DelightfulTTS: The Microsoft Speech Synthesis System for Blizzard Challenge 2021

Yanqing Liu[1], Zhihang Xu[1], Gang Wang[1], Kuan Chen[1], Bohan Li[1], Xu Tan[2], Jinzhu Li[1], Lei He[1], Sheng Zhao[1]

[1]Microsoft Azure Speech, China
[2]Microsoft Research Asia, China

{yanqliu, zhihangxu, gaw, kuan.chen, bohli, xuta, jinzl, helei, szhao}@microsoft.com

Abstract

This paper describes the Microsoft end-to-end neural text to speech (TTS) system: DelightfulTTS for Blizzard Challenge 2021. The goal of this challenge is to synthesize natural and high-quality speech from text, and we approach this goal in two perspectives: The first is to directly model and generate waveform in 48 kHz sampling rate, which brings higher perception quality than previous systems with 16 kHz or 24 kHz sampling rate; The second is to model the variation information in speech through a systematic design, which improves the prosody and naturalness. Specifically, for 48 kHz modeling, we predict 16 kHz mel-spectrogram in acoustic model, and propose a vocoder called HiFiNet to directly generate 48 kHz waveform from predicted 16 kHz mel-spectrogram, which can better trade off training efficiency, modelling stability and voice quality. We model variation information systematically from both explicit (speaker ID, language ID, pitch and duration) and implicit (utterance-level and phoneme-level prosody) perspectives [1]: 1) For speaker and language ID, we use lookup embedding in training and inference; 2) For pitch and duration, we extract the values from paired text-speech data in training and use two predictors to predict the values in inference; 3) For utterance-level and phoneme-level prosody, we use two reference encoders to extract the values in training, and use two separate predictors to predict the values in inference. Additionally, we introduce an improved Conformer [2] block to better model the local and global dependency in acoustic model [3, 4]. For task SH1, DelightfulTTS achieves 4.17 mean score in MOS test and 4.35 in SMOS test, which indicates the effectiveness of our proposed system.

Index Terms: Text to Speech, TTS, 48 kHz, HiFiNet, Fast-Speech, AdaSpeech, Conformer

1. Introduction

The Blizzard Challenge[1] is aimed at advancing the technologies in text to speech (TTS) [5, 6, 3] by comparing and understanding different approaches, and has been organized annually since 2005 [7]. The basic task is to build high-quality TTS systems based on the speech database provided by the organizers. Participants use their developed systems to synthesize audio from the given test set; the generated audio samples are used to evaluate performance of different system through subjective listening tests and objective metrics.

Typical TTS systems consist of three key components [1]: text analysis, acoustic model, and vocoder. Firstly, input text is normalized and transformed into linguistic features of different levels: phone-level, syllable-level, word-level, through a text analysis (TTS front-end) module. Then, these linguistic features are transformed into an intermediate acoustic representation like mel-spectrogram by an acoustic model. At last, the acoustic representation is converted to waveform with a vocoder. Rich unit selection based concatenative speech synthesis [8] and statistical parametric speech synthesis like HMM/DNN/LSTM based models [9, 10, 11], have been the most popular methods in the past. Recently, end-to-end neural TTS has been researched extensively. Tacotron 2 [5] introduced an attention-based encoder-decoder acoustic model to predict mel-spectrogram given a character sequence and a WaveNet model [12] to synthesize waveform from mel-spectrogram, which achieves high voice quality but suffers from robustness issues and slow speed in both training and inference. To improve training speed, TransformerTTS [13] have been proposed to synthesize high quality speech with Transformer [14]. [15] tried to solve the stability problems like repeated phones by using linguistic features and phone duration with autoregressive Transformer, but the inference is still slow due to autoregressive nature. FastSpeech [6] leads a research trend on non-autoregressive TTS, by using feed-forward transformer in encoder and decoder for parallel generation and a duration predictor for length duration, which is much faster and more stable than autoregressive models like Tacotron 2 [5].

