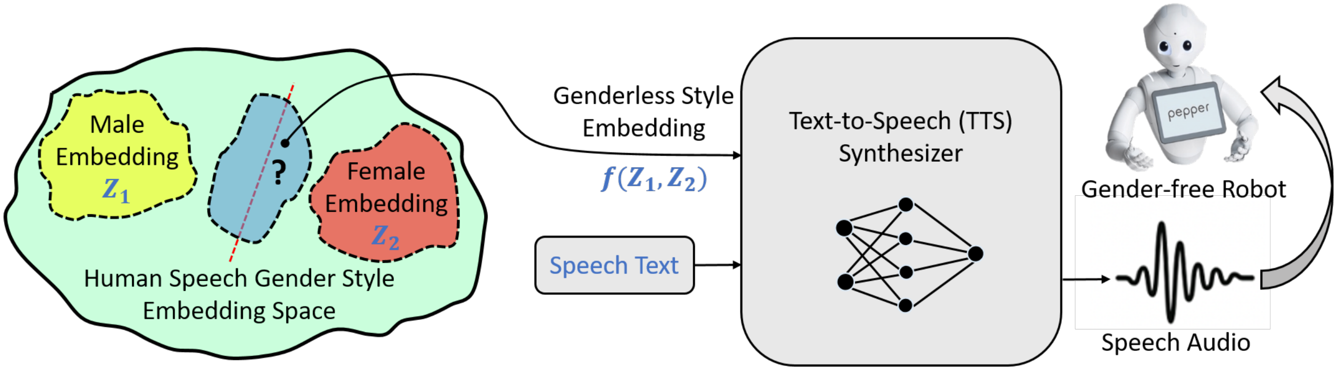
First Attempt of Gender-free Speech Style Transfer for Genderless Robot

Chuang Yu
Cognitive Robotics Lab
University of Manchester
Manchester, United Kingdom
chuang.yu@manchester.ac.uk

Changzeng Fu
Intelligent Robotics Lab
Osaka University
Osaka, Japan
changzeng.fu@irl.sys.es.osaka-u.ac.jp

Rui Chen
Human-Robot Interaction Lab
Kyoto University
Kyoto, Japan
rui.chen.32z@st.kyoto-u.ac.jp

Adriana Tapus
Autonomous Systems and Robotics Lab, U2IS
ENSTA Paris, Institut Polytechnique de Paris
Palaiseau, France
adriana.tapus}@ensta-paris.fr



The pipeline of gender-free robot speech synthesis. The text-to-speech (TTS) synthesizer takes the genderless speech style embedding and text as inputs to output the gender-free voice, which can be used in the genderless robot, for example, the Pepper robot. The genderless speech style embedding is a function of the male and female speech style embedding distributions. The male and female embedding are extracted from a speaker gender encoder. The speaker gender encoder is independent from TTS synthesizer during training. Z_1 is the male speech style embedding and Z_2 is the female speech style embedding. $f(Z_1, Z_2)$ is the gender-free speech style embedding, which is obtained from the existed female and male speech style embedding.

***Abstract*—Some robots for human-robot interaction are designed with female or male physical appearance. Other robots are endowed with no gender characteristics, namely genderless robots, such as Pepper and NAO robot. A robot with male or female physical appearance should possess the mapped speech gender style during a natural human-robot interaction, which can be learned from humans' male or female speech. In this paper, we make a new trial to synthesis gender-free speeches for physically genderless robots, which is promising in order to improve a more natural human-robot interaction with genderless robots. Our gender style-controlled speech synthesizer takes the speech text and gender style embedding as inputs to generate speech audio. A speech gender encoder network is used to extract the embedding of the speech gender style with female and male speeches as input. Based on the distribution of the female and male gender style embedding, we explore the gender-free speech style embedding space where we sample some gender-free embedding vectors to generate genderless speech audio. This is a preliminary work where we show how the genderless speech audio wave will be synthesized from text.**

***Index Terms*—genderless robot speech, speech style transfer, text-to-speech synthesis**

I. Introduction

Gender is an essential characteristic of humans. How about robot gender identity? Robot gender characteristics in robotics design, including robot behavior design and appearance design, make a big difference in human-robot interaction [1]. Several works show that robot's behaviors, including verbal and non-verbal behaviors, should match its gender to make an appropriate human-robot interaction [2] [3]. How about the genderless robots, for example, the Pepper robot and NAO robot, who have no clear gender identity [4] [5]?

