HILvoice:Human-in-the-Loop Style Selection for Elder-Facing Speech Synthesis

Xueyuan Chen[1], *Qiaochu Huang*[1], *Xixin Wu*[2,*], *Zhiyong Wu*[1,2,*], *Helen Meng*[2]

[1] Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] The Chinese University of Hong Kong, Hong Kong SAR, China

{chenxuey20, hqc22}@mails.tsinghua.edu.cn, {wuxx,zywu,hmmeng}@se.cuhk.edu.hk

Abstract

Controllable speech synthesis has made great progresses over the last decades. State-of-the-art systems can provide flexible interfaces for configuring the styles of generated speech for target users. However, for a specific user group, e.g., the older adults, using the available configuration interfaces to select the styles that are favored by the group still needs to be investigated. Two main questions of such a style selection are (i) how to provide various options for the target users to pick; and (ii) how to effectively obtain the opinions from the target users. Since these two questions are highly correlated which makes it difficult to solve them separately, we propose a holistic framework to consider these two questions together by involving the target users in an iterative loop. We demonstrate by experimental results that the proposed framework can successfully select a speaking style preferred by the older adults than the default neutral setting. Analysis results show that the selected style has slower speaking rate, which coincides with previous studies on auditory perception of older adults.

Index Terms: speech synthesis, human in the loop, elder-facing, controlability

1. Introduction

Controllable text-to-speech (TTS) synthesis has made great progresses over the last decades [1–9]. Various approaches have been proposed to provide flexible controlling interfaces for configuring speaking styles of generated speech, based on reference exemplars [4, 7], or on acoustic features [5, 9, 10], etc. Later developments also improve from coarse utterance-level control to finer-grain control on words, phonemes and frames [6, 8, 11]. When these control techniques are applied to a specific scenario or for a specific target user group, one problem is how to determine the desired characteristics of speaking style, so that the techniques can be utilized to configure the TTS systems to generate the corresponding characteristics. Solving this problem is non-trivial, especially when the characteristics of the target users, e.g., older adults, are not well understood by the TTS system developers. We advocate that the TTS system developers and the system users should be involved into the same working loop to together determine the speech characteristics that are preferred by the users and, at the same time, can be rendered by the available control techniques [12].

Traditional TTS system development relies on the selection of speakers with the desired characteristics and requires the speaker to record utterances in a consistent speaking style [13, 14]. However, recruiting the speakers for a specific target user group is difficult and costly. Actually the speaker recruitment and recording control require understanding of the characteristics desired by the target group at the first place.

Recent development of controllable TTS enables another option for style selection. These TTS systems can be controlled to synthesize various speaking styles with different speaking rates, pitch and intensity levels, by presenting an exemplar carrying the desired style(s) to the TTS systems, or directly manipulating the corresponding neural embeddings. A real-time modification system [15] based on these controlling methods can be developed to obtain opinions from individuals of the target user group. However, the modification of the real-time system requires background knowledge, which the target users may not have. The aggregation of the individual opinions and the according optimization of system parameters towards the whole target group are rarely studied in previous works.

In this work, we study the synthesis style selection for the older adults group and propose a human-in-the-loop framework, called HILvoice, based on the sequential model-based optimization (SMBO) algorithm. An iterative process is adopted to gradually update TTS model parameters based on subjective preference feedback from participating target users. In the forward path, the model with certain parameters generates speech samples and some other samples with perceptible difference for comparison. The users provide their preference opinions on the comparison pairs. In the backward path, the model parameters are updated towards the direction that is preferred by the users. Experimental results demonstrate that the proposed framework can effectively address the problems of generating various candidate styles and optimizing model parameters according to the obtained opinions. As far as we know, this is among the first studies on style selection with human involved in the loop. Though we focus on the target user group of older adults, our framework can be smoothly extended to other groups.

The rest of the paper is organized as follows: Section 2 presents related works on style control, hyperparameter selection and TTS systems for older adults. Section 3 introduces the proposed HILvoice framework. Experimental setup and results are given in Section 4. Conclusions are drawn in Section 5.

