ENHANCING MULTILINGUAL TTS WITH VOICE CONVERSION BASED DATA AUGMENTATION AND POSTERIOR EMBEDDING

Hyun-Wook Yoon[1], *Jin-Seob Kim*[1], *Ryuichi Yamamoto*[2†], *Ryo Terashima*[2†],
Chan-Ho Song[1], *Jae-Min Kim*[1], *Eunwoo Song*[1]

[1]NAVER Cloud, Korea, [2]LINE Corp., Japan

ABSTRACT

This paper proposes a multilingual, multi-speaker (MM) TTS system by using a voice conversion (VC)-based data augmentation method. Creating an MM-TTS model is challenging, owing to the difficulties of collecting polyglot data from multiple speakers. To address this problem, we adopt a cross-lingual, multi-speaker VC model trained with multiple speakers' monolingual databases. As this model effectively transfers acoustic attributes while retaining the content information, it is possible to generate each speaker's polyglot corpora. Subsequently, we design the MM-TTS model with variational autoencoder (VAE)-based posterior embeddings. It is to be noted that incorporating VC-augmented polyglot corpora into the TTS training process might degrade synthetic quality, since the corpora sometimes contain unwanted artifacts. To mitigate this issue, the VAE is trained to capture the acoustic dissimilarity between the recorded and VC-augmented datasets. Through the selective choice of the posterior embeddings obtained from the original recordings in the training set, the proposed model enables the generation of acoustically clearer voices.

***Index Terms*—** Multilingual text-to-speech, data augmentation, voice conversion, variational autoencoder (VAE)

1. INTRODUCTION

Recent advancements in text-to-speech (TTS) systems have been particularly noteworthy, especially with regard to monolingual TTS models [1–3]. Therefore, recent research is focused on multilingual TTS systems [4, 5]. A prevailing concern in the development of a multilingual TTS model is the substantial resources required to assemble a polyglot data set. The processes involved in data collection, curation, and validation are both time-consuming and financially burdensome.

To address these limitations, cross-lingual TTS [6–8] systems have been developed, where the TTS model is trained with a data set of speakers speaking different languages. Subsequently, techniques such as knowledge transfer [6] and speaker-content disentanglement [7] have been utilized with the model to extract linguistic information from input audio, distinct from speaker attributes. However, because it is challenging to faithfully separate speaker and linguistic information from the provided audio, the synthetic voice tends to have low speaker similarity [9].

As an alternative, we introduce an effective data augmentation method using a voice conversion (VC) approach. Unlike the conventional methods in [10–12], our proposed method employ a cross-lingual, multi-speaker (CM) VC model by using the monolingual databases collected from different linguistic background. Considering the limited data available per speaker, we anticipated that using a many-to-many voice conversion model to create a database encompassing both multilingual and multi-speaker (MM) aspects would enhance learning stability when building a TTS system [13]. Furthermore, the training process of the MM-TTS model became more straightforward because the VC-augmentation negated the necessity for the additional modules for speaker and language disentanglement.

However, simply combining the VC-augmented data with the TTS training set is not always beneficial for improving synthetic quality, since this might cause negative effects if poorly converted samples are included. To address this issue, our proposed method incorporates a posterior embedding from a variational autoencoder (VAE) [14] into the MM-TTS model. VAEs are well known for capturing latent representations of feature distribution and have been employed in various TTS tasks [5, 15, 16]. Similarly, the proposed VAE model is learned together with MM-TTS to differentiate between acoustic distribution in the original and augmented data in the latent space. During the inference process, by selectively taking the posterior embeddings obtained from the original recordings in the training data, it is possible to further enhance synthetic quality while mitigating the possible degradation caused by the VC-augmentation. This will be further investigated in objective and subjective evaluations, which confirm the superior performance of the proposed method over the conventional ones.

2. PROPOSED METHOD

Our method involves a two-step procedure: Initially, we augment multi-speaker monolingual databases to create multi-speaker polyglot databases by using the CM-VC. Subsequently, we model the MM-TTS system using both recorded and augmented databases.