While there is much progress in the research on TTS and the quality of synthesized speech has been impressive, the perception quality and naturalness still have large space to improve. With this goal in mind, we participate the Blizzard Challenge 2021, and design an end-to-end neural text to speech (TTS) system called DelightfulTTS to advance the perception quality and naturalness. We approach this goal from two aspects: 1) The bandwidth to convey high-quality speech should be enough. We model and generating 48 kHz waveform, since high sampling rate (48 kHz) can have larger range of frequency to convey expression and prosody [16] and thus can result in higher perception quality than 16 kHz or 24 kHz. 2) The speech itself should be expressive. We model variation information [1] in speech through a systematic design to improve prosody and naturalness. TTS is a typical one-to-many mapping problem where there could be multiple varying speech outputs (e.g., different pitch, duration, speaker, prosody, emotion, etc) for a given text input. Modeling these variation information [17, 3, 18, 4] can mitigate the one-to-many mapping problem and thus improve expressiveness and fidelity of synthesized speech. We describe the specific design in each aspect as follows:

- To generate 48 kHz waveform, typical method is to generate 48 kHz mel-spectrogram and then generate

---

48 kHz waveform from the generated 48 kHz mel-sepctrogram. However, according to the recent advances in TTS [1, 19], the quality of vocoder task nearly matches to that of the recording [19], while the quality of acoustic model task still has gap to recording due to the difficulty of prosody modeling. Intuitively speaking, vocoder only needs recover phase information given ground-truth spectrogram, while acoustic model needs to close the large gap between text and spectrogram, such as duration, pitch, formant, voiced/unvoiced part, prosody, etc. When increasing the sampling rate to 48 kHz, the burden of acoustic model is further increased. According to the "cannikin law (wooden barrel principle)", the capacity of a wooden barrel is determined by its shortest stave. Thus, to relieve the burden on acoustic model, we predict 16 kHz mel-spectrogram in acoustic model, and generate 48 kHz waveform from predicted 16 kHz mel-spectrogram using our recently developed HiFiNet [20] vocoder.

- Nevertheless, the burden of acoustic model is still heavy due to one-to-many mapping. Thus, we model variation information systematically from both explicit (speaker ID, language ID, pitch and duration) and implicit (utterance-level and phoneme-level prosody) perspectives [1]: 1) For speaker and language ID, we use lookup embedding in training and inference; 2) For pitch and duration, we extract the values from paired text-speech data in training and use two predictors to predict the values in inference; 3) For utterance-level and phoneme-level prosody, we use two reference encoders to extract the values in training, and use two separate predictors to predict the values in inference. We build the acoustic model based on non-autoregressive generation model FastSpeech [6] and AdaSpeech [4], with an improved Conformer module [2] to better model the local and global dependency in mel-spectrogram.

According to the official evaluation of Blizzard Challenge 2021 in task SH1, our proposed system achieves 4.17 mean score in MOS test and 4.35 in SMOS test(recording has 4.21 and 4.07 mean score in MOS and SMOS test respectively), which is better than other participants and matches to the quality of recording in mean score perspective of MOS and SMOS, which demonstrate the superiority of our system to advance the perception quality and naturalness[2].

The rest of the paper is organized as follows. Section 2 presents the architecture of the proposed system. Section 3 describes the task, data process and training strategy. Section 4 introduces the subjective evaluation results. Finally, conclusion and future work are presented in Section 5.

2. Our Approach

2.1. System Overview

As shown in Figure 1a, our proposed system consists of an acoustic model that generates mel-spectrogram from text and a vocoder that generates waveform from mel-spectrogram. To ensure high fidelity and quality of the synthesized speech, we directly generate waveform in 48 kHz sampling rate, since it can have larger range of frequency to convey expression and prosody [16]. To better trade off the difficulty in acoustic model

---

[2] Audio samples: https://cognitivespeech.github.io/delightfultts

and vocoder, we predict 16 kHz mel-spectrogram with acoustic model, and generate 48 kHz waveform from predicted 16 kHz mel-spectrogram with vocoder, which are described in Section 2.2. The acoustic model consists of an improved Conformer module [2] in the phoneme side and mel-spectrogram side respectively, and a variance adaptor in between the two conformer modules, as shown in Figure 1d. We model the variance information through variance adaptor to improve the prosody and naturalness in both explicit and implicit ways (Figure 1b), as introduced in Section 2.3. We introduce the details of the improved Conformer module (Figure 1c) in Section 2.4. The vocoder is based on our previously proposed HiFiNet [20]. We introduce each design in our proposed system in the following subsections.

