A STUDY ON ALTERING THE LATENT SPACE OF PRETRAINED TEXT TO SPEECH MODELS FOR IMPROVED EXPRESSIVENESS

Mathias Vogel

Media Technology Center, ETH Zürich, Switzerland

ABSTRACT

This report explores the challenge of enhancing expressiveness control in Text-to-Speech (TTS) models by augmenting a frozen pretrained model with a Diffusion Model that is conditioned on joint semantic audio/text embeddings. The paper identifies the challenges encountered when working with a VAE-based TTS model and evaluates different image-to-image methods for altering latent speech features. Our results offer valuable insights into the complexities of adding expressiveness control to TTS systems and open avenues for future research in this direction.

Index Terms— Machine Learning, Signal Processing, Text-to-Speech Synthesis

1. INTRODUCTION

Significant progress has been made in the development of Text-to-Speech (TTS) systems, with models such as VITS [1] achieving a mean opinion score comparable to that of genuine speech recordings. However, the difficulty of having precise control over the prosodic features of generated speech samples remains an unsolved issue. Many TTS models lack mechanisms to control prosodic and emotional nuances, which are essential for a wide range of applications.

In this paper, we present an exploratory study in which we enhance the VITS model with expressiveness control by adding a Denoising Diffusion Model (DDM) [2] conditioned on joint audio/text embeddings such as CLAP embeddings [3] to alter the latent VITS encodings. We chose DDM because these models are known to be easy to condition and are currently state-of-the-art in many computer vision tasks [4, 5, 6]. The final goal is to be able to change the generated speech by providing a target style by either providing a recording by text promts describing the style. Contrary to our expectations, the method did not produce the desired improvements in expressiveness control. However, we believe that the findings of our study offer valuable insight into the complexities and challenges associated with adding expressiveness control to TTS systems. Our findings could help future research in designing systems that allow greater control over prosodic and emotional features.

Our contributions are as follows:

1. We identify challenges of working with pretrained VAE based TTS models.
2. We apply and compare different image-to-image methods to change latent speech features, highlighting their strengths and weaknesses.
3. We open up discussion on further research directions of controlled emotional TTS.

2. METHOD PRELIMINARY

We chose the VITS model as our backbone TTS model due to its fast inference speed and quality. This model is trained using adversarial learning, as illustrated in fig. 1. The architecture consists of several parts, namely:

1. A conditional variational auto-encoder (CVAE) with a WaveGlow [7] based encoder and a HiFi-GAN [8] based decoder acting as a neural vocoder. The CVAE embeds a linear spectrogram x_{lin} in a lower dimensional space Z to lower the computational costs of the model.
2. A normalizing flow that provides an invertible mapping between the complex distribution Z and the simpler distribution $f_\theta(\mathrm{Z})$.
3. A text encoder that processes input phonemes c_{text} to statistics μ_θ and σ_θ that represent the probabilities of learned representations h_{text} given the input phonemes and the spectrogram x_{lin} of the corresponding audio.
4. A stochastic duration predictor that learns the length of the phoneme representations.

Figure 2 illustrates the use of VITS during inference, where the model does not require the posterior encoder component of the CVAE and directly transforms a text input into speech.

Work done as part of an internship at ETHZ MTC

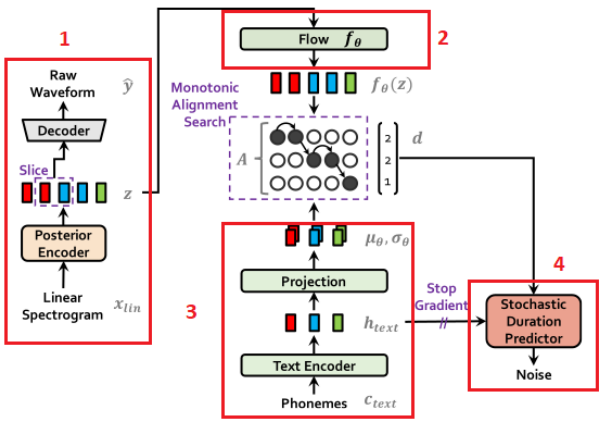

Fig. 1. The training procedure of VITS.

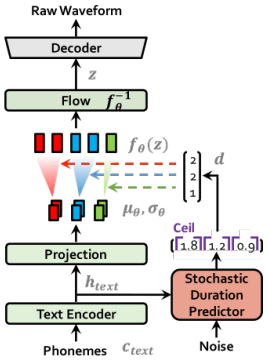

Fig. 2. VITS inference.

3. METHOD

In our proposed method, we alter the CVAE bottleneck embeddings Z of VITS. Because samples of Z resemble a spectrogram fig. 3, we apply approaches from Image-to-Image translation.

Fig. 3. Visualization of a latent sample Z which reminds of a spectrogram.

The goal is to transform a neutral Z into a stylized Z' while preserving the content. To be able to control the style, we condition the diffusion process on semantic text-audio embeddings c obtained using CLAP. CLAP is trained using a contrastive loss, which should result in embeddings that are close together for audio recordings and text prompts that match semantically, such as a recording of someone shouting and the phrase "person shouting". Since there are no semantic descriptions available for the LJS [9] and VCTK [10] datasets on which VITS was pre-trained, we use embeddings of the audio only during training. However, a data set with style annotation would be very beneficial.

