Cross-lingual Style Transfer with Conditional Prior VAE and Style Loss

Dino Ratcliffe, You Wang, Alex Mansbridge, Penny Karanasou, Alexis Moinet, Marius Cotescu

Alexa AI, Amazon

{dinoratc, wayou, mansbra, pkarana, amoinet, cotescu}@amazon.com

Abstract

In this work we improve the style representation for cross-lingual style transfer. Specifically, we improve the Spanish representation across four styles, Newscaster, DJ, Excited, and Disappointed, whilst maintaining a single speaker identity for which we only have English samples. This is achieved using Learned Conditional Prior VAE (LCPVAE), a hierarchical Variational Auto Encoder (VAE) approach. A secondary VAE is introduced, conditioned on one-hot encoded style information, resulting in a structured embedding space of the primary VAE. This places utterances of the same style in similar locations of the latent space irrespective of language. We also experiment with extending this model by incorporating a style loss. We perform subjective evaluations for style similarity using native Spanish speakers, and show an average relative improvement over the baseline of 3.5% with statistical significance (p-value<0.01) across all four styles. Interestingly the more expressive styles achieve a higher relative improvement of 4.4% compared to 2.6% for styles that are closer to neutral speech. We also demonstrate that this is whilst maintaining speaker similarity and in-lingual performance in all styles. Accent performance is maintained in three out of four styles with the exception of Excited, while naturalness performance is maintained in News and Disappointed styles.

Index Terms: cross-lingual style transfer, Learned Conditional Prior for VAE, style loss, neural TTS

1. Introduction

Through the popularity of AI assistants with primarily audio interfaces, Text To Speech (TTS) is having to represent a wide variety of textual content [1]. It is desirable to have these systems match the style of their speech with the textual content. However achieving this across languages whilst maintaining a fixed speaker identity is challenging.

Improvements have been made in TTS as deep learning approaches are developed and larger datasets become available [2]. This includes work focused on multi-lingual and multi-speaker models that aim to produce speech for a given speaker in multiple languages [3, 4, 5]. Such models are usually constructed by incorporating additional language specific information. In [6, 7], the input to the text encoder is a set of discrete phonological features derived from phonemes. Alternatively, a trainable utterance-level language embedding can be combined with the output of the text encoder before passing through the decoder [3, 4, 5]. Our work is based on such approaches as it doesn't require additional data labelling. We extend Tacotron 2, a popular architecture for Neural TTS [8], as our baseline.

Style transfer is frequently framed in the state-of-the-art TTS research as a disentanglement problem of the speech representation which is seen as a mixture of multiple factors including speaker, language, style, semantic meaning, etc. Extensive research work has been done regarding speech factor disentanglement with monolingual speech synthesis systems [9, 10, 11, 12], however in a cross-lingual domain this work is limited. In [13, 14, 15] style information is encoded and represented in a latent space through a reference Variational Auto Encoder (VAE). [11] uses adversarial learning to achieve speech disentanglement by minimising the correlation between different speech factors. The style information in speech representation can also be extracted and eliminated from the text encoder by utilising a pre-trained style classifier [10].

As mentioned above, VAEs have been widely utilized in the latest TTS research in acoustic space modelling [13], expressive speech synthesis [15], style transfer [14, 16], controllable speech synthesis [17] and intonation modelling [18]. There are also extensions applied to traditional VAEs such as adjusting the prior [19, 20] or stacking a sequence of VAEs together to better model hierarchical features in the training data [17, 21, 22]. Our paper is based on the Learned Conditional Prior VAE (LCP-VAE) architecture in [13], where a Conditioning Secondary VAE (Conditioning Secondary VAE (CSVAE)) is introduced to the standard VAE (Conditional Prior VAE (Conditional Prior VAE (CPVAE))) reference encoder. CSVAE takes as input a condition, which in [13] is a one-hot speaker ID vector, making the architecture able to model speaker information with controllability. We further explore this architecture by applying it on modelling style information in speech representations.