It is not suitable to use the speech with male or female style on a genderless robot. We explore here the area of gender-free speech synthesis for the genderless robots. In the paper, the terminology $genderless$ and $gender-free$ mean that a robot does not identify with male or female and that it is difficult or not possible to recognize it as male or female from the generated speech.

Recently, supervised and unsupervised text-to-speech (TTS)

978-1-6654-0731-1/22/$31.00 ©2022 IEEE

synthesis and speech style transfer have attracted many AI and Robotics researchers' attention to the development of speech processing research and Natural Language Processing (NLP) research [6] [7] [8] [9]. We can use these technologies to do speech gender style transfer and male or female speech generation with human speech recording. However, it is not easy to get gender-free speech samples recorded from humans. Therefore, it is very challenging to get genderless embedding for gender-free speech generation as there are no related databases. Furthermore, the relation between male/female speech style embedding distribution and the distribution of genderless speech embedding distribution is still not clear, which makes the task even more difficult.

This paper presents a preliminary work on synthesizing the gender-free speech style with text and gender style embedding as inputs. The pipeline, as shown in the Figure . The model contains two parts. The first part is genderless speech style embedding with a speaker gender encoder, which can extract the speech gender style embedding from male or female speeches. The speaker gender encoder is independent of the TTS synthesizer training. The female and male embeddings are applied to look for the genderless style embedding. Namely, the genderless speech style embedding is the function of the male and female speech style embedding distributions. The second part is the text-to-speech (TTS) synthesizer that takes the genderless speech style embedding generated in the first part and the speech text as inputs in order to output the gender-free speech audio.

The rest of the paper is organized as follows: Section II presents the related works. Section III shows the methodology of the robot gender-free speech synthesis. The results are shown in Section IV. The discussion and future works are part of Section V.

II. Related works

Text-to-speech (TTS) generation research has made significant progress recently with some main methods, for example, Tacotron [9], Tacotron 2 [10], FastSpeech [7], and so on. Furthermore, style-controlled TTS is a new trend [11] [12].

The authors in [9] came up with a new method for style controllable speech generation from speech text. It is the first time when the concept of ”style tokens” is introduced to Tacotron, which is widely used [13] [14]. The style tokens, a kind of independent prosodic styles, are extracted from training data with style attention networks.

In [11] a speaker-style controllable text-to-speech generation model is presented. Their model contains three independently trained models: a speaker encoder network, a sequence-to-sequence synthesis network based on Tacotron 2, and an auto-regressive WaveNet-based vocoder network [15]. The Speaker Encoder is used for the speaker style embedding, which computes a fixed dimensional vector from a speech signal. Furthermore, the style encoder network is the head part of a speaker verification network [16]. In this paper, a good performance on the high-quality speaker style representation is obtained, which lets the TTS model synthesis natural and new speech never seen in the training process.

The reference [12] introduced the Variational Autoencoder (VAE) to the TTS model for style-controllable speech synthesis from text. The VAE model is for learning latent style representation of speaker styles with an unsupervised learning method. It is a recognition model or inference network, which encodes reference audio into a fixed-length short vector of latent representation, which stands for style representation. The TTS model is based on the Tacotron 2, which converts the speech style embedding (extracted from reference speech audio) and the speech text to the target speech with a specific style. In order to overcome the KL collapse problem, the paper uses the KL annealing trick and makes the KL loss be considered every k steps. The model also shows an ability of disentanglement to some degree. Namely, some latent features of the speaker style embedding can be used to control one of the specific speech styles independently. Finally, the paper shows a good performance on the style control. The paper also inspired us to deal with the high-quality gender style embedding problem of speech audio.