We obtain training data for Z and Z' using different modules of VITS. To generate Z, we simply apply the frozen VITS text encoder to the transcript of a speech recording. We denote the resulting embeddings by Z_{text}. Encoding the corresponding speech recording with the VITS PosteriorEncoder we obtain Z_{audio} embeddings. Because Z_{audio} is produced from a speech recording, it should contain more prosodic features than Z_{text} since Z_{text} just matches the average style of the speaker. Both Z_{audio} and Z_{text} are of shape [C, H, W], where zero padding is applied to the width dimension when needed. Figure 4 shows a complete sketch of our training method.

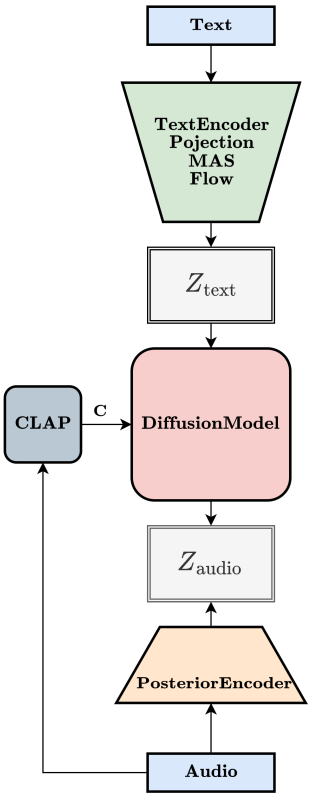

Fig. 4. Training of the proposed method.

4. EXPERIMENTAL EVALUATION

The resulting audio of all the experiments reported can be heard on GitHub. We divide the results by their Image-to-Image modeling approach. Independent of the

approach, we train our models using the Adam optimizer, a constant learning rate of 1e-4, and batch size of 64. We train all our models for around 50k steps using the mean-squared error loss. We use the v-objective introduced in [11] and DDIM sampling [12] using 10 sampling steps, which was experimentally found to be a good compromise between speed and quality.

4.1. Palette

For our baseline we adapt the Palette [13] Image-to-Image diffusion model for our task by allowing the U-Net [14] backbone, inspired by recent works [15, 16], to process non-squared data effectively. We add the information from Z_{text} by channel-wise concatenation with diffusion noise ϵ for each diffusion timestep such that the input of the model has a shape of [B, 2C, H, W]. Conditioning on the CLAP embedding c is implemented through cross-attention following [6].

The resulting speech samples using this method do not contain any intelligible content, although there is a certain melody and style to them. The potential cause of this could be that the model is only able to concentrate on style alterations when the content of the source Z_{text} and the target Z_{audio} are the same. However, due to the stochastic nature of VITS, specifically its Stochastic Duration Predictor, the size of the source and target latent spectrogram is not the same and sometimes differs by as much as 50%. This could potentially lead the training objective to consist mainly of misalignment errors rather than differences in style. We hypothesize that conditioning by channel-wise concatenation does not lead to good results, as this approach originates from image super-resolution [17, 18], where the underlying content is aligned by definition. Instead, the content could be provided to the model using a pre-trained model such as Wav2Vec [19] and concatenate the extracted embeddings with the CLAP embeddings.

However, even if we were capable of providing the content and target style to the diffusion model in an effective manner, there is no assurance that the final audio generated by the altered VITS still matches our target style. This is because the VITS decoder is trained as a person-specific vocoder, which could remove potential style changes and produce the same average person-specific speech. This hypothesis is tested in the following experiment, where we also explore a more advanced Image-to-Image method.

4.2. I2SB

I2SB stands for Image-to-Image Schrödingers Bridge [20] and is capable of producing state-of-the-art results in many image-to-image tasks such as deblurring, (freeform) inpainting or super-resolution. In contrast to the Palette diffusion process, which maps isotropic noise to a sample by conditioning on an image via concatenation, I2SB does not rely on concatenation and directly maps one distribution $p_0(x)$ into another distribution $p_1(x)$. This process, as well as the difference to approaches such as Palette, is visualized in fig. 5. I2SB allows for more general image-to-image translation tasks by adapting the way the bridge is constructed and the way to condition the process.

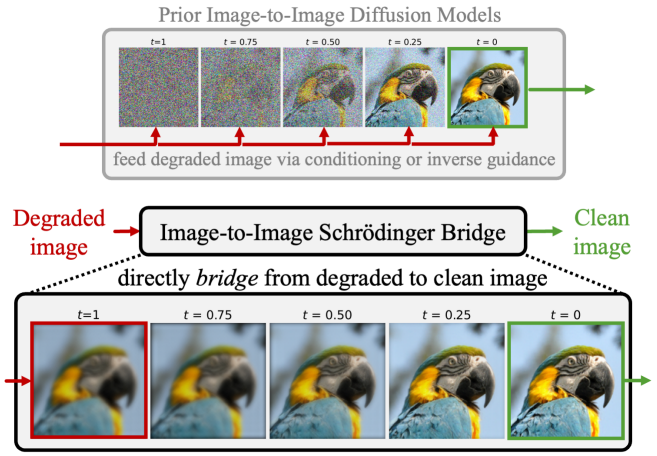

Fig. 5. Comparison of a traditional image-to-image translation task shown on top with the I2SB method, where a direct bridge between the initial distribution and the target distribution is learned.