The representation of different speaking styles can also be improved by incorporating style loss associated with the model's ability to represent the style [23, 24, 25]. Style loss can be implemented by matching the layer activations of a style classifier network [25] or using gram matrices in order to quantify style [26, 27].

In this work we improve the style representation across languages when applied to Spanish speech whilst maintaining a certain English speaker identity. The styles investigated range from highly expressive (Excited, Disappointed) to less expressive (Newscaster and DJ). With Newscaster being a style that imitates traditional news broadcast presenter speech and DJ representing radio presenter speech. As explained above, we use LCPVAE for the first time in cross-lingual style modelling. Using LCPVAE we can force clustering of utterances based on style, and then represent variations between languages within those style clusters. This helps transferring styles between languages by encoding the same style, irrespective of language, in similar areas of the embedding space. This clustering is achieved by conditioning the CSVAE of LCPVAE on style one-hot encodings at utterance level instead of speaker identity encodings. In addition we propose combining LCPVAE with a style loss in order to force more style information to be embedded in the style encoding.

Our main contribution is using LCPVAE within the task of cross-lingual style transfer. We show, through MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) evaluations, that this approach increases the cross-lingual

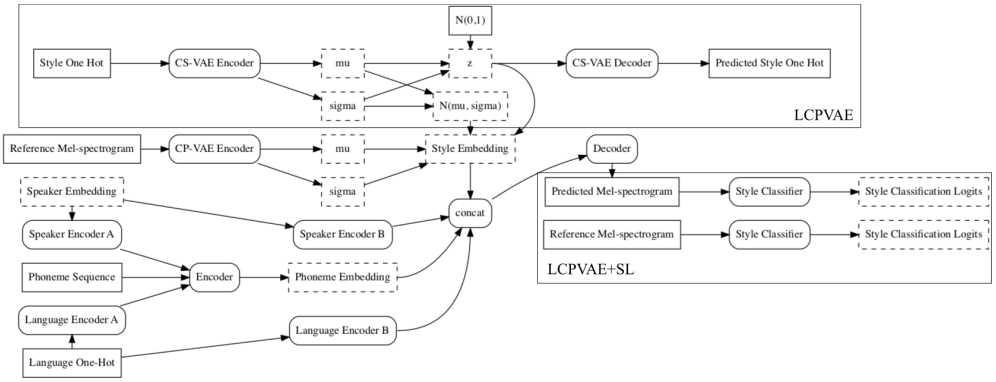


Figure 1: *This diagram represents the full architecture of the LCPVAE+SL model. When removing the style classifier this becomes the standard LCPVAE model. Removing the CSVAE and using a unit Gaussian prior for the CPVAE results the baseline model. Note seperate speaker and language encoders are used pre and post the encoder.*

style similarity from English to Spanish by 3.5% on average, ranging from 2.3% to 5.3% across all four investigated styles with statistical significance. This is whilst maintaining speaker similarity and in-lingual performance in all styles. Accent performance is maintained in three out of four styles with the exception of Excited, while naturalness performance is maintained in News and Disappointed styles. We also show that incorporating a style loss does not give a statistically significant improvement, with the majority of the improvement in cross-lingual performance coming from the use of LCPVAE.

2. Methodology

In this section we present each of the approaches used for cross-lingual style transfer and highlight their differences. For all experiments we use a phoneme representation of the text as input to the model. This representation has separate sets of phonemes for each language. We make use of speaker embeddings extracted from a speaker classifier trained on the LibriTTS dataset by Karlapati et al. [28]. All models output mel-spectrogram representations of the output speech.