2.1. Data augmentation by voice conversion

Among the state-of-the-art VC models, we adopt a non-parallel many-to-many Scyclone model as a base architecture thanks to its straightforward and stable generation [17]. To effectively train the CM-VC model, we also apply pitch-shift data augmentation method [18], where each of monolingual training corpus is reproduced by adjusting pitch values in several semitone-levels. In our

[†]Current affiliation is LY Corp. due to a merger with Yahoo Japan Corp.

979-8-3503-4485-1/24/$31.00 ©2024 IEEE

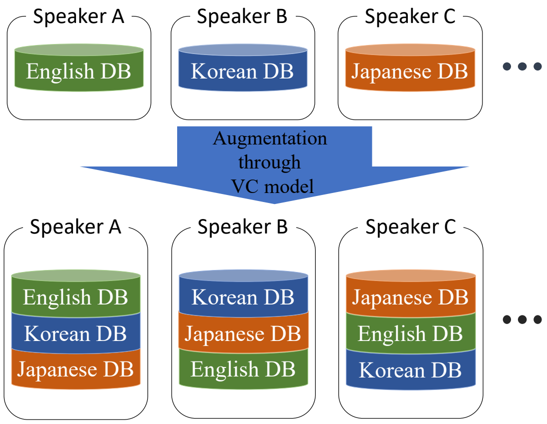

Fig. 1: Data augmentation process using CM-VC.

preliminary experiments, the quality of generated polyglot corpora becomes more stable and natural because this process enables the CM-VC to cover a variety of prosodies from multiple speaker and languages.

Fig. 1 depicts the generation process of multi-speaker polyglot dataset. The pretrained CM-VC model converts the source speaker's monolingual (e.g., English) corpus into the target speaker's voice of which original recordings only contain another language (e.g., Korean).

2.2. Multilingual, multi-speaker TTS

2.2.1. Unified phoneme representation

After data augmentation, we proceed to model the MM-TTS model. However, to train a model across distinct languages, it is essential to merge phoneme definitions, which differ by language, into a single unified one. Therefore, we implement rule-based unified phoneme representations to combine three distinct languages. Specifically, our target dataset utilizes English, Korean, and Japanese data, each with 42, 47, and 50 phoneme definitions, respectively. They are integrated into a single set containing 102 phonemes considering the International Phonetic Alphabet (IPA) [19], and some phonemes with similar pronunciations are merged. For instance, several unvoiced phonemes (including the glottal stops and nasal sounds) that exhibit minimal pronunciation differences across languages are unified. Additionally, certain vowels (e.g., a, e, i, o, u) that are similar in Korean and Japanese are also unified. Ultimately, this standardized set of phonemes is used in the TTS model. Details of merged unified phoneme representations can be found in Table 1.

2.2.2. Posterior encoder

As described in section 1, simply incorporating augmented data into the training set can lead to quality degradation[1]. To address this issue, we integrate a VAE posterior encoder within the TTS acoustic model following the base architecture proposed in [20]. In general, the posterior encoder takes the acoustic features during training and maps their posterior distributions into latent space. Similar to this, we guide the posterior encoder to focus solely on capturing the distributions of the recorded and augmented data by providing explicit speaker and language information to the system (Fig. 2). Consequently, as described in the t-SNE plot in Fig. 3, the data clusters in the latent space become clearly distinguishable based on whether it is derived from the recorded or augmented data.

Before synthesis, the posterior encoder selectively extracts posterior embeddings from all the recorded data in the training set while discarding those from the augmented ones. Subsequently, the TTS model takes their average [21] as the posterior embedding in the synthesis step. As this process enables the TTS outputs to have distributions as similar as those of original recordings, the synthetic quality is significantly improved. This will be further discussed in section 3.3.

Table 1: Unified phoneme definitions comparing with the corresponding IPA symbols.

