as Equation 3. The unvoiced speech is generated from a white Gaussian noise signal.

$$har_i = i * f_0 \quad i \in [1, N_h] \quad (3)$$

where har_i means the i^{th} harmonic frequency of speech, N_h is the number of harmonics, and $f0$ is the periodic (fundamental) frequency.

In the excitation spectrogram, energy is supposed evenly distributed on each harmonic frequency of voiced segments or the whole frequency band of unvoiced segments. It can help to learn the accurate distribution of energy in the frequency of mel-spectrogram. Thus, the excitation spectrogram can be defined as Equation 4.

$$els_{i,j} = \begin{cases} e_i/N_f & if\ i \in unvoiced \\ e_i/N_h & if\ i \in voiced,\ j \in harmonics \\ 0 & if\ i \in voiced,\ j \notin harmonics \end{cases} \quad (4)$$

where $els_{i,j}$ means the j^{th} linear spectrum at i^{th} frame, e_i is the total energy of i^{th} frame, and N_f is the fft number in the calculation of linear spectrum.

Finally, the linear excitation spectrogram els is converted to mel excitation spectrogram esp by Equation 5

$$esp = els \otimes W^{l2m}_{N_f \times N_m} \quad (5)$$

where $W^{l2m}_{N_f \times N_m}$ is the transformation matrix from linear to mel spectrogram and N_m is the dimension of mel-spectrogram.

2.3. Conditional Gated LSTM

Text content is the most important feature of speech due to its decisive role in intelligibility. In addition, speech can be characterized in terms of timbre, style, speed, and emotion, etc. Many researches aim to change or control some of these characteristics without a negative influence on the intelligibility and voice quality. For this purpose, the embedding of characteristic information is usually added or concatenated to the text content embedding, which is fed as the inputs of network. However, in this way, the characteristic information plays a similar role to the text content. Both of them would directly take effect in the same way on the intelligibility, voice quality, and other characteristics at the same time, which is not the way we expected. Consequently, we propose conditional gated LSTM (CGLSTM) where the characteristic information is used to re-weight the gates and the text content flows in the hidden/cell states. Thereby the characteristic information will directly play a part in the gates-based flow of the text content without operating the text content in itself.

Compared with LSTM, which is frequently used in speech synthesis tasks due to its good capacity of learning long dependencies, CGLSTM uses the characteristic information to re-weight the input/output/forget gates as Equation 6.

$$g_t = \sigma((W_{xg} \otimes [h_{t-1}, x_t] + b_{xg}) \odot (W_{cg} \otimes c_t + b_{cg})) \quad (6)$$

where g_t can be the forget, input or output gate; x_t, c_t, and h_{t-1} are the current text content inputs, current characteristic information embedding, and previous hidden state; W and b are the corresponding weights and biases.

3. EXPERIMENTS

3.1. Corpus

The data set of our experiments is the public multi-speaker mandarin corpus AISHELL-3 [20], which contains roughly 85 hours of recordings spoken by 218 native Chinese mandarin speakers. Among them, recordings from 173 speakers have Chinese character-level and pinyin-level transcripts and a total of 63263 utterances. This part of the transcribed data will be used in our experiments, which is divided into the training set and test set without overlapping.

- **Training set**: contains 57304 utterances from 165 speakers, with 133 females 46915 utterances and 32 males 10389 utterances. The training set is used to pre-train the multi-speaker generative model, which is further adapted using the test set.
- **Test set**: contains 4 females and 4 males, and only 20 utterances of each speaker are randomly chosen for speaker adaptation.

The recordings are mono, 16bit, and down-sampled from 44100HZ to 16000HZ. Preprocessing is conducted on both the training and the test sets to reduce the diversity of them: 1) Energy normalization by scaling the maximal amplitude of utterance. 2) Silence trimming by keeping 60ms silence at the head and tail of utterance.

3.2. Setup

The pipeline of our experiments includes 1) Pre-training: train the multi-speaker generative model using the training set. 2) Speaker adaptation: train the target model by transfer learning using the single-speaker data from the test set and 3) Synthesis: infer the mel-spectrogram and synthesize the waveform by Hifi-Gan vocoder[21].