2. Related Works

As investigated in many previous works, style control approaches can be roughly divided into two categories, i.e., supervised and unsupervised methods. Supervised methods use corpora with predefined styles to train TTS models with style information as input to the models [1, 3]. Unsupervised approaches aim to discover styles from corpora without predefined annotations [2, 4, 7]. Recent developments improve the utterance-level control to finer-grain controls on smaller units,

* Corresponding author. This research is supported by National Natural Science Foundation of China (62076144), the CUHK Stanley Ho Big Data Decision Analytics Research Centre and the Centre for Perceptual and Interactive Intelligence

e.g., words, phonemes or frames [6, 8, 11]. Efforts have also been devoted to improve the interpretability by performing control on style-related acoustic features, e.g., speaking rate, pitch and intensity [5, 9, 10]. However, based on these controllable models, how to determine the controlling configuration, e.g., the suitable embedding, for a specific user group, is rarely studied. Perrotin and McLoughlin [15] developed a real-time system for modifying voice quality. This system can be added at the end of a speech synthesis pipeline to generate the style that is tuned by target users. However, the operation of the system requires background knowledge of signal processing, which hinders the system application to the user groups without the required knowledge. Udagawa, Saito and Saruwatari [16] proposed a human-in-the-loop method for speaker adaptation. However, their work focuses on speaker voice selected by an individual listener, while our framework focuses on speaking style preferred by a target user group.

Our framework is also related to the topic of hyperparameter selection. There are a range of hyperparameter selection approaches that have been widely used in various research areas, e.g., grid search [17], random search [18] and SMBO [19]. However, it is difficult to directly apply these algorithms to synthesis style selection due to the expensive evaluation based on subjective experiments. In this work, we adapt the SMBO algorithm for style selection by approximating the loss using subjective preference test results and updating the model parameters according to the approximated loss.

Regarding development of elder-facing TTS systems, there is less work in this direction due to lack of corpora specifically designed for elderly listeners. [14] recruited exemplar speakers, who work regularly with elderly adults and/or have certification as senior peer counselors, to collect a Japanese corpus for elder-facing TTS development. The relationship between inter-sentence distances and pause lengths was analyzed based on this corpus. [20] investigated the audio preferences of the elderly from five aspects: volume, pitch, speed, timbre and music genre. Their analysis shows that older adults prefer 72.9–79.2 dB sounds, 440.0 Hz-830.6 Hz pitch, and relatively slower speaker rate of about 190 words per minute. These are useful clues for elder-facing TTS corpus collection and system development. In our framework, no target corpus is required, instead, target users are involved in the style selection of controllable TTS systems. The users' preference opinions are taken into account in the model parameter selection, such that the TTS systems are able to generate speaking styles desired by them.

3. Proposed HILvoice Framework

To involve the target users in the style selection for TTS systems, we propose a novel human-in-the-loop framework, named HILvoice, based on the SMBO algorithm. As shown in Figure 1, the framework iteratively updates the TTS model parameters $\boldsymbol{\lambda}_t$ by taking into consideration the preference opinions of target users. In the following, we will introduce the SMBO algorithm and the forward and backward paths of the iteration loop in the HILvoice framework.

3.1. Sequential model-based optimization

Sequential model-based optimization (SMBO) [19] is a versatile stochastic optimization framework that can effectively optimize model hyperparameters. As shown in Algorithm 1, the SMBO first builds a model $\mathcal{M}$ according to the initial parameters $\boldsymbol{\lambda}_0$, and then iterates the following steps: (i) uses $\mathcal{M}$ to determine the next parameters $\boldsymbol{\lambda} \pm \boldsymbol{\epsilon}$ to be explored (line 3); (ii) evaluates the loss $\mathcal{L}(\mathcal{M}, \boldsymbol{\lambda}, \boldsymbol{\epsilon})$ of the model with the new parameters $\boldsymbol{\lambda} \pm \boldsymbol{\epsilon}$ (line 4); and (iii) uses the loss information $(\boldsymbol{\lambda}, c)$ to update the model $\mathcal{M}$ (line 5 & 6). After the resource budget is exhausted and the iteration ends, the $\boldsymbol{\lambda}$ corresponding to the lowest loss is chosen as the selected parameters.

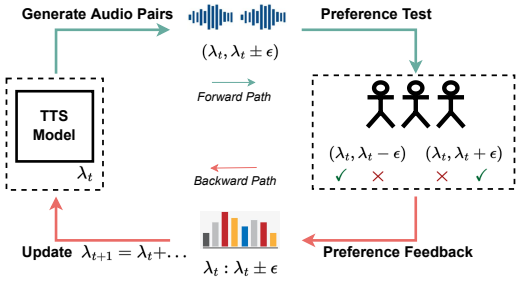

Figure 1: *The proposed HILvoice framework iteratively updates the TTS model parameters $\boldsymbol{\lambda}_t$ by taking into consideration the preference test results. The target users are involved in the loop of selecting the desired style.*

Our framework adapts the SMBO algorithm to find the optimal control parameters of acoustic feature transformation coefficients (e.g., 1.2 times slower) or neural embeddings (e.g., global style tokens). Target user subjects are invited to provide preference scores for the various control parameters. A proxy loss function is derived from the preference scores and used to backpropagate the loss to update the control parameters (Sec. 3.3).