Consonants (Pulmonic)		Original IPA symbol			Unified symbol
		Korean	Japanese	English	
Plosive	Bilabial	p^h ㅍ (파랑)	p パ (パン)	p p (pack)	p
		b ㅂ (바람)	b ば (ばね)	b b (back)	b
	Alveolar	t^h ㅌ (타다)	t た (たび)	t t (time)	t
		d ㄷ (다수)	d ど (どう)	d d (dog)	d
	Velar	k^h ㅋ (크기)	k く (くる)	k k (kiss)	k
		g ㄱ (가방)	g が (がく)	g g (gaggle)	g
Nasal	Bilabial	m ㅁ (마을)	m ま (まあ)	m m (much)	m
	Alveolar	n ㄴ (나무)	n な (なみ)	n n (note)	n
Fricative	Labiodental		ɸ ふ (ふく)	f f (fish)	f
	Alveolar	s ㅅ (사랑)	s さ (さよ)	s s (soup)	s
		tʒ ㅈ (자유)	z ざ (ざん)	z z (zip)	z
	Alveolo-palatal & Postalveolar		ɕ し (しき)	ʃ sh (ship)	sh
	Glottal	h ㅎ (하늘)	h は (はな)		h
Affricate	Postalveolar		ʧ ち (ちゃ)	ʧ ch (chair)	ch
Trill & Approximant	Labiodental		ɾ ら (らく)	ɹ r (run)	r

2.2.3. TTS synthesis

Fig. 2 demonstrates the synthesis framework of the proposed MM-TTS model that consists of context encoder, autoregressive decoder and an external duration model [23–25]. As described in section 2.2.1, the input text is mapped to the unified phoneme sequence and converted into linguistic vectors[2]. At the same time, the posterior, speaker and language embeddings are extracted from the

[1] In our experiments, this is manifested as a "muffled sound".

[2] The linguistic vectors consist of 432-dimensional linguistic contents that encapsulate phoneme identity, accent, break, and positional information, etc.

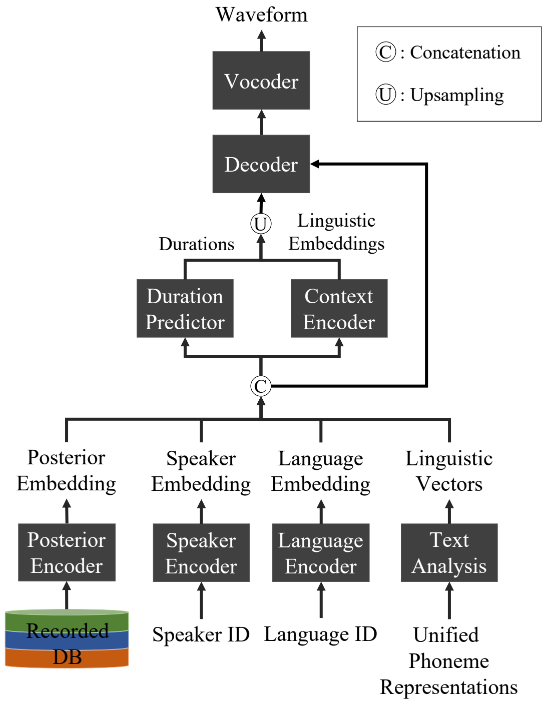

Fig. 2: MM-TTS Framework

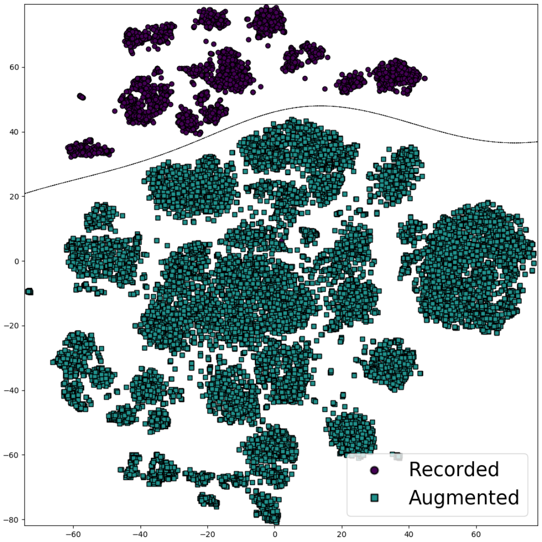

Fig. 3: t-SNE visualization of posterior embeddings obtained from the training data. Deep blue circles and green squares represent those from the recorded and augmented data, respectively. The black line represents marginal boundary calculated by a support vector machine [22].

corresponding encoders, concatenated with the linguistic vectors and fed into a context encoder to obtain high-level linguistic embeddings. They are also fed into a duration predictor to upsample the resulting linguistic embeddings to the frame-level. The autoregressive decoder transforms those features into acoustic vectors and finally the vocoder generates the corresponding speech.

