ON-THE-FLY DATA AUGMENTATION FOR TEXT-TO-SPEECH STYLE TRANSFER

Raymond Chung[1,2] and Brian Mak[1]

[1]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
[2]Logistics and Supply Chain MultiTech R&D Centre, Pok Fu Lam, Hong Kong
{mhchungae, mak}@cse.ust.hk

ABSTRACT

Recent advanced text-to-speech (TTS) systems synthesize natural speeches. However, in many applications, it is desirable to synthesize utterances in a specific style. In this paper, we investigate synthesizing audios with three styles — newscasting, public speaking and storytelling — for a speaker who provides only neutral speech data. Firstly, considerable speech data were collected from the neutral speaker, and small amounts of speech from the wanted styles were collected from other speakers such that no speakers uttered in more than one style. All the data were used to train a basic multi-style multi-speaker TTS model. Secondly, augmented audios were created on-the-fly with the latest TTS model during its training and were used to further train the TTS model. Specifically, augmented data were created by 'forcing' a speaker to imitate stylish speeches of other three speakers by requiring their attention alignment matrices as similar as possible. Objective evaluation on the rhythm and pitch profile of the synthesized speech shows that the TTS model trained with our proposed data augmentation successfully transfers speech styles in these aspects. Subjective ABX evaluation also shows that stylish speeches synthesized by our proposed method are overwhelmingly preferred than those from a baseline TTS model by 40-60%.

Index Terms— text-to-speech, neural speech synthesis, scenario-based speech synthesis, newscasting speech, storytelling speech, public speaking speech

1. INTRODUCTION

Text-to-speech (TTS) system has a long research history. A well-known example is Stephen Hawking, who relied on a TTS system to speak for him. The voice of his TTS system speaks in the same tone in any scenarios.

In recent years, Shen et al. developed a deep learning model called Tacotron2 [1]. It is an end-to-end encoder-decoder model in which the training inputs are the raw texts and the outputs are the mel-spectrograms of the corresponding audios. An attention mechanism is used to align the decoded mel-spectrograms with the input texts. The mel-spectrograms are converted back to audios by the WaveNet [2] vocoder. This was a major break-through of the field as they showed that the mean opinion score (MOS) of their generated speeches were as natural as humans'. Besides improving the naturalness of synthesized speech, TTS research also looks into creating personalized synthetic speech with styles.

The perception of a speech to human is greatly influenced by two factors: the pitch changing profile and rhythm of speech. One may manipulate these two factors in a TTS system so as to generate speech of different styles. For instance, Robinson et al. [3] modified the fundamental frequency of an audio to convert it from a neutral speech to an emotional speech. In some other complex scenarios, we notice the following: In the newscasting scenario, utterances are usually spoken a little faster than neutral speech [4]; in the public speaking scenario, keywords may be stressed, emphasised or lengthened; in the storytelling scenario, speeches are more rhythmic.

To creating a multi-style TTS model, it is plausible to record high-quality audio training data by a voice talent who can manage to speak with different styles that we want. However, this only works for voice talents; most common people cannot speak in all wanted styles. To overcome this, we propose a novel way to train a multi-style TTS model for anyone from whom we can collect a relatively large amount of his/her speech spoken in neutral voice by generating augmented data from the person spoken in wanted styles on-the-fly. Our idea is simple: a person learns to speak expressively by imitating other people's speech. In Tacotron2 [1], the attention alignment encapsulates the rhythm of the generated speech. The augmented data from the neutral speaker should follow the same rhythm of a speaker speaking in the target style. With the additional use of scenario embedding computed from trained global style tokens (GST) [5, 6], our proposed TTS model could generate speeches which match the style of the target scenarios even with limited stylish audio training data.

Our main contribution are as follows: We show that (1) scenario-specific audio data readily found from the internet can be used for multi-style text-to-speech model training for a neutral speaker, from whom only neutral speeches are available; (2) stylish augmented training data can be generated from a neutral speaker by imitating the attention matrix of the stylish speech from another speaker on-the-fly; (3) we could use a style selection to synthesize utterances of unseen texts with a suitable rhythm and pitch profile with a neutral voice with a particular style.

