Accented Text-to-Speech Synthesis with a Conditional Variational Autoencoder

Jan Melechovsky[1], *Ambuj Mehrish*[1], *Berrak Sisman*[2], *Dorien Herremans*[1]

[1]Singapore University of Technology and Design, Singapore
[2]Speech & Machine Learning Lab, The University of Texas at Dallas, Texas, USA

jan_melechovsky@mymail.sutd.edu.sg, ambuj_mehrish@sutd.edu.sg

Abstract

Accent plays a significant role in speech communication, influencing one's capability to understand as well as conveying a person's identity. This paper introduces a novel and efficient framework for accented Text-to-Speech (TTS) synthesis based on a Conditional Variational Autoencoder. It has the ability to synthesize a selected speaker's voice, which is converted to any desired target accent. Our thorough experiments validate the effectiveness of the proposed framework using both objective and subjective evaluations. The results also show remarkable performance in terms of the ability to manipulate accents in the synthesized speech and provide a promising avenue for future accented TTS research.

Index Terms: Text-to-Speech, Accent, Conditional Variational Autoencoder, Controllable Speech Synthesis, Accent Conversion

1. Introduction

Accent in speech is a way of speaking a certain language that can be described on a phoneme, rhythmic, intonation and structural level [1]. As part of one's idiolect, accent carries information about a person's background, such as education, region, and mother tongue [1, 2]. As such, disentangling the accent from other speaker characteristics that form one's idiolect, such as pitch and vocal tract shape, remains a challenge.

Even though accented TTS has numerous real-world applications, it has not been the main focus of the Text-to-Speech (TTS) field. Incorporating accent into TTS models would allow for more customizability, e.g., a better identity representation of people with speaking disabilities. Moreover, changing the accent of any conversational AI system has the potential to allow specific users to understand its produced speech better. We hypothesize that learning an accent representation as an auxiliary attribute of a TTS system requires fewer training data for each particular accent, given that we have data from other accents to supplement, compared to training an accent-specific system.

In recent years, the field of TTS has seen advancements in terms of performance with the arrival of attention-based deep learning models [3, 4]. Some examples include the well-known Tacotron and Tacotron2 models [5, 6], FastSpeech [7], or FastSpeech2 [8]. When it comes to controllable speech synthesis in TTS, selected major contributions include GST [9], GMVAE-Tacotron [10], and VAE for speaking style modeling [11]. In GST, a set of tokens is learnt in an unsupervised manner from the input reference audio files and these tokens can learn different attributes in speech, such as pitch, pace, or noisiness of the signal. In [10], the ability of GMVAE to capture different attributes was shown as well. Other work that uses a VAE module to capture different speaking styles is that of [11, 12], where the focus was on latent prosody attributes such as affect and intent. In [13], a Multi-Level VAE is used in an attempt to disentangle speakers and accents. [14]'s interesting work on accented TTS focuses on changing the accent intensity of L2 (second language) speakers to make them sound as L1 (native) speakers. This was achieved through an accent variance adaptor that modifies phoneme energy, duration, and pitch. Motivated by the success of VAE, we propose a novel TTS framework based on conditional VAE in this paper, that allows flexible manipulation of accent.

Most of the research on accented speech focuses on foreign accent conversion (FAC) [15, 16, 17], accent identification, and accented automatic speech recognition (ASR) [18, 19, 20, 21]. FAC is a special case of voice conversion (VC) [22, 23, 24], in which an input audio sample is converted into an output audio sample that is modified, e.g., different emotion, speaker identity, or accent. In FAC, the goal is to convert the input L2 speech into an L1-like speech. An interesting work is that of [15], in which the accent and speaker representations are provided via two pre-trained classifiers to achieve disentanglement. In [17], the disentanglement is obtained through the use of adversarial learning with an auxiliary speaker classifier. Unlike in FAC, our aim is to develop a system capable of generating speech in any accent, for any speaker (no L2- or L1-specific conversion). Moreover, our proposed method is a TTS system, which can generate speech from any text without the need for source audio.

In this work, we propose an efficient and reliable TTS system based on Tacotron2 that utilizes a Conditional Variational Autoencoder (CVAE) [25] to allow accent conditioning while retaining speaker identity. The CVAE architecture allows for the generation of speech with controllable features such as speaker identity, emotion, or style. The main objective of the proposed architecture is to synthesize speech in the style of the source speaker converted to a target accent. The contributions of this work are as follows: 1) We present a novel framework for controllable speech synthesis with a focus on accent conversion; 2) The proposed framework allows disentanglement of accent and speaker characteristics; and 3) Converting a speaker's accent is simple and does not require any reference audio.