Before the iterations, a controllable TTS model is well trained and used as the initialized model $\mathcal{M}$. The default control parameters (i.e., a neutral style) are used as the initial parameters $\boldsymbol{\lambda}_0$. Each iteration of the adapted SMBO algorithm consists of a forward path and a backward path. The forward path generates exploratory parameters $\boldsymbol{\lambda} \pm \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a small perceptible perturbation on the current control parameters $\boldsymbol{\lambda}$. Preference tests between $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda} \pm \boldsymbol{\epsilon}$ are conducted to obtain the preference score loss c. The backward path updates the control parameters $\boldsymbol{\lambda}$ according to loss c. The details of both paths in the loop are discussed in the following sections.

Algorithm 1 SMBO

1: initialise model $\mathcal{M}$; $\boldsymbol{\lambda} \leftarrow \boldsymbol{\lambda_0}$; $\mathcal{H} \leftarrow \emptyset$
2: **while** resource budget for optimization not exhausted **do**
3: $\quad \boldsymbol{\lambda} \pm \boldsymbol{\epsilon} \leftarrow$ candidate configuration from $\mathcal{M}$ // Sec. 3.2
4: $\quad$ Compute $c = \mathcal{L}(\mathcal{M}, \boldsymbol{\lambda}, \boldsymbol{\epsilon})$ // Sec. 3.3
5: $\quad \mathcal{H} \leftarrow \mathcal{H} \cup \{(\boldsymbol{\lambda}, c)\}$
6: $\quad$ Update $\mathcal{M}$ given $\mathcal{H}$ // Sec. 3.3
7: **end while**
8: **return** $\boldsymbol{\lambda}$ from $\mathcal{H}$ with minimal c

3.2. The forward path

The forward path uses the pretrained TTS model to provide various options for target users to pick. Given the TTS model $\mathcal{M}$ and the current control parameters $\boldsymbol{\lambda}_t$, a small perturbation $\boldsymbol{\epsilon}$ is added to the current control parameters $\boldsymbol{\lambda}_t$. The purpose of this operation is to generate different speech utterances around the current control parameters for users to pick, so as to stably obtain the direction and step size for further model updates. One

thing to note is that the perturbation should not be too small to be perceived by the target group. In each loop, audio pairs $(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}_t - \boldsymbol{\epsilon})$ and $(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}_t + \boldsymbol{\epsilon})$ are generated by the above well-trained TTS model, and presented to the target users for ABX preference tests. Data processing and analysis will then be conducted to obtain the ratios of picking $\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}_t \pm \boldsymbol{\epsilon}$ among all $(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}_t \pm \boldsymbol{\epsilon})$ pairs.

3.3. The backward path

The backward path obtains opinions from target users and update the TTS model for the next loop. Here, we design a simple strategy and mainly focus on the overall preference ratios, which indicate the direction and step size for the next model update. Specifically, let P_b and P_s denote the ratios of picking $\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}_t - \boldsymbol{\epsilon}$ among all $(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}_t - \boldsymbol{\epsilon})$ pairs respectively, Q_b and Q_s denote the ratios of picking $\boldsymbol{\lambda}_t + \boldsymbol{\epsilon}$ and $\boldsymbol{\lambda}_t$ among all $(\boldsymbol{\lambda}_t, \boldsymbol{\lambda}_t + \boldsymbol{\epsilon})$ pairs, and P_n and Q_n denote the ratios of no preference between the compared pairs (i.e., $P_b + P_s + P_n = 1$). Note that, bigger values of P_b and Q_b indicate the preference in the direction of $\boldsymbol{\epsilon}$, and bigger values of P_s, Q_s indicate the $-\boldsymbol{\epsilon}$ direction. Hence, the updating direction depends on the relative values of $P_b + Q_b$ and $P_s + Q_s$. If $P_b + Q_b > P_s + Q_s$ then move the current parameters $\boldsymbol{\lambda}_t$ towards $\boldsymbol{\epsilon}$ with a step size determined by the values of $P_b + Q_b$, and vise versa. The updating operation can be represented as:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + r_t \boldsymbol{\epsilon} \tag{1}$$

$$r_t = \begin{cases} \frac{P_b + Q_b}{2}, & P_b + Q_b > P_s + Q_s \\ 0, & P_b + Q_b = P_s + Q_s \\ \frac{-(P_s + Q_s)}{2}, & P_b + Q_b < P_s + Q_s \end{cases} \tag{2}$$