3. EXPERIMENTS

3.1. Database

We utilized an internal dataset comprising six speakers: male and female native speaker for each language (English, Korean, and Japanese). For each speaker, 500, 100, and 50 sentences were used for training, validation, and testing, respectively. After training the CM-VC model, the augmentation was performed for each speaker using the data from other speakers as the source. As a result, each speaker had a training set comprising 500 recorded and 2,500 augmented audio samples. All of those corpora were used to train the proposed MM-TTS model. Note that the validation and test sets were the same as those used in the CM-VC task.

3.2. Model settings

3.2.1. CM-VC model

For the CM-VC model, the detailed setups followed our previous work in [18] except for two changes: One was modifying the structure from one-to-one to many-to-many by adding speaker's information (e.g., one-hot representation) to the generation process. The other was incorporating a single-Gaussian WaveRNN-based vocoder [26] to generate time-domain waveforms.

3.2.2. MM-TTS model

For the MM-TTS model, we used duration-informed Tacotron 2 model as described in section 2.2.3. The major architecture followed our previous work in [25] but we additionally introduced three auxiliary encoders such as posterior, speaker and language encoders. The posterior encoder consisted of six 2D convolution layers, a GRU layer with 128 units, and two projection layers for obtaining mean and log variance of latent variables. After the reparameterization trick [14], the output is fed to additional projection layer to form 32-dimensional posterior embedding. Learnable lookup-tables were used for the speaker and language encoders, converting input IDs into a 64-dimensional speaker embedding and an 8-dimensional language embedding, respectively.

3.3. Evaluations

To validate the performance of our method, we implemented three different systems: **CM-TTS**, **MM-TTS** and **MM-TTS$_{vae}$**. The CM-TTS was a baseline cross-lingual and multi-speaker TTS model that trained only with monolingual databases from multiple speakers without any augmentation method. Therefore, language transfer was conducted in the unspoken language. The MM-TTS$_{vae}$ was the proposed multilingual and multi-speaker TTS system trained with the CM-VC-augmented polyglot corpora from multiple speakers. Note that the MM-TTS was the same with MM-TTS$_{vae}$ but did not use the VAE-based posterior embedding.

Table 2: Objective evaluation results with respect to different TTS systems. Note that the results within a same language were averaged together (two speakers per a language).

Model	English				Korean				Japanese			
	WER(%)	CER(%)	F0$_{rmse}$(Hz)	LSD(dB)	WER(%)	CER(%)	F0$_{rmse}$(Hz)	LSD(dB)	WER(%)	CER(%)	F0$_{rmse}$(Hz)	LSD(dB)
CM-TTS	**3.11**	**1.29**	37.58	4.58	19.88	6.88	29.64	4.64	16.04	10.50	25.75	4.53
MM-TTS	16.74	10.28	37.73	4.22	27.76	11.74	26.41	**4.42**	21.24	14.01	24.72	**4.27**
MM-TTS$_{vae}$	4.87	2.34	**36.57**	**4.15**	**15.13**	**4.36**	**26.28**	4.59	**14.45**	**9.51**	**24.24**	4.36

Table 3: Naturalness MOS test results with 95% confidence interval with respect to different TTS systems. Note that the results within a same language were averaged together (two speakers per a language).

Model	First language : English			First language : Korean			First language : Japanese		
	English	Korean	Japanese	English	Korean	Japanese	English	Korean	Japanese
CM-TTS	2.71 ± 0.12	1.96 ± 0.11	2.16 ± 0.11	1.70 ± 0.08	2.75 ± 0.10	1.75 ± 0.09	1.77 ± 0.10	1.84 ± 0.10	2.93 ± 0.12
MM-TTS	2.93 ± 0.12	1.47 ± 0.08	1.91 ± 0.11	1.52 ± 0.08	2.15 ± 0.10	1.89 ± 0.09	1.96 ± 0.12	2.31 ± 0.13	2.98 ± 0.12
MM-TTS$_{vae}$	$\mathbf{3.13 \pm 0.12}$	$\mathbf{2.15 \pm 0.12}$	$\mathbf{2.20 \pm 0.12}$	$\mathbf{2.13 \pm 0.09}$	$\mathbf{3.03 \pm 0.10}$	$\mathbf{2.34 \pm 0.11}$	$\mathbf{2.30 \pm 0.12}$	$\mathbf{2.66 \pm 0.12}$	$\mathbf{3.15 \pm 0.12}$
Recorded	4.65 ± 0.08	-	-	-	4.94 ± 0.03	-	-	-	4.73 ± 0.06