2.1. Baseline

We use a baseline model based on the work by Zhang et al. on multilingual speech synthesis [5]. In their work they extended the Tacotron 2 [8] architecture with language and speaker embeddings, along with a residual encoder, and an adversarial speaker classifier. In order to better match the models across experiments we made a few changes to the architecture. In Figure 1 we show the full architecture of the models used in our experiments. The CPVAE in Figure 1 is the same as the residual encoder in Zhang et al. [5]. The CSVAE and style classifier are not used for the baseline model. Note that we have removed the adversarial classifier and also feed the speaker and language embeddings before and after the phoneme encoder. This is to factor out the speaker and language information from the latent space but leave the style information in. In our experiments we treat the residual encoder as an encoding of the style of the utterance. This residual encoder is a traditional VAE encoder with the standard unit Gaussian prior and using the standard Evidence Lower Bound (ELBO) loss. This model has been shown experimentally to produce high quality multi-lingual speech across speakers and languages. Finally, we remove the adversarial loss in order to understand the extensions we test in isolation as we do not use an adversarial loss with the LCPVAE models.

2.2. Learned Conditional Prior VAE (LCPVAE)

The LCPVAE architecture used is based on the work by Karanasou et al. [13] where they introduced a hierarchical architecture for the residual encoder. LCPVAE consists of a primary VAE, the CPVAE, that receives a mel-spectrogram reference and produces a single style embedding vector of length 256. This primary encoder is conditioned on the output of a secondary VAE, the CSVAE. This is done by defining the prior of the CPVAE to be that of the posterior of the CSVAE, and by applying the reparametrization trick [13]. This results in the model having two Kullback–Leibler Divergence (KLD) losses, one between the unit Gaussian and the CSVAE, and another between the CSVAE and the CPVAE. The relationship between the networks and their priors are shown in more detail in Figure 1. This approach has been previously used for multi-speaker TTS, with this work being the first time it is applied to a style transfer task. Instead of conditioning the CSVAE on the one-hot speaker encoding we condition it on the style of the utterance. The other modification made to LCPVAE is that we no longer concatenate the one-hot encoding to the mel-spectogram as input to the CPVAE. This was done to give the flexibility of running inference on the CPVAE for styles that were not present in the training data. With this approach we explicitly embed the hierarchy within the style embedding space, clustering utterances of the same style in similar regions irrespective of language. We believe this helps the model transfer style across languages.

2.3. LCPVAE + Style Loss

Finally we present an LCPVAE model extended with a style loss (SL). This style loss is composed of a style classifier that predicts the style of an utterance from the mel-spectogram. This classifier is pre-trained with the same training set as the TTS system. Once this classifier has been trained we utilise the style information learnt in order to define a style loss. Differently from Liu et al. in [25] we define a style loss given the output of the classifier, instead of matching the activations. Our style loss is defined as follows:

$$Loss_{style}(\mathbf{Y}, \hat{\mathbf{Y}}) = L_1(C_\theta(\mathbf{Y}), C_\theta(\hat{\mathbf{Y}})), \quad (1)$$

where $\mathbf{Y}$ represents the ground truth mel-spectrogram, $\hat{\mathbf{Y}}$ represents the mel-spectogram output from the model and C_θ represents a style classifier with the parameters θ. This loss is then summed with the standard Tacotron 2 loss and the negative ELBO.

3. Experiments

3.1. Training

We trained three models: **Baseline**, **LCPVAE** and **LCPVAE+SL**, using the architectures described in Sections 2.1, 2.2 and 2.3, respectively. All systems were trained on an internal dataset of high quality studio recordings described in Table 1. Even though we were only interested in evaluating 4 speaking styles, we included all available styles in the training set, to increase the variance in the data. All models were trained for 400,000 steps using a batch size of 32 utterances and all other hyper-parameters were held constant unless otherwise specified. For the baseline model we annealed the KLD loss linearly from 0 to 1 between steps 25,000 and 150,000, while for the LCPVAE we obtained the best results when the annealing was applied between steps 12,500 and 75,000. This value was also used for the LCPVAE+SL model.