The rest of the paper is organized as follows: In Section 2 we present the related work, followed by a description of our proposed method. We describe the training procedure and experiments in Section 4, and Section 5 concludes the study.

2. Related Work

For our baseline encoders in the experiment, we chose to use GST [9] and GMVAE [10], both combined with the same Tacotron2 system as the proposed method. The GST module uses 10 tokens with 8 attention heads to produce an embedding

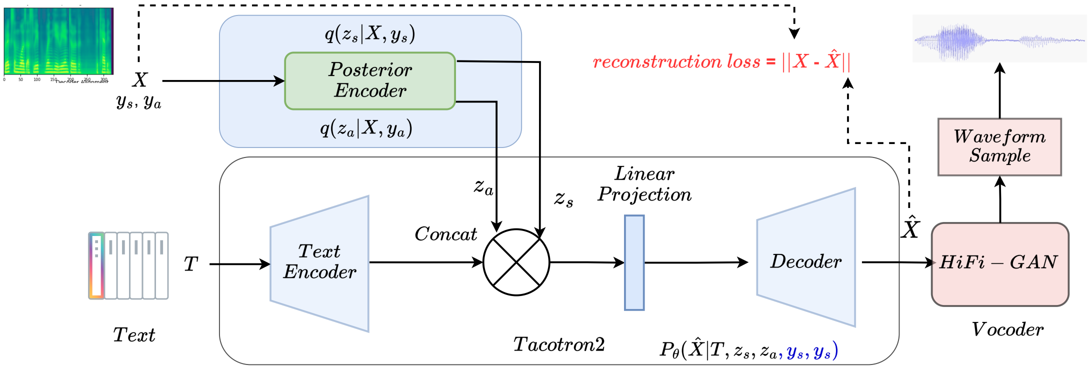


Figure 1: *An illustration of the training phase and overall architecture of the proposed framework, Tacotron2 with CVAE encoder.*

of size 256. As in the original paper, the training is unsupervised. For the GMVAE baseline, we follow the architecture from [10], but use two observed encoders – one for accent, one for speakers – and no latent encoder. The embedding size is 16 for each encoder and the number of mixtures is 6 for accents and 24 for speakers. As such, this framework uses both the speaker and accent labels as input. Our proposed method, which is described in the next section, uses a Conditional VAE (CVAE) to capture accent and speakers.

3. Proposed Method

The architecture of the proposed method is shown in Fig. 1. It consists of Tacotron2 [6] and Posterior Encoder (Fig. 2). As the Posterior Encoder, we have opted to use a CVAE architecture [25] with the objective of maximizing the evidence lower bound (ELBO) of the intractable marginal log-likelihood of data $\log p_\theta(X|\mathbf{y})$:

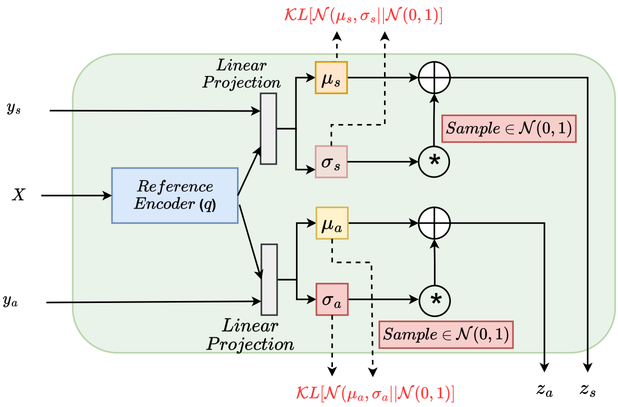


Figure 2: *Posterior Encoder architecture based on CVAE.*

$$\log p_\theta(X|\mathbf{y}) \geq \mathbb{E}_{q_\phi(z|X)}[\log p_\theta(X|\mathbf{z}) - \log \frac{q_\phi(z|X)}{p_\theta(X|\mathbf{y})}] \quad (1)$$

where θ and ϕ represent the parameters used for the decoder and posterior encoder respectively and $p_\theta(X|\mathbf{y})$ denotes a posterior distribution of latent variable $z = [z_s, z_a]$ with given label condition $y = [y_s, y_a]$ for a speaker s and accent a. The negative ELBO is then used as training loss, which can be viewed as the sum of the reconstruction loss $-\log p_\theta(X|\mathbf{z})$ and the KL divergence loss $\log q_\phi(z|X) - p_\theta(X|\mathbf{y})$.