In our experiments, the frame hop size is set to 12.5ms, the window size 50ms, and the number of mel-bank 80 for mel-spectrogram. Mean Absolute Error (MAE) is calculated to measure the reconstruction error of pitch and energy while Mean Square Error (MSE) is applied to mel-spectrogram. Besides, the task of speaker classification uses cross-entropy as the loss function. The setup of our experiments is described as follows:

- **Baseline**: A modified Tacotron2 with a length regulator replacing attention-based alignment. In order to cope with the case of multi-speakers, a speaker-encoder is added. In detail, following modifications are made on the proposed system (Figure 1): 1) The excitation spectrogram generator is removed. 2) CGLSTM in the decoder is replaced with the standard LSTM while an affine transformation is conducted on the speaker embedding and the text content before they are fed to the decoder.
- **System-1**: Baseline + excitation spectrogram generator, which is the system Msdtron without CGLSTM-decoder.
- **System-2**: Baseline + excitation spectrogram generator + CGLSTM decoder. It is our proposed system Msdtron in Figure 1.

3.3. Multi-speaker Model

Figure 2 shows the reconstruction error of mel-spectrogram of different systems in the pre-training stage. Compared with

the Baseline, the excitation-spectrogram generator brought an obvious improvement in terms of reconstruction error in System-1. The reconstruction error reduced further in System-2 when the CGLSTM-decoder was introduced. It indicates that the excitation spectrogram and CGLSTM can greatly improve the modeling capability of systems for multi-speaker corpus.

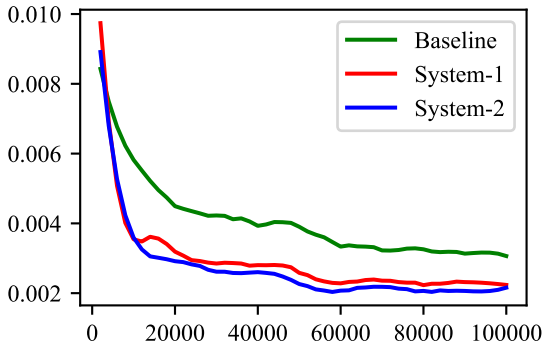

Fig. 2. Reconstruction Errors of mel-spectrogram in different systems. (x-label: steps of training; y-label: mel-error)

We also compared the number of parameters of systems in Table 1. It drops by 10% from Baseline to System-2. In other words, we can achieve better performance with less computation and less memory.

Table 1. The number of parameters of systems

Baseline	System-1	System-2
11.24 M	9.5 M	10.08 M

3.4. Speaker Adapted Model

For unseen speakers in the test set, we adapted the multi-speaker model using data of the target speaker[1] . The Mean Opinion Score (MOS) test was carried out to evaluate its performance in voice quality of speech and speaker similarity. 9 native Chinese testers participated in it. The MOS results are shown in Table 2.

According to the MOS results, System-1 outperformed the Baseline in both aspects of voice quality and speaker similarity. It indicates that the excitation spectrogram is much more effective than the simple use of pitch and energy. It can reduce the noise or signal distortion caused by insufficient modeling capabilities for complex data, thus improving the clearness of pronunciation. Besides, System-2 achieved the best performance, which proves that CGLSTM can control the specific characteristics of voice better than LSTM while the voice quality is improved at the same time.

[1]Audio examples: https://Msdtron.github.io/

Table 2. MOS of voice quality (Quality) and speaker similarity (Similarity) for unseen-speaker after speaker adaptation. In the evaluations, scores range from 1 to 5. 1) For voice quality, score=1 means the voice has strong and annoying noise, or bad pronunciation while score=5 means the voice is clean, pleasant, and clear pronounced. 2) For speaker similarity, score=1 means the compared two voices don't sound like from the same person at all while score=5 means it's easy to make a judgment that they are from the same person.

	Quality		**Similarity**	
	female	male	female	male
Ground-truth	4.70	4.42	-	-
Baseline	2.71	2.76	3.39	3.28
System-1	3.67	3.06	3.89	3.31
System-2	**3.97**	**3.58**	**4.06**	**3.64**

Furthermore, more improvement was observed on females by comparing Baseline and System-1 while more was observed on males from System-1 to System-2. One possible reason is the unbalanced data for females and males in the training set, with a rough ratio female : male = 9 : 2. When the excitation spectrogram was used, it helped System-1 to learn more information of females than of males and thus to achieve better performance on females. Meanwhile, in System-2 with CGLSTM-decoder, the speaker embedding is attempted to control the characteristics of speakers without a negative impact on voice quality. Therefore, System-2 could share knowledge better apart from speaker information between males and females, which brought obvious improvement on males.