This update improves the control parameters towards the direction that is preferred by the users. With the updated parameters $\boldsymbol{\lambda}_{t+1}$, the loop is repeated until the difference between $\boldsymbol{\lambda}_{t+1}$ and $\boldsymbol{\lambda}_t$ is smaller than a preset threshold or the resource budget is exhausted. The $\boldsymbol{\lambda}$ with a minimal loss $\mathcal{L} = |P_b + Q_b - P_s - Q_s|/2$ will be chosen according to Algorithm 1.

4. Experiments

4.1. Corpus

An internal single-speaker Cantonese corpus is used for the experiments, containing around 12 hours of speech data spoken by a male Cantonese native speaker with neutral speaking style. The corpus is designed for all generations, not specifically for older adults. The corpus has a total of 10,000 audio utterances, of which 200 utterances are used for validation, 100 for test, and the rest for training.

4.2. Experimental Settings

We adopt the FastSpeech 2 [5] as our basic TTS model, which contains a variance adaptor to model the pitch, energy and duration features. For feature extraction, 80-dimensional Mel-spectrograms are extracted with 16kHz sampling rate. The frame size is set to 1,200 and the hop size is set to 240. The ground truth phoneme duration is extracted by the Montreal Forced Aligner [21].

We control the prosodic characteristics in the generated speech via the model parameter set $\boldsymbol{\lambda} = [p, e, d]$, where p, e, d are the there parameters related to pitch, energy and duration. Two different control methods are investigated in the experiments. The first is an *explicit* control directly on the outputs of variance predictors, i.e., the prosodic acoustic features, by scaling up or down the values of pitch, energy or duration for each phoneme proportionally. Here, the parameters p, e and d are the scaling coefficients. For example, $d = 0.2$ means scaling the duration values to 1.2 times larger, and the speaking rate decreases accordingly. The second method is an *implicit* control upon the hidden embeddings before the variance predictors. A small perturbation is added to the embeddings and the modified embeddings are fed to the predictors.

With initialized control parameters $\boldsymbol{\lambda}_0 = [0, 0, 0]$, we use different perturbation $\boldsymbol{\epsilon}$ for explicit and implicit controls respectively. In the explicit control, the pitch, energy and duration perturbation vectors are $\boldsymbol{\epsilon}_p = [0.1, 0, 0], \boldsymbol{\epsilon}_e = [0, 0.3, 0]$ and $\boldsymbol{\epsilon}_d = [0, 0, 0.15]$ respectively. For the implicit control, the vectors are $\boldsymbol{\epsilon}'_p = [5, 0, 0], \boldsymbol{\epsilon}'_e = [0, 5, 0]$ and $\boldsymbol{\epsilon}'_d = [0, 0, 0.1]$. These perturbation vectors are set empirically so that the perturbation granularity is sufficiently small while at the same time the generated speech with perturbed control parameters is different from the original speech in perception. In each iteration, we update the explicit and implicit control parameters according to Eq. (1) with the above different $\boldsymbol{\epsilon}_{\{p,e,d\}}$ for the updates of pitch, energy and duration parameters respectively.

4.3. Preference Tests

We conduct ABX preference tests to obtain feedback for updating control parameters and also for the final effectiveness evaluation of the proposed framework. The speech quality and naturalness are not evaluated because we only slightly perturb the control parameters and the speech quality and naturalness are not affected.

In the preference tests, the older adult subjects are invited to listen to pairs of audios with same text content but different control parameters (e.g., $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda} + \boldsymbol{\epsilon}_d$). For each pair, they are asked to provide a preference choice: (i) the former is better; (ii) the latter is better; or (iii) no preference (the difference between the paired utterances is difficult to be perceived). For different pairs, utterances with different text content are synthesized. All the tests are conducted on the same device (ASUS E18534 laptop) in the same room with low background noise. The subjects listen to the audios with the built-in speaker, and the volume is set to the maximum (100%) when the audios are played.