2.2. Tradeoff between Acoustic Model and Vocoder

We directly generate waveform in 48 kHz sampling rate to ensure high fidelity and quality for speech synthesis. There is a tradeoff in the task difficulty between acoustic model and vocoder, which influences the design choices on acoustic model and vocoder. Intuitively speaking, the burden of acoustic model is larger than that of vocoder, according to several perspectives: 1) Information gap. Vocoder only needs to recover phase information given spectrogram, while acoustic model needs to close the large gap between text and spectrogram, such as duration, pitch, formant, voiced/unvoiced part, prosody, etc. Thus, the large information gap between text and spectrogram makes the task of acoustic model more difficult than that of vocoder. 2) Current research progress. According to the recent advances in TTS [1, 19], the quality of vocoding task nearly matches to that of the recording [19], while the acoustic modeling task still has large improvement space due to the difficulty of prosody modeling. 3) Model capacity. We conduct a brief survey on the size of model parameters of the mainstream acoustic models and vocoders (mainly GAN-based) [1], and find that the parameter sizes of acoustic models are usually 5x (or more) larger than that of vocoders.

If we generate 48 kHz mel-spectrogram with acoustic model and then generate 48 kHz waveform with vocoder, the task difficulties of both acoustic model and vocoder are increased, compared with 16 kHz counterparts. However, the burden of acoustic model is even larger according to the analysis above. According to the "cannikin law" (i.e., "wooden barrel principle"), the capacity of a wooden barrel is determined by its shortest stave. In our case, the quality of speech synthesis would be constrained by the acoustic model if not well handled. Thus, to relieve the burden on acoustic model, we predict 16 kHz mel-spectrogram with acoustic model, and generate 48 kHz waveform from predicted 16 kHz mel-spectrogram using our recently developed HiFiNet [20] vocoder, which is optimized with a generator and discriminator-based network [21]. In this way, the task difficulty of acoustic model is reduced and that of vocoder is slightly increased (with both phase reconstruction and super-resolution), resulting in a more balanced way.