Table 1: *Training dataset breakdown*

	English	Spanish	French	Total
Hours	411	175	32	618
Speakers	138	56	7	201
Styles	11	7	1	11

3.2. Evaluation

The goal of the proposed method is to produce styled speech in a language different than the one of the source speaker while maintaining the source speaker identity. We generate audio samples for the four styles in both English and Spanish by running inference with the CPVAE output replaced by pre-computed centroids. Each centroid is computed by taking the average embedding over 200 randomly selected training samples of the source speaker speaking in the target style (in English). For each style, we randomly select 50 sentences from the test set and synthesise them using the respective centroid. All stimuli are rated by 30 listeners.

We evaluate the systems' performance by running MUSHRA evaluations across four different axis: i) Style Transfer, ii) Naturalness, iii) Accent and iv) Speaker Similarity. In addition to the three stimuli, we include one hidden anchor in each test. For the Style Transfer test we use neutral TTS samples of the target speaker in Spanish (lower anchor). For naturalness, we use recordings of the target speaker in the target style in English (higher anchor). For accent and speaker similarity, we use recordings of a similar speaker speaking in Spanish as high and lower anchors, respectively. In the Naturalness test, we do not present any external reference and we ask the following question: "Please rate the audio samples in terms of their naturalness, ignoring any non-native accent issues. Do they sound like a human speaker?" In the Style Transfer and Speaker Similarity evaluations we present the listeners with a random recording of the target speaker speaking in English, in the corresponding style, while for Accent we present them with a random recording of a Spanish speaker. In these three tests, we ask the listeners to evaluate the stimuli by answering the following question: "Please rate how similar the audio samples sound to the reference in terms of their {style, accent, speaker}".

The listeners rate the stimuli on a scale from 0 to 100, where 0 represents the least desired result (e.g. completely different style) and 100 represents the most desired result (e.g. completely the same style). We ran t-tests with Bonferroni-Holm corrections at a 95% significance level to measure statistical significance with null-hypothesis being that the average MUSHRA scores of the two systems compared are equal.

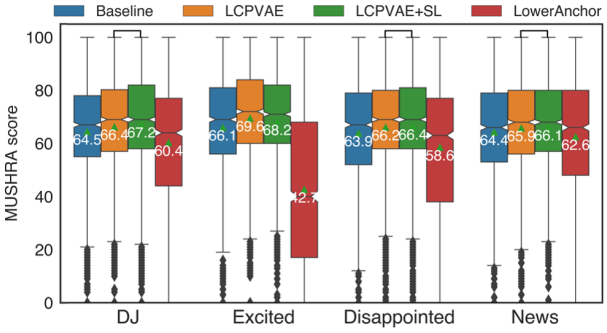


Figure 2: *MUSHRA Scores of Style Transfer.*

4. Results

We aim at improving the style transfer ability of the proposed systems while maintaining the same performance in Naturalness, Accent and Speaker Similarity compared with the baseline model. To make sure that our model doesn't introduce regression for in-lingual expressive speech synthesis, we also run evaluation with English samples in all four styles in the target speaker. The results didn't show significant regression for 14 out of 16 evaluations we conducted. We will not include the in-lingual results in this paper for brevity.

4.1. Style Transfer

We evaluate our systems in four different styles which are DJ, Excited, Disappointed and Newscaster. The MUSHRA scores for style transfer evaluation are shown in Figure 2. Both of our proposed systems LCPVAE and LCPVAE+SL outperform the baseline models significantly with p<0.01 across all four styles. The improvement of average score in percentage number and the p-value are shown in Table 2.