2. RELATED WORK

The Tacotron2 model was very extensible. To make its output more expressive, it was extended by conditioning on the output of a global style tokens (GSTs) layer, which extracted the style information from a reference audio, so that it could generate speech with a similar global style as the reference audio [5]. Stanton et al. [7] demonstrated that the GSTs could be predicted instead from texts using a TP-GST network. However, they only conducted experiments on storytelling audios.

In the area of multi-style text-to-speech research, Prateek et al. [4] collected 20 hours of neutral speaking audios and 4 hours of newscasting audios from one speaker to train a bi-style text-to-speech model. It used one-hot style embedding and word contextual embedding as additional inputs for its decoder to generate scenario-specific audios. Hu et al. [8] recorded three speaking styles for a speaker on the same texts to create a multi-style text-to-speech model of the speaker. It used a well pre-trained speaker verification model to extract acoustic features from utterances and observed that utterances of the same style clustered together well. The TTS model then based on the mean vector of each style cluster to generate stylish utterances of the speaker. Whitehill et al. [9] proposed an adversarial cycle consistency training method for multi-reference style transfer using disjoint datasets. Speaker embeddings and style embeddings were separated trained using the scheme with paired and all possible combinations of unpaired triplets. A triplet consists of synthesizing text, and two reference audios with matched or unmatched styles. The paper shows good performance for an internal dataset consisting of only two speakers, one speaking in neutral style whereas another speaker speaking with four different emotions.

Data augmentation is very common in deep learning. In image classification, the AlexNet [10] applied on-the-fly data augmentation by translating and flipping the training image data. Zhu et al. [11] generated some virtual training samples for improving the performance of a speaker verification model. There were few works that used data augmentation in text-to-speech model training. Lee et al. [12] first pretrained a TTS model and an automatic speech recognition (ASR) model, and then applied the ASR model to improve the clarity of augmented speeches generated from unpaired combinations of GSTs and text contents. Dipjyoti et al. [13] applied a classic signal processing technique, Spectral Shaping and Dynamic Range Compression (SSDRC), to convert 2 hours of normal speech of a speaker to Lombard-style speech. Then they finetuned a Tacotron model pre-trained with LJSpeech [14] with 0.5 hours of real Lombard speech uttered by the speaker with the additional two hours of converted speech. Huybrechts et al. [15] proposed a 3-steps approaches on leveraging some augmented data. They trained a voice conversion (VC) model, CopyCat [16], to convert available stylish speech to the target speaker's voice. Then they trained their text-to-speech model with these additional augmented data. Because of the quality issue of some converted audios, they had to finetune their model with real stylish speech from the target speaker in order to obtain synthesized stylish speech with better quality. Our approach does not depend on any external ASR nor VC model; it is a single model training approach.