2.3. Variation Information Modelling

TTS is a typical one-to-many mapping problem where there could be multiple varying speech outputs (e.g., different pitch, duration, speaker, prosody, emotion, etc) for a given text input. It is critical to model these variation information in speech so as to improve expressiveness and fidelity of synthesized speech. While previous works [17, 3, 18] have tried different methods to model the information, they focus on a specific aspect
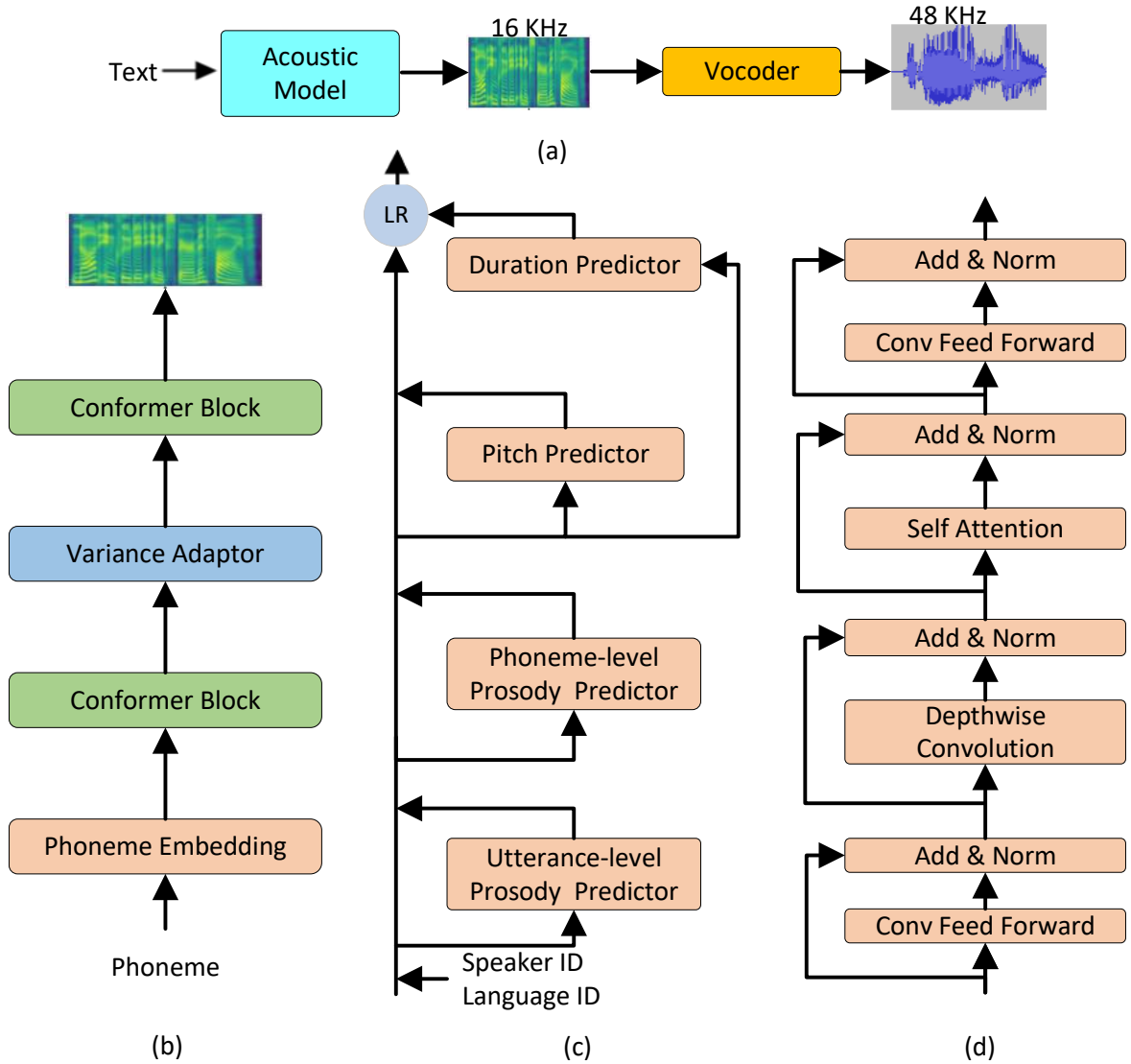
Figure 1: *Overview of our proposed TTS system. Figure (a) shows the overall pipeline for DelightfulTTS. Figure (b) shows the architecture of the acoustic model. Figure (c) shows the variance adaptor with explicit and implicit variation information modeling. Figure (d) shows the improved Conformer block.*

and cannot model in a comprehensive and systematic way. In this paper, considering that different variation information can be complementary to each other, we propose a unified way to model them in the proposed variance adaptor (as shown in Figure 1c). Observing that some variation information can be obtained implicitly (e.g., pitch can be extracted by some tools) or explicitly (e.g., utterance-level prosody can only be learned by the model), we categorize all the information we model as follows: 1) Explicit modeling, includes language ID, speaker ID, pitch, duration. 2) Implicit modeling, including utterance-level and phoneme-level prosody.