Table 2: *Improvement of Average MUSHRA Score over Baseline Model in Percentage for DJ, Excited, Disappointed and Newscaster Style. The p-value is computed between each model against the baseline.*

	DJ	Excited	Disappointed	News
LCPVAE	+2.8	+5.3	+3.5	+2.3
LCPVAE+SL	+4.1	+3.1	+3.9	+2.8
p-value	<0.001	<0.001	<0.001	<0.01

Note that although LCPVAE+SL outperforms LCPVAE in three out of four styles, the differences are not statistically significant. There is an insignificant increase in average score over LCPVAE models in DJ (+1.3%), Disappointed (+0.4%), and Newscaster style (+0.5%). However, there is a significant lower average score in Excited style (-2.2%). It thus seems that in a highly expressive style, like Excited, LCPVAE alone performs better. For this style, we observed that the training process of LCPVAE+SL model is not as stable as LCPVAE. We also examined the latent embedding space but did not see any clear difference between the two models in modelling the Excited style. We hypothesize that the undesired performance of LCPVAE+SL can be attributed to two reasons. First of all, the style loss architecture that we currently apply is still in its most rudimentary form where we compare the final output of the style

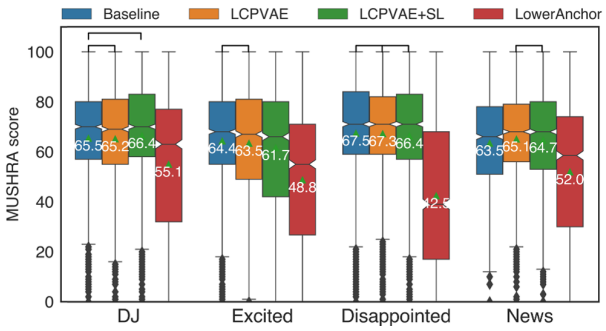


Figure 3: *MUSHRA Scores of Speaker Similarity.*

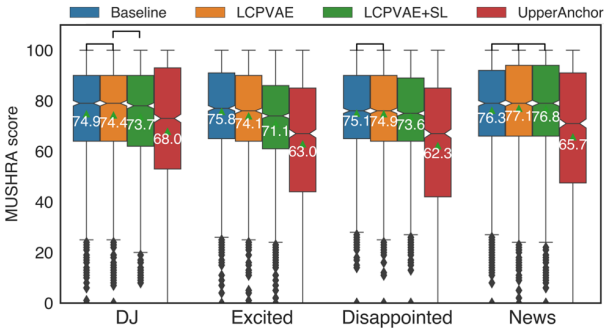


Figure 4: *MUSHRA Scores of Accent.*

classifier directly. This can be problematic as the style classifiers might be using content information instead of style information for classification. Using lower layers instead of final output would be more reasonable in such case. The second reason is lack of extensive hyper-parameter tuning. The results are reported with the first model that we are able to train until convergence without any VAE collapse and there is still space for exploration. Further investigation in style loss design and an extensive hyper-parameter search would likely improve its performance in style transfer. This is left as future work.

4.2. Naturalness, Accent and Speaker Similarity

In Figure 3, 4 and 5 we show the MUSHRA scores of the evaluated systems within each style respectively. For Speaker Similarity and Accent in Figure 3 and 4, the results of LCPVAE models either significantly outperform the baseline model (Newscaster style in Speaker Similarity) or show no significant difference from the baseline, with Excited style being the only exception. For Naturalness as shown in Figure 5, we have observed statistically significant regression in DJ (-1.7%) and Excited (-4.8%) style, and no significant difference in Disappointed and Newscaster style. Note that all of the candidate models in the evaluation of Accent outperform the upper anchor system which is recordings from a native Mexican Spanish speaker in neutral style. We suspect that this is because the speaker in the recordings has less strong Mexican accent compared with the synthetic speech, as the training data contains speech from many different native speakers. We have observed similar patterns in all three metrics that the proposed LCPVAE-based systems show a regression in the Excited style. The results confirm the findings in [14] that maintaining voice quality for voices with higher expressiveness is more difficult in general. A promising next step would be to test the architecture with more highly expressive styles for further investigation.