Finally, we notice that the voice quality of Baseline is relatively low. The main flaw of the synthesized voices is the annoying noise. It can be partly explained by the fact that the corpus AISHELL-3 contains a certain reverberation and background noise, which often degrades the performance of TTS systems. We can try to add additional professional recordings into the training set, just as the M2VoC-2021 challenge did. It also proves that Tacotron2 and its simple variants don't perform well on diverse multi-speaker data.

4. CONCLUSIONS

In this paper, we have proposed the excitation spectrogram to represent the harmonic structure of speech, and CGLSTM to better control the specific characteristics of speech with less impact on the voice quality than LSTM. The proposed system (Msdtron) outperformed the baseline largely both on the voice quality and the speaker similarity. More researches will be conducted to enhance the cleanliness and expressiveness of synthesized voices with more diverse data of different noise levels and speaking styles in the future.

5. REFERENCES

[1] Yi Ren, C. Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *ArXiv*, vol. abs/2006.04558, 2020.

[2] Jonathan Shen, R. Pang, Ron J. Weiss, M. Schuster, Navdeep Jaitly, Z. Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. Skerry-Ryan, R. A. Saurous, Yannis Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783, 2018.

[3] Bajibabu Bollepalli, Lauri Juvela, and P. Alku, "Lombard speech synthesis using transfer learning in a tacotron text-to-speech system," in *INTERSPEECH*, 2019.

[4] Qiong Hu, E. Marchi, David Winarsky, Y. Stylianou, Devang K. Naik, and Sachin S. Kajarekar, "Neural text-to-speech adaptation from low quality public recordings," in *Speech Synthesis Workshop 10*, 2019.

[5] Mengnan Chen, Minchuan Chen, S. Liang, J. Ma, Lei Chen, Shaojun Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding," in *INTERSPEECH*, 2019.

[6] D. Paul, P. V. M. Shifas, Yannis Pantazis, and Y. Stylianou, "Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion," *ArXiv*, vol. abs/2008.05809, 2020.

[7] Seungwoo Choi, Seungju Han, Dong-Young Kim, and Sungjoo Ha, "Attentron: Few-shot text-to-speech utilizing attention-based variable-length embedding," *ArXiv*, vol. abs/2005.08484, 2020.

[8] Chung-Ming Chien, Jheng hao Lin, Chien yu Huang, Po chun Hsu, and Hung yi Lee, "Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech," *ArXiv*, vol. abs/2103.04088, 2021.

[9] E. Cooper, Cheng-I Lai, Y. Yasuda, Fuming Fang, Xin Eric Wang, Nanxin Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6184–6188, 2020.

[10] Ye Jia, Y. Zhang, Ron J. Weiss, Q. Wang, Jonathan Shen, Fei Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez-Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, 2018.

[11] David Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[12] Zexin Cai, C. Zhang, and Ming Li, "From speaker verification to multispeaker speech synthesis, deep transfer with feedback constraint," *ArXiv*, vol. abs/2005.04587, 2020.

[13] Wei-Ning Hsu, Yu Zhang, Ron J. Weiss, Yu-An Chung, Yuxuan Wang, Y. Wu, and James R. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5901–5905, 2019.

[14] Wei-Ning Hsu, Y. Zhang, Ron J. Weiss, H. Zen, Y. Wu, Yuxuan Wang, Yuan Cao, Y. Jia, Z. Chen, Jonathan Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," *ArXiv*, vol. abs/1810.07217, 2019.

[15] N. Kumar, Srishti Goel, A. Narang, and Brejesh Lall, "Few shot adaptive normalization driven multi-speaker speech synthesis," *ArXiv*, vol. abs/2012.07252, 2020.

[16] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. Skerry-Ryan, Eric Battenberg, Joel Shor, Y. Xiao, Fei Ren, Ye Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *ICML*, 2018.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[18] Yaroslav Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," *ArXiv*, vol. abs/1409.7495, 2015.

[19] Masanori Morise, Fumiya Yokomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99-D, pp. 1877–1884, 2016.

[20] Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *ArXiv*, vol. abs/2010.11567, 2020.

[21] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *arXiv preprint arXiv:2010.05646*, 2020.