For speaker and language ID, we use lookup embedding in training and inference. For pitch and duration, we extract the values from paired text-speech data in training and use two predictors [3] to predict the values in inference. For utterance-level and phoneme-level prosody, we use two reference encoders to extract the values in training [4] and use two separate predictors

to predict the values in inference. The two reference encoders are both made up of convolution and RNN layers. Utterance-level prosody vector is obtained by the last RNN hidden and a style token layer [22]. Phoneme-level prosody vectors are obtained by using the outputs of phoneme encoder (phoneme-level) as query to attend to the outputs of mel-spectrogram reference encoder (frame-level). Different from [23], we do not use VAE, but directly use the latent representation as phoneme-level vector for training stability [4]. The utterance-level prosody predictor contains a GRU layer followed by a bottleneck module to predict the prosody vector. The phoneme-level prosody predictor takes both the outputs of text encoder and utterance-level prosody vector as input. With the help of utterance-level prosody vector, we do not need an autoregressive prosody predictor as [23] for faster inference. By unifying explicit and implicit information in different granularities (language-level, speaker-level, utterance-level, phoneme-level) [1] in the vari-

ance adaptor, we can achieve better expressiveness in prosody and controllability in pitch and duration.

2.4. Conformer in Acoustic Model

Conformer [2] is a Transformer variant that integrates both CNNs and Transformers components. Originally used for end-to-end speech recognition, it exhibits better accuracy with fewer parameters than previous work on several open ASR datasets, and achieves a new state-of-the-art performance. The multi-headed self-attention (MHSA) is conformer integrates an important technique from Transformer-XL[24] with relative sinusoidal positional encoding scheme. Conformer proposes a novel combination of self-attention and convolution, in which self-attention learns the global interaction while the convolutions efficiently capture the local correlations. We think that the global and local interactions are especially important for TTS considering it has a longer output sequence than machine translation or speech recognition in decoder. For non-autoregressive TTS model, a powerful modelling unit is critical because each frame in decoder cannot see its history as autoregressive model does. In our proposed system, we make several improvements to the original Conformer architecture. First, we replace Swish [25] with ReLU and observe better generalization especially on long sentences. Second, we switch the order of self-attention and depthwise convolution for faster convergency. Third, we replace the linear layers in feed forward-modules of Macron [26] structure with convolution layers, which results in better prosody and audio fidelity. Altogether, our improved conformer block is composed of four modules stacked together, i.e., a convolutional feed-forward module, a depthwise convolution module, a self-attention module and a second convolutional feed-forward module in the end, illustrated in Figure 1d.

3. Task Description, Data Processing, and Training Strategy

3.1. Task Description

The task in Blizzard Challenge 2021 we joined is Hub task 2021 - SH1. There're about 5 hours of speech data with 48 kHz sampling rate from a female native speaker of European Spanish. The hub task is to build a voice from the provided European Spanish data to synthesize texts containing only Spanish words.

3.2. Data Processing

Phoneme representation is widely used as the input of end-to-end TTS systems like recent neural TTS models [13, 15, 6, 27], which can avoid ambiguity depending on the text context. In this task we leverage an internal text preprocess module that converts the input Spanish text into a phoneme sequence. First, we split the text into sentences through sentence separation module. Second, we perform text normalization (TN) with a rule-based TN module. Third, we convert the normalized text to phoneme sequence with a G2P [28] model and a Spanish lexicon. The duration target is extracted by an internal force alignment model. For acoustic model training, audio used in our system is down-sampled to 16 kHz. Then mel-spectrograms are computed through a short time Fourier transform (STFT) using a 50 ms frame size, 12.5 ms frame hop, and a Hann window function. For vocoder training, the original 48k audios were used in the training of HiFiNet [20].