We invite 21 older adults in our experiments, including 10 males and 11 females. The average age of the 21 elderly people is 72 years old, among whom the youngest is 62 and the oldest is 83. We conduct three rounds of iterative updates, and each round of updates involves 4, 4, and 8 people in the ABX tests. For the comparison between $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_{\{1,2,3\}}$, 4, 12 and 5 people participate in the preference tests respectively. We specifically invite more people to evaluate the comparison between $\boldsymbol{\lambda}_0$ and $\boldsymbol{\lambda}_2$ to support the effectiveness evaluation, as $\boldsymbol{\lambda}_2$ is the output selection of our framework.

4.4. Experimental Results

Table 1 shows the iteratively updated control parameters $\boldsymbol{\lambda}_{\{0,1,2,3\}}$ for both explicit and implicit controls. The control parameters for duration increase more significantly compared to the parameters for pitch and energy, and consistently in both explicit and implicit controls, as shown in Figure 2. The increase of duration control parameters reflects that a slower speaker rate is preferred by older adults, this coincides with intuition and previous studies in auditory perception of older adults [14, 20].

Table 1: *Control parameters $\boldsymbol{\lambda}_t$ in the t-th iteration for explicit and implicit controls. "norm" means the raw control parameters are normalized by the corresponding perturbation unit $\boldsymbol{\epsilon}$.*

	explicit							implicit						
	pitch		energy		duration		loss	pitch		energy		duration		loss
	raw	norm	raw	norm	raw	norm		raw	norm	raw	norm	raw	norm	
$\boldsymbol{\lambda}_0$	0	0%	0	0%	0	0%	0.306	0	0%	0	0%	0	0%	0.209
$\boldsymbol{\lambda}_1$	0.04	40%	-0.17	-56.7%	0.11	73.3%	0.222	3.13	62.6%	0	0%	0.09	90%	0.153
$\boldsymbol{\lambda}_2$	-0.03	-30%	-0.01	-3.3%	0.20	133.3%	**0.118**	0	0%	-2.7	-54%	0.04	40%	**0.141**
$\boldsymbol{\lambda}_3$	0.03	30%	-0.17	-56.7%	0.20	133.3%	–	-2.88	-57.6%	-5.2	-104%	0.09	90%	–

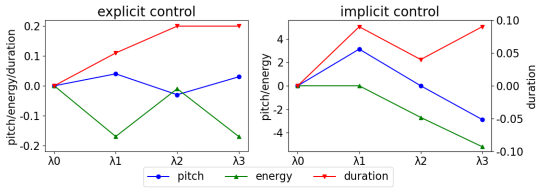

Figure 2: *Explicit and implicit control parameters in various iterations. For both control methods, duration parameters increase consistently.*

The energy values are consistently scaled down in both explicit and implicit control. One possible reason is that the original sound volume is already sufficiently high[1] so that the test participants prefer a lower volume. According to the SMBO algorithm, $\boldsymbol{\lambda}_2$ is selected as the final parameters for both explicit and implicit controls.

The comparison between the original parameters and the selected parameters using both control methods by ABX preference test is shown in Figure 3. For both implicit and explicit controls, the selected parameters are better than or comparable with the original parameters. This demonstrates the effectiveness of the proposed HILvoice framework. The improvement in the implicit control is marginal, due to lack of interpretability of the hidden embedding space where the control is performed. The prosodic features are entangled and the setting of proper perturbation is difficult. This also implies the challenges of optimizing a pretrained model using black-box feedback of preference scores on the model outputs.

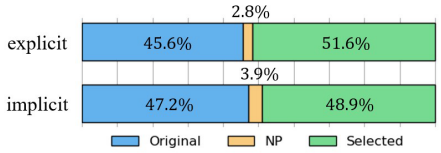

Figure 3: *Preference test results of explicit and implicit controls.*

To further explore the impact of the selected parameter sets on the synthesized speech, a case study is conducted and shown in Figure 4. The Mel-spectrograms and pitch contours of the utterances synthesized with the original control parameters and the selected control parameters using explicit and implicit methods are presented. It can be found that the duration values of both explicit and implicit controls are larger than that of the original setting. This shows that the generated utterances corresponding to the updated parameter sets have slower speaking rates and are expected to be easier for older adults to follow. This supports that our proposed HILvoice framework can select a better speaking style and improve the user experience of the target user group of older adults.