About genderless voice synthesis, the institute $EqualAI$ created the genderless voice based on frequency manipulation [17]. However, the rule-based method is limited because the gender representation in voice contains a lot of properties including vocal intensity(loudness), frequency(pitch), and so on. Our paper uses a deep learning method to extract gender-related voice representation that can contain more gender information in speech, which leads to a more expressive speech generation.

III. Methodology

A. Gender-free TTS architecture

In our task of gender-free speech generation from the text, the gender-free speech means that humans cannot recognize the gender when hearing the speech audios. Our model contains three sub-models, including a speech gender encoder, a TTS synthesizer, and a $Vocoder$ network and a rule-based genderless speech style embedding extraction model. These first three submodels are trained independently. The speech gender encoder takes female and male speech samples to get the fixed dimensional vectors as the speech gender style embedding, respectively. The rule-based genderless speech style embedding extraction model uses the female and male speech style embedding from the speech gender encoder to get the gender-free speech embedding. The TTS synthesizer is a $seq2seq$ model with attention based on Tacotron 2. It takes word embedding of speech text and the condition of gender-free style embedding. The TTS synthesizer to synthesize the Mel spectrograms, which is fed to the Neutral $Vocoder$ network based on WaveNet [15] to generate the gender-free speech audio finally.

B. Gender-free speech style extraction

Actually, our speech gender encoder is a speaker gender recognition model with speech audio, which is built based on one excellent speaker certification model in reference [18]. Before being fed to the speech gender encoder, the reference speech audio is processed towards the 40-channel log-mel spectrograms. Then, the mel spectrograms are input into a stack of 3 LSTM layers of 768 cells in the speech and the following multilayer perceptron with 256-dimensional output in speech gender encoder. The 256-dimensional latent features of the last time step output in the last LSTM are the gender style presentation of reference speech audio.

We trained the speech gender encoder model on the open database LibriTTS [19] but with the gender labels. Because the encoder is a gender recognition model, the fixed dimensional style embedding vectors of the same gender cluster in a similar area in the speaker embedding feature space. As shown in Figure , the female and male speech style embeddings cluster in two areas (male: yellow female: red) and belong to two distributions, respectively. The gender-free embedding area/space should be far from female and male embedding areas, and mostly, it is near the blue region of embedding space. How can we sample from the gender-free speech embedding distribution in the possible blue region? The only way is to use the existed male and female speech latent representation distributions to achieve the goal, as shown in Equation 1. Where, Z_1 obeys the distribution of the male speech style embedding, Z_2 obeys the distribution of the female speech style embedding and Z_0 obeys the distribution of the gender-free speech style embedding.

$$Z_0 = f\left(Z_1, Z_2\right) \tag{1}$$

To simplify the calculation, we only use the linear f function in this paper, as shown in Equation 2. Where, $0 \leq a \leq 1$ and $0 \leq b \leq 1$. In future work, we plan exploring more complex and precise modes.

$$Z_0 = a \cdot Z_1 + b \cdot Z_2 \tag{2}$$

IV. Results

LibriTTS [19] database is used during the training process of the speech gender style encoder. LibriTTS database and VCTK [20] database are used during the TTS synthesizer training. However, we update all the labels with genders (0: female; 1: male) for the database above. The extracted speech gender style embedding from the trained speech gender encoder is 256-dimensional vectors. After Principal Component Analysis (PCA) processing, the 3-dimensional embedding visualization of the speech gender style embedding is shown in Figure 1. Each color corresponds to a gender style. The red is for the female, while green is for the male. From the Figure, we can see that the female speech embedding and male speech embedding cluster are in two distinct regions, which certify our initial hypothesis. The middle zone between the green cluster and the red cluster is the gender-free speech style embedding space, which makes it is possible to sample the speech gender embedding in the whole style representation space. This paper shows some preliminary work on our gender-free speech generation project. The hyperparameter is being adjusted based on the training experiments. The weight $[a, b]$ is equal to $[0.25, 0.75]$, $[0.5, 0.5]$ and $[0.75, 0.25]$ in Equation 2 to generate some speech audios. The Z_1 and Z_2 use the mean vector of female speech style embedding and male speech style embedding, respectively, to try exploring the gender-free speech style embedding space. However, the generated speech is still not natural and with noise.