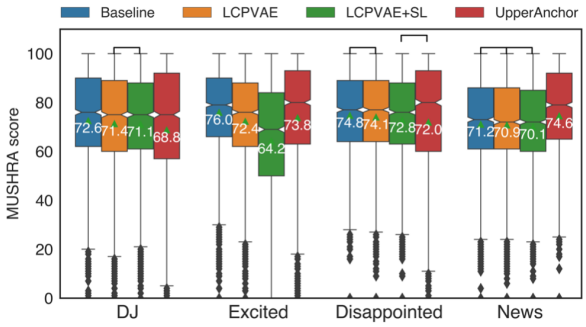
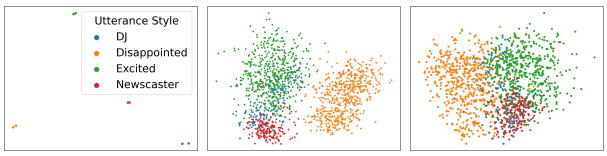


Figure 5: *MUSHRA Scores of Naturalness.*



(a) *LCPVAE CS* (b) *LCPVAE CP* (c) *Baseline*

Figure 6: *Plots of learnt embeddings for the Baseline and LCPVAE, compute on held out examples. Dimension reduction done using Principle Component Analysis (PCA).*

4.3. Qualitative Evaluation

Here we investigate the structure of the learnt embedding space. In Figure 6 we show the embedding space of the target speaker given the output of the VAEs for the Baseline model and LCPVAE. The output of the LCPVAE+SL is not shown due to similarity to the LCPVAE model. We can see that the clustering of the CSVAE is very similar to that observed by Karanasou et al. when applied to speaker identity [13]. Interestingly we see more separation between the styles in the CPVAE embedding space than the Baseline, this shows that conditioning on the CSVAE is having a positive effect on clustering the styles. We also observed local clustering of language within styles, although not shown. We believe this increase in clustering performance is driving the better style reproduction in speech.

5. Conclusions

In this work we have shown that giving explicit hierarchy to the style embedding space using LCPVAE can improve cross lingual style transfer. With MUSHRA evaluations we showed significant improvements for cross-lingual style representation in all 4 styles, DJ (2.8%), Excited (5.3%), Disappointed (3.5%) and Newscaster (2.3%). We also demonstrated that this improvement does not harm speaker similarity and in-lingual style representation. We did however notice a drop in naturalness (DJ, Excited) and Accent (Excited) for some styles. In addition we presented results with combining style loss with the LCPVAE approach, here we saw small however insignificant improvement over the standard LCPVAE model in three of the 4 styles.

For future work we believe that a further exploration into the use of style loss, as well as incorporating more style information into the CSVAE could provide further gains in cross-lingual style transfer. We also think that a more thorough investigation into the irregular performance increase and subsequent naturalness decrees between styles should be explored in order to gain more insights into the weaknesses of these models.

6. References

[1] M. B. Hoy, "Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants," *Medical Reference Services Quarterly*, vol. 37, no. 1, pp. 81–88, 2018, pMID: 29327988. [Online]. Available: https://doi.org/10.1080/02763869.2018.1404391

[2] 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," in *Proc. Interspeech 2019*, 2019, pp. 1526–1530.

[3] E. Casanova, J. Weber, C. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone," *arXiv preprint arXiv:2112.02418*, 2021.

[4] J. Yang and L. He, "Towards Universal Text-to-Speech," in *INTERSPEECH*, 2020, pp. 3171–3175.

[5] Y. Zhang, R. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning," in *Proc. Interspeech 2019*, 09 2019, pp. 2080–2084.

[6] G. Maniati, N. Ellinas, K. Markopoulos, G. Vamvoukakis, J. S. Sung, H. Park, A. Chalamandaris, and P. Tsiakoulis, "Cross-Lingual Low Resource Speaker Adaptation Using Phonological Features," in *Proc. Interspeech 2021*, 2021, pp. 1594–1598.