The L_2 loss between the predicted mel spectrogram $\hat{X}$ and ground truth mel spectrogram X is used as the reconstruction loss:

$$L_{recon} = ||\hat{X} - X||_2 \quad (2)$$

where $||.||_2$ denotes L_2 norm. For the CVAE encoder, we propose two variants. The first one follows the traditional CVAE concept of having a label passed as a condition to both the encoder and decoder. The intuition is that the speaker and accent are mainly determined by the provided labels and the latent distribution captures minor differences inside these categories, like prosody. The second variant uses labels only in the encoder. Thus, the whole accent and speaker representation is captured by the latent variables z_a for accent and z_s for speaker. We name these two variants CVAE-L, and CVAE-NL, indicating 'label', and 'no-label', respectively.

The generated z_a and z_s, each of size 128, are concatenated with the text embeddings and passed through a single linear layer. The output is then passed to the decoder to generate a mel spectrogram, which is converted into audio with a pre-trained HiFi-GAN.

4. Experiments

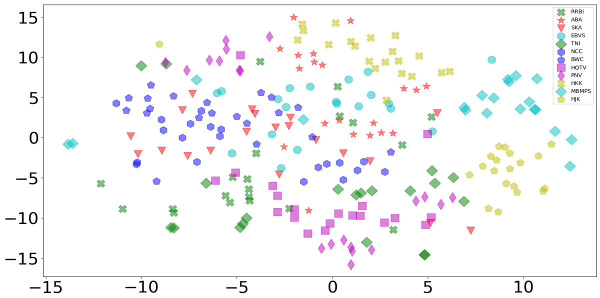
4.1. Dataset

In our experiments, we use the L2Arctic dataset [26], which contains 27 hours of recorded speech from 24 speakers with 6 distinct accents – Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese. Each accent is represented by two female and two male speakers and the data is mostly paired except for a few missing utterances in some speakers. We picked 10 unseen (never seen by the model from any speaker) utterances from all the speakers as our test dataset, and we use 15 seen (seen by the model from a different speaker) and 5 unseen utterances for each speaker as our validation set. The rest of the data is used for training.

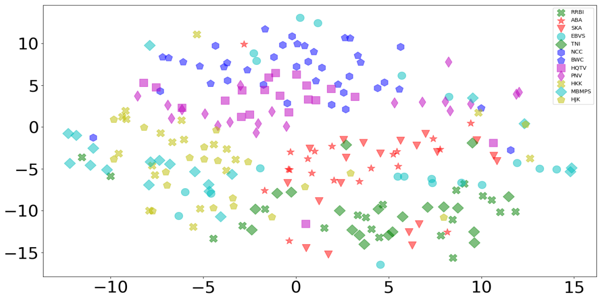
4.2. Training and Inference

We train all models with a batch size of 64 and use ADAM optimizer for 150k steps. The KL loss coefficient is set to 1×10^{-4} at the beginning of training. After reaching 10k steps, it linearly increases to 5×10^{-4} until 35k steps and is kept constant afterwards. All the models are trained with the target mel spectrogram both as the reference input and target output, as well as the relevant text input.

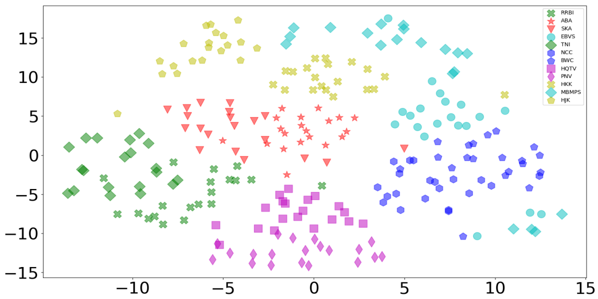
During inference, the model can use reference audio files for speaker and accent modelling. Additionally, with the posterior encoder being a VAE module, we can sample from the distribution to generate the embeddings. For our experiments, we first use the validation set to extract the μ_a and μ_s values (Fig. 2)