3.3.1. *Objective evaluation*

To evaluate the prediction accuracy of the acoustic model, we measured root mean square error of the fundamental frequency ($F0_{rmse}$; Hz) and log spectral distance (LSD; dB). Note that no corresponding ground truth audio was available for augmented audio with transferred or converted language, therefore, we only conducted evaluations on audio samples that had ground truth recordings. The evaluation results shown in Table 2 verify that our proposed methods (i.e., MM-TTS and MM-TTS$_{vae}$) that incorporated data augmentation mostly outperformed the baseline (i.e., CM-TTS) in both $F0_{rmse}$ and LSD. This suggests that the data augmentation technique enhanced the generation accuracy of acoustic features.

We also measured word error rate (WER; %) and character error rate (CER; %) to evaluate the intelligibility of the synthetic waveforms. We used the *whisper-large-v3*[3] model to predict scripts from generated audio samples and calculated the WER and CER against the ground truth script. The results of this evaluation are presented in Table 2. In English, the performance of proposed model (i.e., MM-TTS$_{vae}$) did not exceed that of the baseline, though the difference was marginal. Conversely, our proposed method surpassed the baseline performance in both Korean and Japanese. We also found that the model, which simply incorporated augmented data (i.e., MM-TTS), showed higher error rate compared to the proposed model.

3.3.2. *Subjective evaluation*

We conducted naturalness MOS listening tests to evaluate the perspective quality of the synthetic speech.[4] We synthesized audio by randomly sampling 15 sentences for each of the 6 speakers from the test set across 3 systems, resulting in a total of 270 audio samples. Given that there were two speakers per language, a total of 300 audio samples per language were evaluated. In this evaluation, 10 participants per language were assigned and asked to rate each audio sample on a 5-point scale: 1 = Bad, 2 = Poor, 3 = Fair, 4 = Good, and 5 = Excellent. We asked participants to specifically rate the overall naturalness of each audio sample.

The results are shown in Table 3 of which trends can be analyzed as follows: Firstly, when comparing models with and without data augmentation (i.e., CM-TTS and MM-TTS), we observed that simply applying data augmentation led to some level of quality degradation. This demonstrates that indiscriminately mixing augmented data can lead to a degradataion in synthesized quality. Secondly, by utilizing VAE-based posterior embeddings to incorporate latent information from recorded data, the MOS score significantly improved across all languages (i.e., MM-TTS and MM-TTS$_{vae}$). Lastly, this trend remained consistent when compared without data augmentation system (i.e., CM-TTS and MM-TTS$_{vae}$), demonstrating that our proposed method can produce better-quality synthesized audio.

4. CONCLUSION

In this paper, we proposed an effective MM-TTS system that leveraged polyglot data augmentation using a pretrained CM-VC model. Our method addressed the inherent challenges associated with collecting polyglot data from multiple speakers by utilizing a CM-VC model that was trained on a monolingual database from various speakers. To address potential quality degradation stemming from naively incorporating VC-augmented polyglot corpora, we employed the VAE-based posterior embedding that was trained to discern the acoustic differences between recorded and VC-augmented datasets. Consequently, our proposed model produced better acoustic outputs by emphasizing posterior embeddings derived from the original recordings during the inference phase.

5. ACKNOWLEDGMENTS

This work was supported by Voice team, NAVER Cloud, Seongnam, Korea.

[3] https://huggingface.co/openai/whisper-large-v3

[4] Generated audio samples are available at the following URL:https://christophyoon.github.io/MMV-TTS

6. REFERENCES

[1] 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 *Proc. ICASSP*, 2018, pp. 4779–4783.

[2] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A diffusion probabilistic model for text-to-speech," in *Proc. ICML*, 2021, pp. 8599–8608.