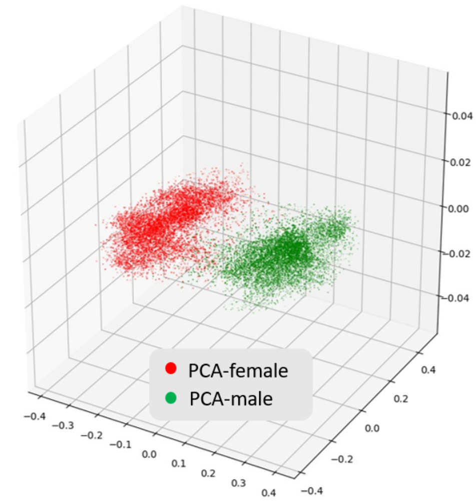


Fig. 1. **Visualization of speech gender style embedding extracted from LibriTTS [19] database. The original embedding dimension is 256. Here the 3-dimensional features are extracted by Principal Component Analysis (PCA) processing. Each color denotes a gender style. The red is for female, while green is for male.**

V. Discussion and Future Works

The work is still at a preliminary stage, and there are many developments needed to be completed in the future. However, the first result is promising for further exploration in the robot gender-free speech synthesis. This paper shows a basic solution to generate gender-free speech audio from text with genderless speech style embedding as the condition. We trained a speech gender recognition model for speech gender style embedding. The visualization of the female and male embedding shows that the two kinds of embeddings are classifiable, and the middle zone exists between the male embedding and the female embedding. Besides, we simplified the projection process from existed male/female embedding distributions to unseen gender-free embedding distribution to sample some style embedding and get some promising results.

Future work will focus on looking for a more suitable method to get the approximate distribution of gender-free speech style embedding using the existed distributions of female and male embedding. We will also focus on getting better gender-free speech results and use them with a genderless robot in a real human-robot interaction scenario. We can then explore the effects of the human perception of the genderless robot.

References

[1] T. Nomura, “Robots and gender,” *Gender and the Genome*, vol. 1, no. 1, pp. 18–25, 2017.

[2] B. Tay, Y. Jung, and T. Park, “When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction,” *Computers in Human Behavior*, vol. 38, pp. 75–84, 2014.

[3] J. Carpenter, J. M. Davis, N. Erwin-Stewart, T. R. Lee, J. D. Bransford, and N. Vye, “Gender representation and humanoid robots designed for domestic use,” *International Journal of Social Robotics*, vol. 1, no. 3, p. 261, 2009.

[4] R. A. Søraa, “Mechanical genders: how do humans gender robots?” *Gender, Technology and Development*, vol. 21, no. 1-2, pp. 99–115, 2017.

[5] Pepper is a humanoid-robot. [Accessed 12-Jan-2022]. [Online]. Available: https://developer.softbankrobotics.com/pepper-qisdk/design/pepper-humanoid-robot

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, “Tacotron: Towards end-to-end speech synthesis,” *arXiv preprint arXiv:1703.10135*, 2017.

[7] 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.

[8] N. Tits, F. Wang, K. E. Haddad, V. Pagel, and T. Dutoit, “Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis,” *arXiv preprint arXiv:1903.11570*, 2019.

[9] Y. Wang, R. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, “Uncovering latent style factors for expressive speech synthesis,” *arXiv preprint arXiv:1711.00520*, 2017.

[10] 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.

[11] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, “Transfer learning from speaker verification to multispeaker text-to-speech synthesis,” *arXiv preprint arXiv:1806.04558*, 2018.

[12] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 6945–6949.

[13] 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.

[14] 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.

[15] 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.

[16] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, “End-to-end text-dependent speaker verification,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2016, pp. 5115–5119.

[17] Meet q: The first genderless voice. [Accessed: 12-Jan-2022]. [Online]. Available: https://www.genderlessvoice.com/

[18] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2018, pp. 4879–4883.

[19] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “Libritts: A corpus derived from librispeech for text-to-speech,” *arXiv preprint arXiv:1904.02882*, 2019.

[20] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, “Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit,” 2016.