3.3. Training Strategy

The overall loss function for the proposed acoustic model is

$$L = L_{utt} + L_{phone} + L_{pitch} + L_{dur} + L_{iter} + L_{ssim} \quad (1)$$

where L_{utt} is the L1 loss between predicted utterance-level prosody vector and the vector extracted from utterance-level reference encoder, L_{phone} is the L1 loss between the predicted phoneme-level prosody vectors and the vectors extracted from phoneme-level reference encoder, L_{pitch}/L_{dur} is the L1 loss between predicted pitch/duration and the ground-truth pitch/duration. For better convergence and voice quality, the output of each Conformer block in the mel-spectrogram side is projected into 80-dimension mel-spectrogram respectively, which is used to calculate mel-spectrogram loss with the ground-truth mel-spectrogram. Thus, L_{iter} is the sum of mel-spectrogram L1 loss between the predicted and ground-truth mel-spectrogram in each Conformer block [23]. For better audio fidelity, we use SSIM (Structural SIMilarity) [29] to measure the similarity between predicted and ground-truth mel-spectrogram in the final Conformer block, denoted as L_{ssim}. We use the predicted mel-spectrogram in the final Conformer block as the final mel-spectrogram prediction. Besides the provided 5 hours data, we also use an extra internal dataset from Microsoft, which contains about 40 hours European Spanish data and 40 hours Mexican Spanish data. Acoustic model and vocoder are both pretrained on the full dataset and then fine-tuned on the provided single speaker data. The Conformer module in both phoneme and mel-spectrogram sides consist of 6 improved Conformer blocks, with attention dimension 384 and hidden size of convolutional feed-forward module 1536. The kernel size in the depthwise convolution module is 7. We train DelightfulTTS on 4 NVIDIA V100 GPUs and each GPU has a batch size of about 6000 speech frames.

4. Results

The evaluation of Blizzard Challenge 2021 includes four aspects, including naturalness test with MOS (mean opinion score), similarity test with SMOS (similarity mean opinion score), SUS intelligibility test with WER (word error rate) and Sharvard intelligibility test with WER. In MOS part listeners listened to one sample and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 to 5. For SMOS they chose a response that represented how similar the synthetic voice sounded to the voice in the reference samples on a scale from 1 to 5. Sentences were designed to test the intelligibility of the synthetic speech in intelligibility test, listeners heard one utterance in each part and typed in what they heard. Listeners were allowed to listen to each sentence only once. The synthesized and natural audios were carefully rated by three types of listeners who are involved by paid listeners, online volunteers, and speech experts. Our system is denoted as F, and natural speech recorded by the original speaker is denoted as System R. Overall, the naturalness(MOS) of system F is significantly higher than all other systems, and is not significantly different to natural speech; speaker similarity(SMOS) of system F is better than all other systems.

4.1. Naturalness Test

Naturalness test evaluation results of all systems is showed in figure 2, scored by all listeners. In synthesized systems, the naturalness(MOS) of system F(mean score 4.17) is significantly higher than all other systems, and is not significantly different to

natural speech(mean score 4.21), while mean score of others is below 4.0. This shows the superiority of our end-to-end speech synthesis system over other systems.

4.2. Similarity Test

Figure 3 presents the mean opinion score of similarity evaluation results for all systems, scored by all listeners. Speaker similarity(SMOS) of system F is better than all other systems. Mean score of our system F is 4.35 while recording R is 4.07, other systems are all below 4.0.

4.3. Intelligibility of Sentences on SUS

Intelligibility test result of SUS set is presented in figure 4 by paid listeners. The SUS [30] set for this test was manually generated using the grammar structures. In this section, the result shows that our proposed system F was significantly better than system L and N, and no systems were significantly better than system F.