(a) *Speaker embeddings z_s.*



(b) *Accent embeddings z_a.*



(c) *Combined embeddings of z_s and z_a.*

Figure 3: *A t-SNE projection of the CVAE-NL embeddings. Each colour represents a different accent, whereas each shape represents a different speaker.*

for each sample and store them. Then, to represent a specific speaker and a specific accent during inference, we average the stored μ_s and μ_a values across the respective speaker and accent, leaving us without the need for further reference audio. In our baseline models, we adopted a similar approach of taking the average representation for each accent and speaker during inference. In addition, to improve speech quality and minimize signal distortion, we apply post-processing to speech samples synthesized by all models[1]. The post-processing improves the overall sound quality of the signal by removing unwanted background noise and enhances the listening experience.

4.3. Accent and Speaker Modelling Analysis

We visualized the embedding space of the CVAE-NL variant by encoding reference audio of the validation set and performing a t-distributed stochastic neighbor embedding (t-SNE) for 12 of the 24 speakers. In Fig. 3a, we can see that the speaker embeddings clearly form clusters per speaker. In Fig. 3b, we observe overlap between speakers of the same specific accents. Interestingly, the combined (concatenated) embeddings in Fig. 3c show even more compact clusters. This shows that the identity of each speaker is determined by both of the embeddings to a certain degree. One can imagine that if we move inside the combined embedding space by changing accent embeddings only, we might get a representation of a different speaker too. Naturally, we have observed this phenomenon in some of the synthesized audio samples with accent conversion, which influenced the design of the subjective evaluation tests, described in Section 4.5.

[1] https://pypi.org/project/noisereduce/

Table 1: *Objective Evaluation results (the lower, the better).*

Metric	GT	CVAE-NL	CVAE-L	GMVAE	GST
MCD ↓	-	7.10	7.176	7.375	7.502
WER ↓	0.1549	0.2311	0.2008	0.2228	0.1959

4.4. Objective Evaluation

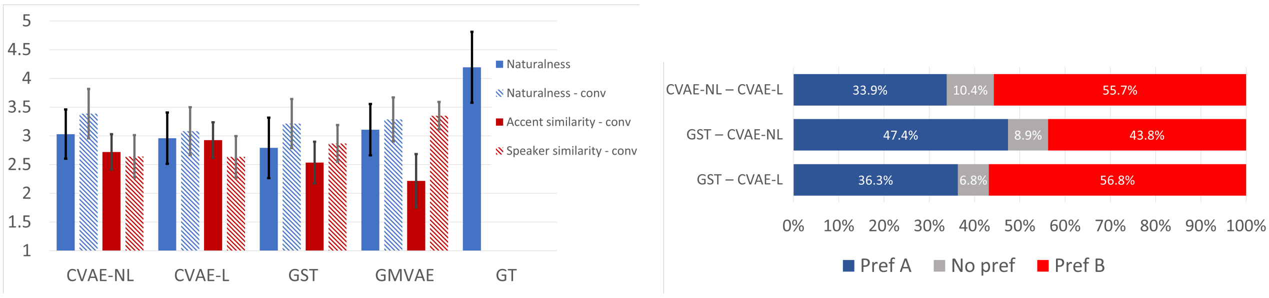
Mel Cepstral Distortion (MCD) [27] is used to evaluate the mel spectrogram reconstruction capabilities of each model, while Word Error Rate (WER) is used to evaluate the intelligibility of synthesized speech. The silero speech-to-text pre-trained models [28], designed for enterprise-grade use, are used for this purpose. The objective evaluation results presented in Table 1 show that all models perform similarly in terms of MCD and WER. However, our proposed methods, CVAE-NL and CVAE-L, slightly outperform GST in terms of MCD but lag behind in terms of WER. These results demonstrate that our proposed method achieves state-of-the-art speech quality.

4.5. Subjective Evaluation

We evaluate the performance of the proposed method[2] and baselines via listening tests. A total of 16 participants attended our experiments and each listened to 192 samples. The naturalness of audio is assessed through Mean Opinion Score (MOS) [29]. Listeners listened to samples from the models both in non-conversion setting (accent is of the original speaker), and conversion setting (we altered the accent of the original speaker to a new one). The results for naturalness (and other MOS tests that we explain later) are reported in Fig. 4a. To assess the statistical significance in the naturalness results, we deploy a paired t-test to test between: 1) each of the non-conversion models, 2) each of the conversion models, 3) each of the models' non-conversion result to their respective conversion counterpart.