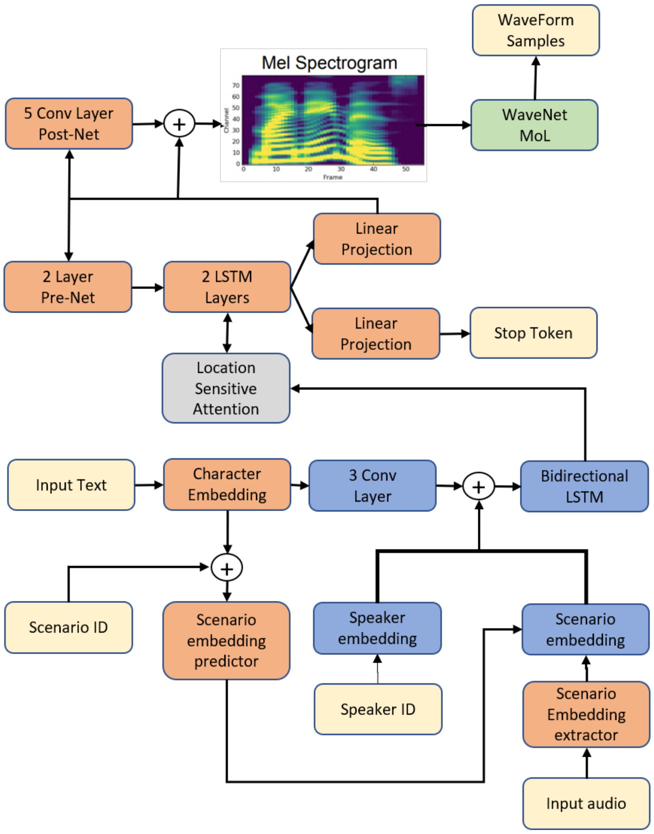

Fig. 1. Proposed Style imitated Tacotron2 model

3. PROPOSED METHOD

Our model is based on the text-predicted GST-Tacotron2 [7]. It consisted of a text encoder, a location-sensitive attention module and a decoder. We further added a trainable speaking embedding model as shown in Fig. 1. We denote the ground-truth mel-spectrogram of an audio as MEL and the Tacotron2 model as $Taco$.

$$MEL = Taco(speaker, text) \ , \tag{1}$$

where $Taco$ consists of an encoder $Taco^{enc}$ and a decoder $Taco^{dec}$. The loss function of the Tacotron2 model is defined as:

$$Loss_{taco} = |MEL - \hat{MEL}|_2 + BCE(Stop, \hat{Stop}) \ , \quad (2)$$

where $\hat{MEL}$ is the model's predicted mel-spectrogram; $Stop$ is the binary "stop token" which signals the end of decoding. Based on the length of the target audio, each output frame is assigned 0 or 1 to represent whether it is the stopping frame. The binary cross entropy loss BCE is used to predict this token.

3.1. Global style tokens module with style selection

We use the same GST module design as in [5] to compute our style embeddings $\boldsymbol{S_{gst}}$. We denote the ground-truth GST extracted from a reference audio during training as $\boldsymbol{S_{gt}}$:

$$\boldsymbol{S_{gt}} = GST(audio) \ . \quad (3)$$

As there will be no ground-truth reference audio during inference, we follow the same idea of [7] and added a text-predicted global style token network (TPGST) to get the predicted style embedding $\boldsymbol{S_{tp}}$. We modified its input by concatenating a one-hot style label with the text embedding to compute $\boldsymbol{S_{tp}}$:

$$\boldsymbol{S_{tp}} = TPGST(text, style\ label) \ . \quad (4)$$

During model training, we require the text-predicted GST as close to the ground-truth GST extracted from the corresponding reference audio as possible.

$$Loss_{tpgst} = |\boldsymbol{S_{gt}} - \boldsymbol{S_{tp}}| \ . \quad (5)$$

3.2. On-the-fly data augmentation

In the original Tacotron2 training process, an attention mechanism learns an alignment matrix for the alignment relationship between the text input and the generated audio frames. We will use $\boldsymbol{h_{text}}$ to represent the hidden state sequence of an input text, $\boldsymbol{h_p}$ to represent the trainable embedding of a speaker p, $\boldsymbol{S_p}$ to represent the style embedding of a speech from speaker p, $\boldsymbol{A_p}$ to represent the corresponding alignment matrix produced by the Tacotron2. During our Tacotron2 encoding step, its encoder generates the hidden state sequence, the style embedding as well as the speaker embedding for the given inputs as follows:

$$\boldsymbol{h_{text}}, \boldsymbol{S_p}, \boldsymbol{h_p} = Taco^{enc}(speaker, audio, text) \ . \quad (6)$$

During our Tacotron2 decoding step, its decoder takes the outputs from its encoder, and predicts the output mel spectrograms using autoregression with the attention mechanism:

$$\boldsymbol{MEL_p}, \boldsymbol{A_p} = Taco^{dec}(\boldsymbol{h_{text}}, \boldsymbol{S_p}, \boldsymbol{h_p}) \ . \quad (7)$$