[7] M. Staib, T. H. Teh, A. Torresquintero, D. S. R. Mohan, L. Foglianti, R. Lenain, and J. Gao, "Phonological Features for 0-Shot Multilingual Speech Synthesis," in *Proc. Interspeech 2020*, 2020, pp. 2942–2946.

[8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2018, pp. 4779–4783, iSSN: 2379-190X.

[9] X. Dai, C. Gong, L. Wang, and K. Zhang, "Information Sieve: Content Leakage Reduction in End-to-End Prosody Transfer for Expressive Speech Synthesis," in *Proc. Interspeech 2021*, 2021, pp. 131–135.

[10] X. He, J. Chen, G. Rizos, and B. W. Schuller, "An Improved StarGAN for Emotional Voice Conversion: Enhancing Voice Quality and Data Augmentation," in *Proc. Interspeech 2021*, 2021, pp. 821–825.

[11] J. Wang, J. Li, X. Zhao, Z. Wu, S. Kang, and H. Meng, "Adversarially Learning Disentangled Speech Representations for Robust Multi-Factor Voice Conversion," in *Proc. Interspeech 2021*, 2021, pp. 846–850.

[12] X. Xu, L. Shi, J. Chen, X. Chen, J. Lian, P. Lin, Z. Zhang, and E. R. Hancock, "Two-Pathway Style Embedding for Arbitrary Voice Conversion," in *Proc. Interspeech 2021*. York, 2021, pp. 1364–1368.

[13] P. Karanasou, S. Karlapati, A. Moinet, A. Joly, A. Abbas, S. Slangen, J. Lorenzo-Trueba, and T. Drugman, "A Learned Conditional Prior for the VAE Acoustic Space of a TTS System," in *Interspeech 2021*. ISCA, Aug. 2021, pp. 3620–3624.

[14] V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, "Using VAEs and Normalizing Flows for One-shot Text-to-speech Synthesis of Expressive Speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6179–6183.

[15] F. Yang, J. Luan, and Y. Wang, "Improving Emotional Speech Synthesis by Using SUS-Constrained VAE and Text Encoder Aggregation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8302–8306.

[16] 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.

[17] W.-N. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Cao, and Y. Wang, "Hierarchical Generative Modeling for Controllable Speech Synthesis," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rygkk305YQ

[18] Z. Hodari and O. Watts, "Using Generative Modelling to Produce Varied Intonation for Speech Synthesis," in *10th ISCA Speech Synthesis Workshop*, 09 2019, pp. 239–244.

[19] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, A. Rosenberg, B. Ramabhadran, and Y. Wu, "Generating Diverse and Natural Text-to-speech Samples Using a Quantized Fine-grained VAE and Autoregressive Prosody Prior," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6699–6703.

[20] Y. Ren, J. Liu, and Z. Zhao, "PortaSpeech: Portable and High-Quality Generative Text-to-Speech," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[21] S. Zhao, J. Song, and S. Ermon, "Learning Hierarchical Features from Deep Generative Models," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 4091–4099. [Online]. Available: https://proceedings.mlr.press/v70/zhao17c.html

[22] P. Liu, Y. Cao, S. Liu, N. Hu, G. Li, C. Weng, and D. Su, "VARA-TTS: Non-autoregressive Text-to-speech Synthesis Based on Very Deep VAE with Residual Attention," *arXiv preprint arXiv:2102.06431*, 2021.

[23] X. An, F. K. Soong, and L. Xie, "Improving Performance of Seen and Unseen Speech Style Transfer in End-to-End Neural TTS," in *Proc. Interspeech 2021*, 2021, pp. 4688–4692.

[24] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable Emotion Transfer for End-to-end Speech Synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[25] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive TTS Training with Frame and Style Reconstruction Loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.

[26] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-time Style Transfer and Super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[27] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer using Convolutional Neural Networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[28] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "CopyCat: Many-to-Many Fine-Grained Prosody Transfer for Neural Text-to-Speech," *Proc. Interspeech 2020*, pp. 4387–4391, 2020.