4.4. Intelligibility of Sentences (INT) on Sharvard

The sentences for this section came from the Sharvard corpus [31]. Note that this means the natural recordings are of a different speaker to that used to build the TTS systems. Presented in figure 5, the results show that there were no significant differences between system F and other systems. This indicates that our end-to-end acoustic model has more work to do in optimization for intelligibility test on Sharvard corpus.
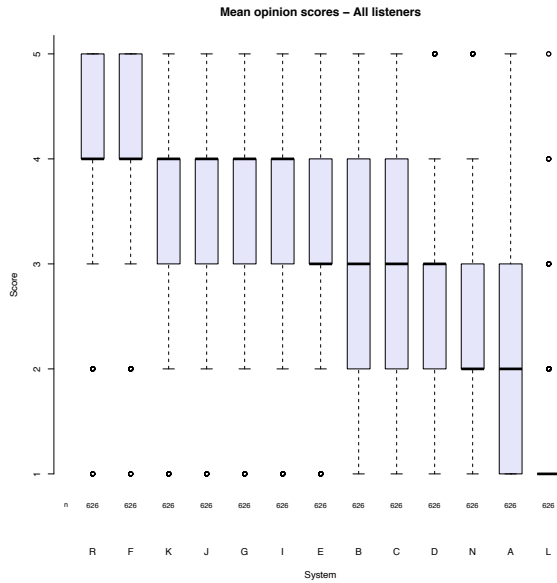
Figure 2: *Boxplot of naturalness scores of each submitted system for all listeners. Our System is identified as F.*

5. Conclusions

This paper presents the Microsoft speech synthesis system for Blizzard Challenge 2021. We approach the goal of synthesizing natural and high-quality speech from text in two aspects: The first is to directly model and generate waveform in 48 kHz sampling rate, with a good tradeoff of the task difficulties
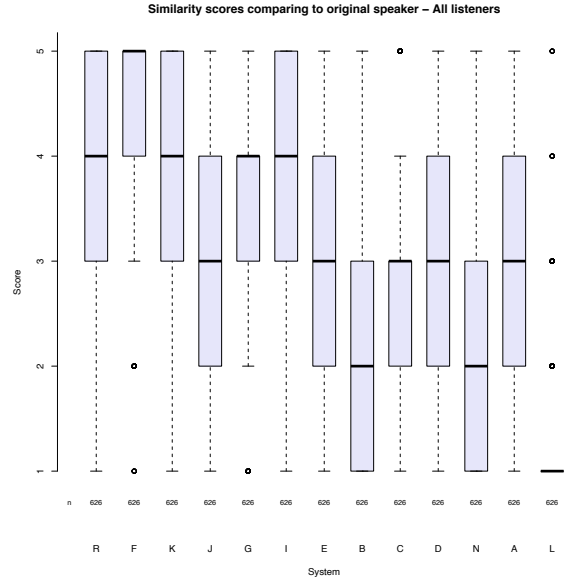
Figure 3: *Similarity scores compared to original speaker. Our System is identified as F.*

between the acoustic model and the HiFiNet vocoder, which brings higher perception quality than previous systems with lower sampling rate; The second is to model the variation information in speech through a systematic design including both explicit and implicit modeling, which improves the prosody and naturalness.
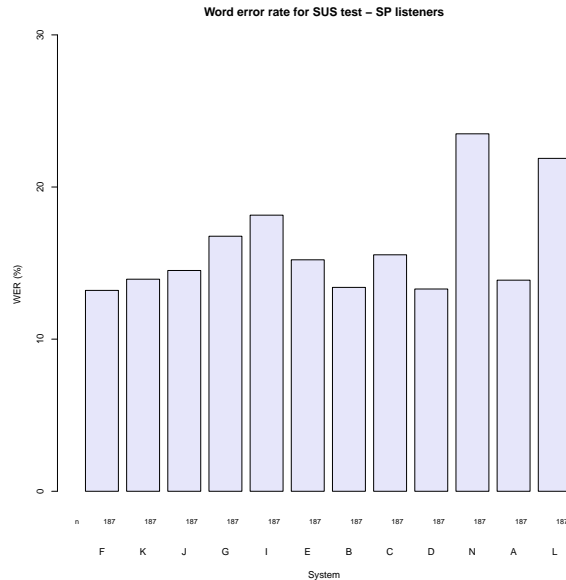
Figure 4: *Word error rates for SUS test.*