[3] X. Tan, J. Chen, H. Liu, J. Cong, C. Zhang, Y. Liu, X. Wang, Y. Leng, Y. Yi, L. He *et al.*, "NaturalSpeech: End-to-end text to speech synthesis with human-level quality," *arXiv preprint arXiv:2205.04421*, 2022.

[4] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-TTS: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, pp. 8067–8077, 2020.

[5] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. ICML*, 2021, pp. 5530–5540.

[6] T. Tu, Y.-J. Chen, C.-c. Yeh, and H.-Y. Lee, "End-to-end text-to-speech for low-resource languages by cross-lingual transfer learning," in *Proc. Interspeech*, 2019.

[7] D. Xin, Y. Saito, S. Takamichi, T. Koriyama, and H. Saruwatari, "Cross-lingual text-to-speech synthesis via domain adaptation and perceptual similarity regression in speaker space," in *Proc. Interspeech*, 2020, pp. 2947–2951.

[8] J. Yang and L. He, "Cross-lingual text-to-speech using multi-task learning and speaker classifier joint training," *arXiv preprint arXiv:2201.08124*, 2022.

[9] ——, "Towards universal text-to-speech," in *Proc. Interspeech*, 2020, pp. 3171–3175.

[10] P. Vijayalakshmi, B. Ramani, M. A. Jeeva, and T. Nagarajan, "A multilingual to polyglot speech synthesizer for indian languages using a voice-converted polyglot speech corpus," *Circuits, Systems, and Signal Processing*, pp. 2142–2163, 2018.

[11] Q. Sun and K. Nagamatsu, "Building multilingual TTS using cross lingual voice conversion," *arXiv preprint arXiv:2012.14039*, 2020.

[12] J. Wu, A. Polyak, Y. Taigman, J. Fong, P. Agrawal, and Q. He, "Multilingual text-to-speech training using cross language voice conversion and self-supervised learning of speech representations," in *Proc. ICASSP*, 2022, pp. 8017–8021.

[13] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep Voice 2: Multi-speaker neural text-to-speech," *Proc. NeurIPS*, 2017.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. ICLR*, 2014.

[15] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Proc. Interspeech*, 2018, pp. 3067–3071.

[16] S.-H. Lee, S.-B. Kim, J.-H. Lee, E. Song, M.-J. Hwang, and S.-W. Lee, "HierSpeech: Bridging the gap between text and speech by hierarchical variational inference using self-supervised representations for speech synthesis," in *Proc. NeurIPS*, 2022, pp. 16 624–16 636.

[17] S. Lee, B. Ko, K. Lee, I.-C. Yoo, and D. Yook, "Many-to-many voice conversion using conditional cycle-consistent adversarial networks," in *Proc. ICASSP*, 2020, pp. 6279–6283.

[18] R. Terashima, R. Yamamoto, E. Song, Y. Shirahata, H.-W. Yoon, J.-M. Kim, and K. Tachibana, "Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation," in *Proc. Interspeech*, 2022, pp. 3018–3022.

[19] I. P. Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press, 1999.

[20] E. Song, R. Yamamoto, O. Kwon, C.-H. Song, M.-J. Hwang, S. Oh, H.-W. Yoon, J.-S. Kim, and J.-M. Kim, "TTS-by-TTS 2: Data-selective augmentation for neural speech synthesis using ranking support vector machine with variational autoencoder," in *Proc. Interspeech*, 2022, pp. 1941–1945.

[21] 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 *Proc. ICASSP*, 2019, pp. 6945–6949.

[22] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, pp. 1565–1567, 2006.

[23] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems," in *Proc. ASRU*, 2019, pp. 214–221.

[24] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive Tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling," *arXiv preprint arXiv:2010.04301*, 2020.

[25] H.-W. Yoon, O. Kwon, H. Lee, R. Yamamoto, E. Song, J.-M. Kim, and M.-J. Hwang, "Language model-based emotion prediction methods for emotional speech synthesis systems," in *Proc. Interspeech*, 2022.

[26] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-time neural text-to-speech with sequence-to-sequence acoustic model and waveglow or single gaussian wavernn vocoders." in *Proc. Interspeech*, 2019, pp. 1308–1312.