In the non-conversion setting (1), GST naturalness shows a statistically significant difference ($p < 0.05$) from GMVAE and CVAE-NL, as GST scored the lowest in terms of naturalness. In the conversion setting (2), there is a statistically significant difference between the results of CVAE-NL and CVAE-L ($p < 0.05$), which could suggest that CVAE-NL, which scored the highest in naturalness after conversion, could be preferred over CVAE-L. In the non-conversion to conversion setting (3), the naturalness of CVAE-NL and GST models shows a statistically significant difference when compared to their conversion counterparts ($p < 0.001$). This is a very interesting observation as the accent conversion seems to have increased the perceived naturalness of audio. In the case of CVAE-L and GMVAE, the difference is not statistically significant. These results show that the audio of the proposed framework does not suffer from any loss in naturalness when performing the accent conversion, which is an important attribute of the system. Finally, all the

[2] Audio samples and code available via https://dapwner.github.io/CVAE-Tacotron/



(a) *MOS results with 95% CI for naturalness without and with accent conversion; and accent and speaker similarity after accent conversion.*

(b) *XAB general preference tests of accent-converted audio.*

Figure 4: *Subjective evaluation results*

models' naturalness (with or without conversion) show statistically very high significance when compared to the ground truth ($p < 0.001$), which is to be expected.

We further assessed the performance of the proposed method in terms of accent and speaker similarity. We note that MOS has been used in both experiments, as we want to quantify the perceived trade-off between accent and speaker identity after conversion. In the accent similarity test, listeners are given two reference samples – one of the source speaker S1 to get an idea about the original accent A1, and one of the target accents A2 represented by a different speaker S2. Then, they are presented with an audio sample of the source speaker S1 in a target accent A2, and are to rate the accent similarity to accent A2. A paired t-test was employed to evaluate the statistical significance between the models. In the accent similarity results (Fig. 4a), there is a statistically significant difference in CVAE-L compared to GST and GMVAE, both with ($p < 0.001$). With $p < 0.05$, all the models show a statistically significant difference in accent similarity from each other. This shows that the proposed methods perform better for accent conversion than the baselines, with CVAE-L being the superior option. On the other hand, GMVAE performs very poorly in this experiment.

In the speaker similarity test, similarly to the previous test, listeners are presented with an S1 source speaker's speech in accent A1. Then they are presented with converted samples of speaker S1 in target accent A2. Participants are to judge "how well the original speaker identity is retained after the accent conversion" on a 5-point scale. In this test, the paired t-test shows a statistically significant difference between GMVAE and all the other models ($p < 0.001$), and GST shows a statistically significant difference to other models with ($p < 0.05$). While the performances of the proposed CVAE-L and CVAE-NL are almost on par with GST, they lag behind GMVAE. This reflects the trade-off between retaining speaker identity and converting accent as the GMVAE baseline converts accent only very little if at all.

Furthermore, we conducted an XAB [30] preference test to assess general preference of accent-converted audio without specific focus on either accent or speaker identity. In this test, listeners are presented with a reference sample of the source speaker in their original accent and a reference sample of the target accent. Then, we ask the following question: "Imagine you are the original speaker J and you want your speech to be converted to a new accent K. Which one of the samples of A or B would you prefer for your new accent audio?" The purpose of this test is both to evaluate performance, but also to investigate the perception by the listeners – whether a more accented audio is preferred even though there might be an identity change, or whether keeping the identity is more important. For this test, we excluded the GMVAE model since its conversion capability was reported low as shown in the previous results. The results in Fig. 4b show that the GST and CVAE-NL models are on par in terms of preference. However, the CVAE-L model is preferred over both the GST and the CVAE-NL models, making it the superior choice. This shows that the proposed CVAE framework is an improvement to the state-of-the-art in accented TTS.

4.6. Discussion on accent-identity balance

The proposed CVAE model achieves promising results in terms of objective and subjective evaluation. As shown in Section 4.3, the embedding space allows for strong accent conversion but this can come at the price of identity disruption. This accent-identity balance is a complex issue, as accent partially forms a person's identity. On this note, the baseline of GMVAE seems to have captured speaker identity too well, hence not allowing for a significant change in accent. We note that this issue might be amplified by the dataset only having 4 speakers per accent. In an extreme case, where one would have just 1 speaker per accent, it would not be possible to distinguish between accent and speaker identity at all. Finally, it is worth mentioning that the accent-identity balance preference can depend on the application. If the primary concern is to convert an audio for better intelligibility, then a change in identity can be a bearable downside; while in some applications, such as TTS for people with muteness, retaining speaker identity might be the more important aspect of the two. Ideally, speaker identity and accent should be completely disentangled. In future work, we aim to focus on designing stronger disentanglement mechanisms to better separate accent from speaker identity.