Overall, the naturalness(MOS) of system F is significantly higher than all other systems, and is not significantly different to natural speech; speaker similarity(SMOS) of system F is better than all other systems, which demonstrate the superiority of our proposed system. For future work, we will

further improve the variation information modeling on multi-speaker multi-lingual multi-style datasets and investigate the style transfer capability across different languages and speakers, and style/speaker/language disentanglement.
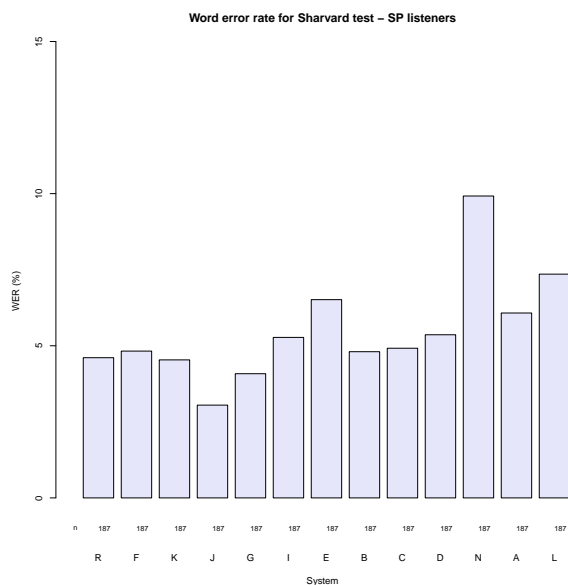

Figure 5: *Word error rates for Sharvard test.*

6. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[3] Y. Ren, C. Hu, X. Tan, 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.

[4] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu, "Adaspeech: Adaptive text to speech for custom voice," *arXiv preprint arXiv:2103.00993*, 2021.

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

[6] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.

[7] A. W. Black and K. Tokuda, "The blizzard challenge-2005: Evaluating corpus-based speech synthesis on common datasets," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[8] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 1. IEEE, 1996, pp. 373–376.

[9] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for hmm-based speech synthesis," in *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, vol. 3. IEEE, 2000, pp. 1315–1318.

[10] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[11] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, "Tts synthesis with bidirectional lstm based recurrent neural networks," in *Fifteenth annual conference of the international speech communication association*, 2014.

[12] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[13] 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*, vol. 33, no. 01, 2019, pp. 6706–6713.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[15] N. Li, Y. Liu, Y. Wu, S. Liu, S. Zhao, and M. Liu, "Robutrans: A robust transformer-based text-to-speech model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8228–8235.

[16] J. Chen, X. Tan, J. Luan, T. Qin, and T.-Y. Liu, "Hifisinger: Towards high-fidelity neural singing voice synthesis," *arXiv preprint arXiv:2009.01776*, 2020.

[17] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.

[18] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Y. Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6699–6703.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[20] Microsoft, "Azure neural tts upgraded with hifinet, achieving higher audio fidelity and faster synthesis speed," p. 0, Nov 2020, hiFiNet. [Online]. Available: https://techcommunity.microsoft.com/t5/azure-ai/azure-neural-tts-upgraded-with-hifinet-achieving-higher-audio/ba-p/1847860

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[22] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[23] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel tacotron: Non-autoregressive and controllable tts," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5709–5713.

[24] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," *arXiv preprint arXiv:1901.02860*, 2019.

[25] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[26] Y. Lu, Z. Li, D. He, Z. Sun, B. Dong, T. Qin, L. Wang, and T.-Y. Liu, "Understanding and improving transformer from a multi-particle dynamic system point of view," *arXiv preprint arXiv:1906.02762*, 2019.

[27] 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.

[28] K. Bali, P. P. Talukdar, N. S. Krishna, and A. Ramakrishnan, "Tools for the development of a hindi speech synthesis system," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[30] C. Benoît, M. Grice, and V. Hazan, "The sus test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences," *Speech communication*, vol. 18, no. 4, pp. 381–392, 1996.

[31] V. Aubanel, M. L. G. Lecumberri, and M. Cooke, "The sharvard corpus: A phonemically-balanced spanish sentence resource for audiology," *International journal of audiology*, vol. 53, no. 9, pp. 633–638, 2014.