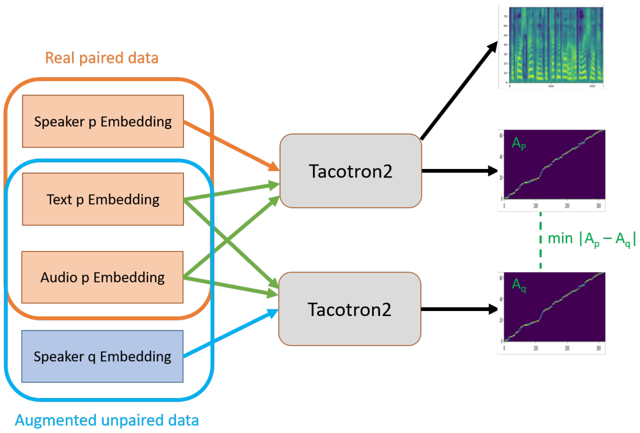

Fig. 2. Proposed on-the-fly data augmentation scheme

A by-product of decoding is the attention alignment matrix, which is usually discarded afterwards in standard Tacotron2 training. We argue that the alignment matrix actually encapsulates useful rhythmic information of the input text/audio which can capture a speaking style. In this paper, we generate augmented data for any speaker q by having him/her imitate the alignment matrix $\boldsymbol{A_p}$ of another speaker p given the text and audio from the latter speaking in style $\boldsymbol{S_p}$ which speaker q does not speak with. That is, if we denote the mel spectrograms of the augmented data from speaker q as $\boldsymbol{MEL_q}$, and the corresponding alignment matrix as $\boldsymbol{A_q}$, we will have

$$\boldsymbol{MEL_q}, \boldsymbol{A_q} = Taco^{dec}(\boldsymbol{h_{text}}, \boldsymbol{S_p}, \boldsymbol{h_q}) \quad (8)$$

As the augmented data may not be perfect, we focus only on its style property captured by the alignment matrices, and only incorporate the alignment loss for augmented data (for any unpaired speaker and style combinations) during training, which is defined as the Frobenius norm of the difference between the two alignment matrices as follows:

$$Loss_{align} = |\boldsymbol{A_p} - \boldsymbol{A_q}|_2 \quad (9)$$

A similar alignment loss was also utilized by [17] in their guided attention method with pre-aligned phoneme sequences to speed up and stabilize their Tacotron2 training.

Note that this data augmentation is performed for any two different speakers speaking in different styles on all training utterances from all speakers from disjoint datasets. This data augmentation scheme is further illustrated in Fig. 2. In the figure, the real paired data contain the speaker embedding of speaker p, together with text and GST extracted from his/her training utterance. The augmented unpaired data is derived from the same text and GST but using another speaker embedding from, say, speaker q. Consequently, in a mini-batch update, our model will be trained with the real sample pairs and augmented sample pairs.

In this paper, we mainly investigate the transfer from a neutral voice to one of the three styles (newscasting, public speaking and storytelling) for TTS. There are two choices for the data augmentation scheme: (1) *data augmentation only for the neutral speaker*: generate stylish augmented data only from originally non-neutral stylish speeches with speaker q being the neutral speaker, Michelle Obama; (2) *data augmentation for all speakers*: generate an augmented sample for any original speech spoken by speaker p for each speaker $q \neq p$, resulting in three additional augmented samples for each original audio. We did both ways and compared their results.