5. Conclusion

This paper introduces a novel framework for accented TTS, which fuses a Conditional Variational Autoencoder with the Tacotron2 system. The proposed framework allows for efficient synthesis of any chosen speaker's speech converted to any of the target accents. We conducted extensive objective and subjective tests to evaluate the efficacy of the proposed approach. The results show a strong performance in terms of the model's ability to synthesize natural-sounding speech in a converted accent, which pushes the state-of-the-art boundaries. We discuss the accent-identity balance and sketch out possible improvements in the development of accented TTS. Overall, the proposed framework has the potential to improve the quality and flexibility of TTS models and could play a significant role in the development of more advanced TTS systems.

6. References

[1] J. C. Wells and J. C. Wells, *Accents of English: Volume 1*. Cambridge University Press, 1982.

[2] R. Lippi-Green, *English with an accent: Language, ideology, and discrimination in the US*. Routledge, 2012.

[3] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint:1409.0473*, 2014.

[4] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "One tts alignment to rule them all," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6092–6096.

[5] 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.

[6] 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.

[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," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[9] 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.

[10] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen *et al.*, "Hierarchical generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1810.07217*, 2018.

[11] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.

[12] M. Sam Ribeiro, J. Roth, G. Comini, G. Huybrechts, A. Gabryś, and J. Lorenzo-Trueba, "Cross-speaker style transfer for text-to-speech using data augmentation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6797–6801.

[13] J. Melechovsky, A. Mehrish, D. Herremans, and B. Sisman, "Learning accent representation with multi-level vae towards controllable speech synthesis," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 928–935.

[14] R. Liu, B. Sisman, G. Gao, and H. Li, "Controllable accented text-to-speech synthesis," *arXiv preprint arXiv:2209.10804*, 2022.

[15] S. Ding, G. Zhao, and R. Gutierrez-Osuna, "Accentron: Foreign accent conversion to arbitrary non-native speakers using zero-shot learning," *Computer Speech & Language*, vol. 72, p. 101302, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230821001029

[16] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting foreign accent speech without a reference," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2367–2381, 2021.

[17] Z. Wang, W. Ge, X. Wang, S. Yang, W. Gan, H. Chen, H. Li, L. Xie, and X. Li, "Accent and speaker disentanglement in many-to-many voice conversion," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[18] A. Aksënova, Z. Chen, C.-C. Chiu, D. van Esch, P. Golik, W. Han, L. King, B. Ramabhadran, A. Rosenberg, S. Schwartz *et al.*, "Accented speech recognition: Benchmarking, pre-training, and diverse data," *arXiv preprint arXiv:2205.08014*, 2022.

[19] F. Tong, T. Li, D. Liao, S. Xia, S. Li, Q. Hong, and L. Li, "The xmuspeech system for accented english automatic speech recognition," *Applied Sciences*, vol. 12, no. 3, 2022.

[20] Y. Qian, X. Gong, and H. Huang, "Layer-wise fast adaptation for end-to-end multi-accent speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2842–2853, 2022.

[21] S. Ghorbani and J. H. Hansen, "Domain expansion for end-to-end speech recognition: Applications for accent/dialect speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 762–774, 2023.

[22] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[23] X. Zang, F. Xie, and F. Weng, "Foreign accent conversion using concentrated attention," in *2022 IEEE International Conference on Knowledge Graph (ICKG)*, 2022, pp. 386–391.

[24] W. Quamer, A. Das, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Zero-shot foreign accent conversion without a native reference."

[25] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[26] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev, J. Levis, and R. Gutierrez, "L2-arctic: A non-native english speech corpus." in *INTERSPEECH*, 2018, pp. 2783–2787.

[27] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Communications, Computers and Signal Processing*, pp. 125–128, 1993.

[28] "Silero models:pre-trained enterprise-grade stt/tts models and benchmarks," accessed: 2022-07-10.

[29] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[30] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.