[1]Demo page: https://thuhcsi.github.io/iscslp2022-HILvoice

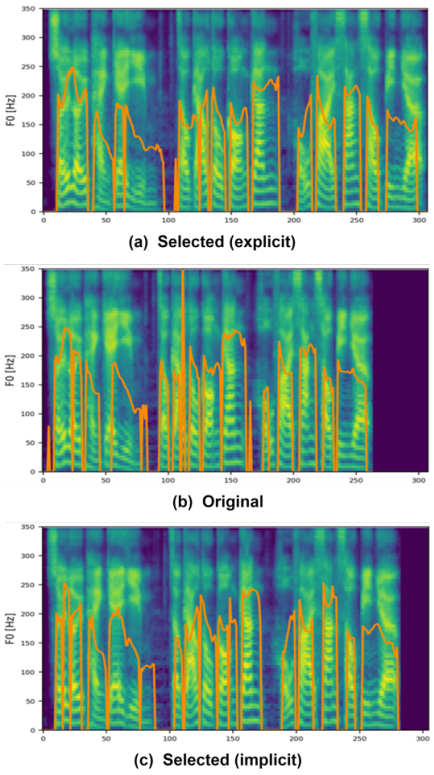

Figure 4: *The Mel-spectrograms and pitch contours of synthesized speech with different control parameters, for the text "The salt in the water causes the steel bars to rust and weaken".*

5. Conclusions

In this paper, we study style selection for elder-facing speech synthesis. A novel human-in-the-loop framework, called HILvoice, is proposed to involve the target user group of older adults in a loop to gradually update the control parameters in pretrained TTS systems. Experimental results demonstrate that the proposed framework can select a speaking style that is preferred by older adults than the original setting. Analysis on the results shows that the selected style has a slower speaking rate, which coincides with intuition and previous studies in auditory perception of older adults.

6. References

[1] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[2] F. Eyben, S. Buchholz, N. Braunschweiler, J. Latorre, V. Wan, M. J. Gales, and K. Knill, "Unsupervised clustering of emotion and voice styles for expressive TTS," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4009–4012.

[3] Y. Lee, S.-Y. Lee, and A. Rabiee, "Emotional end-to-end neural speech synthesizer," in *NeurIPS*, 2017.

[4] 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.

[5] 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," in *International Conference on Learning Representations*, 2020.

[6] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2020, pp. 6264–6268.

[7] X. Wu, Y. Cao, H. Lu, S. Liu, S. Kang, Z. Wu, X. Liu, and H. Meng, "Exemplar-based emotive speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 874–886, 2021.

[8] D. Tan and T. Lee, "Fine-grained style modeling, transfer and prediction in text-to-speech synthesis via phone-level content-style disentanglement," in *INTERSPEECH*, 2021.

[9] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," *arXiv preprint arXiv:2009.06775*, 2020.

[10] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised learning of disentangled and interpretable representations from sequential data," *Advances in neural information processing systems*, vol. 30, 2017.

[11] L.-W. Chen and A. Rudnicky, "Fine-grained style control in Transformer-based text-to-speech synthesis," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7907–7911.

[12] F. M. Zanzotto, "Human-in-the-loop artificial intelligence," *Journal of Artificial Intelligence Research*, vol. 64, pp. 243–252, 2019.

[13] W. Zhu, W. Zhang, Q. Shi, F. Chen, H. Li, X. Ma, and L. Shen, "Corpus building for data-driven tts systems," in *Proceedings of 2002 IEEE Workshop on Speech Synthesis, 2002.* IEEE, 2002, pp. 199–202.

[14] H. Nakajima and Y. Aono, "Collection and analyses of exemplary speech data to establish easy-to-understand speech synthesis for Japanese elderly adults," in *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2020, pp. 145–150.

[15] O. Perrotin and I. McLoughlin, "GFM-Voc: A real-time voice quality modification system," in *Interspeech 2019-20th Annual Conference of the International Speech Communication Association*, 2019, pp. 3685–3686.

[16] K. Udagawa, Y. Saito, and H. Saruwatari, "Human-in-the-loop speaker adaptation for dnn-based multi-speaker tts," *arXiv preprint arXiv:2206.10256*, 2022.

[17] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio, "An empirical evaluation of deep architectures on problems with many factors of variation," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 473–480.

[18] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012.

[19] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *International conference on learning and intelligent optimization*. Springer, 2011, pp. 507–523.

[20] D. Men and L. Wu, "The investigation into design elements of auditory pleasure experience for the elderly based on a testing tools development," in *International Conference on Human-Computer Interaction*. Springer, 2021, pp. 258–276.

[21] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.