Finally, the overall loss function for training our proposed model is the sum of the three losses in predicting the output mel spectrograms, GSTs and alignments:

$$Loss = Loss_{taco} + Loss_{tpgst} + Loss_{align} \; . \qquad (10)$$

4. EXPERIMENTS AND RESULTS

4.1. Dataset

We did not find a voice talent to record audio data of various styles. Instead, we collected from audiobooks or from the Web existing audio data which were spoken under four realistic scenarios: neutral speaking, newscasting, public speaking and storytelling. Specifically, to get a large amount of neutral speaking audios, we purchased an audiobook from Google Play read by Michelle Obama. Her speaking style was considered neutral in reading the audiobook. We then collected some newscasting audios of a news anchorwoman from the Voice of America (VOA) website. Some Hillary Clinton's public speeches and presidential debate videos on the YouTube were collected for the public speaking style. Finally, the children audiobooks from the Blizzard Challenge 2017 were used for the storytelling style.

All the collected data were re-sampled at 16000Hz sampling rate, and encoded in 16-bit PCM WAV format. The online Google speech-to-text engine was employed to obtain the orthographic transcriptions of the collected audios. The Montreal Forced Aligner [18] was further used to obtain the word alignment between each audio and its transcript. Finally, all the audios were segmented into short clips of 2 to 10 seconds which were found to give more stable Tacotron2 training. The final amounts of collected data of the four wanted styles are summarized in Table 1. For testing, we held out 50 utterances from Michelle Obama's data, and 30 utterances from each of the other three stylish speakers' data.

4.2. Experimental Setup

We modified the codebase [1] of Mellotron [19] to implement our idea. The training audios were sampled at 16000Hz. The mel-spectrograms were computed with a FFT size of 800, a hop length of 12.5ms and a window size of 50ms. The dimension of our speaker embeddings was 128. We started with the pre-trained model of the LibriTTS dataset [20] and trained our proposed model with the prepared training data of Table 1. Since each stylish speaker had less audio data than the neutral speaker's, the audio data of each stylish speaker were repeated in each epoch to around the same amount (around 12 hours) as the neutral speaker's to tackle the data imbalance issue. We used a min-batch size of 64 and trained our model for 80 epochs using the Adam optimizer [21] with a learning rate of 5×10^{-4}. Afterwards, we continued our model training with the original training data together with on-the-fly augmented data for another 120 epochs. Finally, We used a pre-trained WaveGlow [22] vocoder [2] to convert the mel-spectograms generated by our model to their audios. (Some audio samples could be found on this webpage. [3])

Speaker	Scenario/Style	Audio data (hr)
Michelle Obama	Neutral	11.7
VOA	Newscasting	2.1
Hillary Clinton	Public speaking	2.5
Blizzard 2017	Storytelling	3.5

Table 1. Summary of collected audios in four speaking styles.

	Speaking Rate	
Speaker	Original	Newscasting
Michelle Obama	13.7	15.8
VOA	16.3	
Hillary Clinton	14.3	16.3
Blizzard 2017	15.0	15.7

Table 2. Speaking rate of the speakers in their original stylish audios and synthetic newscasting audios (of unseen news texts).

For evaluation, we compared the synthesized audios from our proposed model with those from a baseline model which was trained with audios only from the neutral speaker, Michelle Obama. The baseline model is equivalent to a single-speaker text-predicted GST Tacotron2 [7].

4.3. Objective Evaluation

We evaluated the effectiveness of the proposed on-the-fly data augmentation on its effects on the rhythm and pitch profile of the generated stylish audios. Here, data augmentation was performed on all speakers (2nd choice in Section 3.2).

[1] https://github.com/NVIDIA/mellotron

[2] Despite that the WaveGlow vocoder was trained with 22050Hz audios and our proposed model was trained with 16000Hz audios, by matching the FFT and filterbank parameters of the two models, we could reuse the pre-trained WaveGlow model to produce good synthetic speech in our system.

[3] https://raymond00000.github.io/ttsdemo.html

Scenario	Newscasting	Public speaking	Storytelling
F0 change	+16.3	+22.8	+10.1

Table 3. Average increment of mean F0 when the same text was spoken with a style vs. spoken with a neutral voice.

4.3.1. Effect on the Rhythm of Style Transfer to the Newscasting Style

We roughly measured the rhythm of a speech by the speaking rate which is defined as the number of phonemes spoken per second. Utterances were synthesized for each speaker, other than the VOA anchorwoman, from the held-out unseen news texts with the newscasting style using the TTS model trained with the proposed data augmentation for all speaker pairs, and their rhythms were compared with the speakers' speaking rate in their original audios. The results are shown in Table 2. Firstly, we notice that the speaking rate of the newscaster is faster than all the other speakers in their original audios. Secondly, our model successfully synthesizes newscasting audios for other speakers with a speaking rate close to the newscaster's which is greater than the original speaking rate of the speakers. We only performed the rhythm evaluation with newscasting as the target style because from Table 2, we noticed that the speaking rate of newscasting style is significantly greater than that of all the other styles, whereas the speaking rates of the other styles are similar to each other.

4.3.2. Effect on the Pitch

Synthetic audios were generated by our proposed TTS model using Michelle Obama's voice from the held-out unseen texts of each style in that style and neutral style. The mean fundamental frequencies (F0) of each pair of generated utterances were computed and compared in Fig. 3 – 5, and their average difference in each stylish test set is summarized in Table 3.

We observe that the mean F0 is higher in all the other three styles than in neutral speech even though the utterances were all generated from the same speaker, Michelle Obama. Among the three non-neutral styles, public speech has the highest mean F0, while storytelling speech has the lowest mean F0. The same trend is found on the original datasets of the four styles: public speech > newscasting speech > storytelling speech > neutral speech although their texts are different.

4.4. Subjective Evaluation

We performed a subjective ABX evaluation on utterances synthesized by the proposed TTS model and the baseline TTS model via the Amazon Mechanical Turk. (Again, the baseline TTS model was trained with only training data from the neutral speaker, Michelle Obama.) For each scenario, we used the neutral speaker's voice to generate utterances of 15 sentences

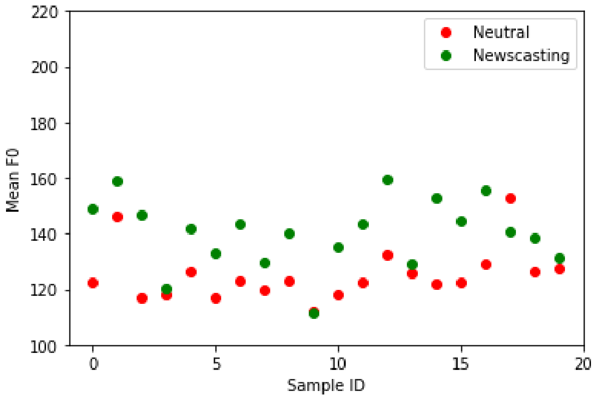

Fig. 3. Mean F0 comparison: newscasting vs. neutral style.

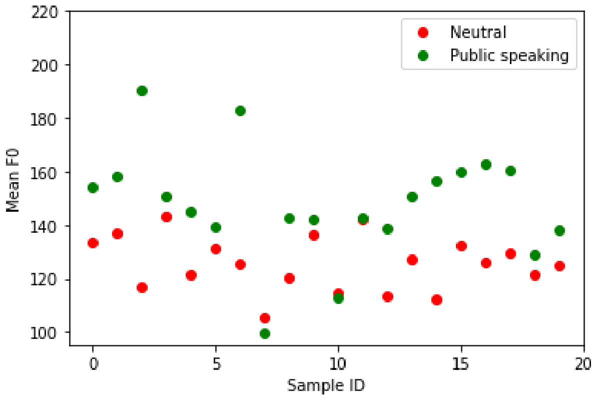

Fig. 4. Mean F0 comparison: public speaking vs. neutral style.

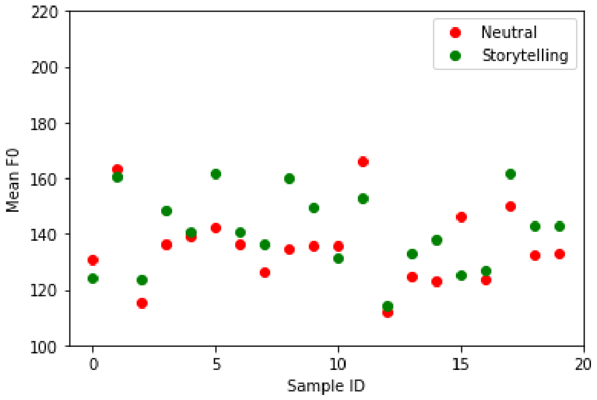

Fig. 5. Mean F0 comparison: storytelling vs. neutral style.

Scenario	Proposed Model	Baseline Model	No Preference
Newscasting	69%	24%	7%
Public speaking	66%	27%	7%
Storytelling	68%	24%	8%

Table 4. ABX results with data augmentation only for the neutral speaker.

Scenario	Naturalness	Intelligibility
Newscasting	3.79±0.15	3.95 ±0.13
Public speaking	3.56±0.10	3.61±0.09
Storytelling	3.74±0.10	3.71±0.12

Table 5. MOS results with data augmentation only for the neutral speaker at 95% confidence level.

in her neutral and scenario style. For each sentence, two synthesized utterances from the two TTS models were played, and the listeners were asked which one they preferred for the desired scenario. The listeners also rated naturalness and intelligibility of the synthesized utterances on a scale from 1 to 5. Fifteen listeners were recruited to do the evaluation for each scenario.

As discussed in Section 3.2, there are two options for data augmentation: (1) data augmentation only for the neutral speaker, Michelle Obama, or (2) data augmentation for all speakers. The preference results and mean opinion scores (MOS) on naturalness and intelligibility of synthesized speech produced by option (1) are shown in Table 4 and Table 5, whereas the results of option (2) are shown in Table 6 and Table 7. From Table 4 and Table 6, we find that generating augmented data with all speakers in various styles may cause more confusions as there are more counts of “no preference”, but there are more subjects preferring our proposed model over the baseline model for style transfer to newscasting style (by 60%), public speaking style (by 46%), and storytelling style (by 41%). The effect is particularly strong for the style transfer to newscasting speech. There can be two reasons for that: (a) data augmentation option 2 generates three times more augmented data for training our TTS models, resulting in a better model; (b) even though our ultimate goal is only to generate synthesized speech from one neutral speaker, Michelle Obama in our case, in other three non-neutral styles, data augmentation option 2 renders our model training method a multi-task learning (MTL) method which tries to learn conversion between any two of the four speaking styles. MTL gives an inductive bias that results in a more robust model for future unseen data.

Scenario	Proposed Model	Baseline Model	No Preference
Newscasting	74%	14%	12%
Public speaking	67%	21%	12%
Storytelling	69%	28%	3%

Table 6. ABX results with data augmentation for all speakers.

Scenario	Naturalness	Intelligibility
Newscasting	3.79±0.08	3.99 ±0.08
Public speaking	3.33±0.11	3.67±0.08
Storytelling	3.80±0.11	3.92±0.12

Table 7. MOS results with data augmentation for all speakers at 95% confidence level.

Table 5 and Table 7 give the MOS results on the naturalness and intelligibility of the generated stylish utterances; they are all acceptable. We observe public speaking speeches have a lower naturalness and intelligibility compared to newscasting and storytelling speeches. A plausible explanation is that the public speech data from Hillary Clinton were obtained from YouTube which have a lower sound quality. Overall speech synthesized by our TTS model has naturalness that is comparable to the reported audio signal quality of the voice-cloning-augmentation TTS model in [15].

5. CONCLUSION

To the best of our knowledge, our proposed method is the first method that builds a multi-style text-to-speech model from a set of single-style disjoint datasets, each spoken by a different speaker. This is made possible by generating augmented speech data for the imitating speaker, and using a loss function over the alignment matrices (from the attention module in the model) of the imitating speaker and the original speaker. The data augmentation is done on-the-fly to utilize the latest model parameters for the data generation. Another benefit is our on-the-fly augmented unpaired data share some embedding variables with the real paired data hence the GPU RAM could be potentially optimized for a larger mini-batch size during training. Objective evaluation on the rhythm and pitch profile of the synthesized stylish speeches shows that our proposed model successfully performs the style transfer. Subjective evaluation also shows that stylish speech generated by our proposed model are overwhelmingly preferred over the baseline model that was trained to generate neutral speech.

6. ACKNOWLEDGEMENTS

This work was supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. HKUST16200118) and the Innovation and Technology Fund of the Hong Kong SAR, China (Project No. ITP/052/19LP).

References

[1] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, and Rj Skerrv-Ryan, “Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 4779–4783, IEEE.

[2] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” *arXiv preprint arXiv:1609.03499*, 2016.

[3] Carl Robinson, Nicolas Obin, and Axel Roebel, “Sequence-to-sequence modelling of F0 for speech emotion conversion,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6830–6834.

[4] Nishant Prateek, Mateusz Lajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, and Trevor Wood, “In other news: A bi-style text-to-speech model for synthesizing newscaster voice with limited data,” *arXiv preprint arXiv:1904.02790*, 2019.

[5] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *International Conference on Machine Learning*. 2018, pp. 5180–5189, PMLR.

[6] Jae-Sung Bae, Hanbin Bae, Young-Sun Joo, Junmo Lee, Gyeong-Hoon Lee, and Hoon-Young Cho, “Speaking speed control of end-to-end speech synthesis using sentence-level conditioning,” *arXiv preprint arXiv:2007.15281*, 2020.

[7] Daisy Stanton, Yuxuan Wang, and RJ 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.

[8] Qiong Hu, Tobias Bleisch, Petko Petkov, Tuomo Raitio, Erik Marchi, and Varun Lakshminarasimhan, “Whispered and Lombard neural speech synthesis,” in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 454–461.

[9] Matt Whitehill, Shuang Ma, Daniel McDuff, and Yale Song, “Multi-reference neural TTS stylization with adversarial cycle consistency,” *arXiv preprint arXiv:1910.11958*, 2019.

[10] Geoffrey E Hinton, Alex Krizhevsky, and Ilya Sutskever, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1106–1114, 2012.

[11] Yingke Zhu, Tom Ko, and Brian Mak, “Mixup learning strategies for text-independent speaker verification.,” in *Interspeech*, 2019, pp. 4345–4349.

[12] Da-Rong Liu, Chi-Yu Yang, Szu-Lin Wu, and Hung-Yi Lee, “Improving unsupervised style transfer in end-to-end speech synthesis with end-to-end speech recognition,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 640–647.

[13] Dipjyoti Paul, Muhammed PV Shifas, Yannis Pantazis, and Yannis Stylianou, “Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion,” *arXiv preprint arXiv:2008.05809*, 2020.

[14] Keith Ito and Linda Johnson, “The LJ Speech dataset,” https://keithito.com/LJ-Speech-Dataset/, 2017.

[15] Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, and Jaime Lorenzo-Trueba, “Low-resource expressive text-to-speech using data augmentation,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6593–6597.

[16] Sri Karlapati, Alexis Moinet, Arnaud Joly, Viacheslav Klimkov, Daniel Sáez-Trigueros, and Thomas Drugman, “Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech,” *arXiv preprint arXiv:2004.14617*, 2020.

[17] Xiaolian Zhu, Yuchao Zhang, Shan Yang, Liumeng Xue, and Lei Xie, “Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis,” *IEEE Access*, vol. 7, pp. 65955–65964, 2019.

[18] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, “Montreal Forced Aligner: Trainable text-speech alignment using kaldi.,” in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[19] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro, “Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6189–6193, IEEE.

[20] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu, “LibriTTS: A corpus derived from librispeech for text-to-speech,” *arXiv preprint arXiv:1904.02882*, 2019.

[21] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[22] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.