We use the same U-Net backbone as described in section 4.1 for an I2SB pipeline using the --add-x1-noise and --ot-ode options. These settings are suggested by the authors of I2SB when dealing with more challenging tasks where the two distributions $p_0(x)$ and $p_1(x)$ are potentially very different from each other. We could validate the suggestions experimentally.

The resulting speech samples using the I2SB approach are qualitatively similar to the Palette approach. The samples contain little noise, but the content is inaudible, which can be heard on our project GitHub. We also evaluate the impact of Classifier-Free Guidance (CFG) [21] which shows that a higher guidance scale changes both the content and the style, however, only the style is expected to change. This experiment shows that our models use the CLAP embedding in a different way than intended. This could be due to the aforementioned lack of content alignment between Z_{text} and Z_{audio} which leads to the model trying to extract missing content from the CLAP embeddings instead of extracting only prosodic speech features.

In addition to changing the CFG scale, we also con-

ducted an experiment to understand the impact of the VITS vocoder. For this experiment, we used ground truth samples from two different speakers A and B of the VCTK test data set. The samples have a different content. We denote ground-truth audio recordings as Audio_A and Audio_B with their corresponding transcripts T_A and T_B and CLAP audio embeddings CLAP_A and CLAP_B. If we want to recreate speaker A we would condition the diffusion model on CLAP_A and set the speaker id to A in the VITS vocoder. On the other hand, if the goal is to obtain a speech sample of speaker A in the style of speaker B, we would condition the diffusion model on CLAP_B while conditioning the VITS vocoder on the id of A. By conditioning the diffusion model on CLAP_B and the VITS vocoder on the id of B, we should obtain a speech sample that sounds exactly like speaker B.

The samples in our GitHub demonstrate that this approach does not work as expected. Conditioning the diffusion model on a CLAP embedding of a different speaker changes the content more than the style, while changing the conditioning id of the VITS vocoder changes the style significantly. These findings suggest that an approach that aims to change the latent space Z of a VAE to achieve style change can not work as intended due to the fact that the decoder is conditioned on a speaker id and that the latent space Z contains mostly information about the speech content instead of style information.

5. CONCLUSION

In this report, we conducted an in-depth examination of adding expressiveness control to Text-to-Speech (TTS) systems, specifically focusing on the VITS model. Our approach aimed to modify the latent embeddings (Z) of a frozen VITS model using a Diffusion Model conditioned on joint audio/text embeddings like CLAP. The goal was to offer a mechanism for controlling the expressiveness of speech by providing a speech sample of reference style or a text prompt describing the desired style.

Contrary to our expectations, our method did not produce significant improvements in expressiveness control. Despite this, the research provides several insights into the complexities and challenges related to augmenting TTS systems with expressiveness control features.

Our contributions can be summarized as follows:

1. We identified challenges tied to working with pretrained VAE-based TTS models like VITS, including the limitations of their latent spaces in encoding stylistic features.

2. We applied and evaluated different image-to-image translation methods, including Palette and I2SB, to alter latent speech features, thus revealing their strengths and weaknesses.

3. We opened the door to future research directions in the field of controlled emotional TTS by outlining the complexities involved.

The most significant limitation encountered was the mismatch in content alignment between Z_{text} and Z_{audio}, which prevented our diffusion models from effectively altering only the style characteristics. Our experiments also demonstrated that the speaker-specific conditioning of the VITS vocoder significantly impacts the stylistic outcome, raising questions about the feasibility of changing the style via the latent space Z.

Given these findings, future research could explore alternative methods to modify the latent spaces of TTS models. This could involve the use of separate and disentangled latent spaces for style and content. Another approach could focus on directly training a style-conditional diffusion model to map from a style-neutral latent space Z to speech audio.

References

[1] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," in Proceedings of Machine Learning Research, 2021, vol. 139.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems, 2020, vol. 2020-December.

[3] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," 2023.

[4] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video Diffusion Models," 4 2022.

[5] E. Hoogeboom, J. Heek, and T. Salimans, "simple diffusion: End-to-end diffusion for high resolution images," 1 2023.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2022-June, pp. 10674–10685, 12 2021.

[7] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A Flow-based Generative Network for Speech Synthesis," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2019, vol. 2019-May.

[8] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in Advances in Neural Information Processing Systems, 2020, vol. 2020-December.

[9] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[10] C. Veaux, J. Yamagishi, and K. MacDonald, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit," 2017.

[11] T. Salimans and J. Ho, "PROGRESSIVE DISTILLATION FOR FAST SAMPLING OF DIFFUSION MODELS," in ICLR 2022 - 10th International Conference on Learning Representations, 2022.

[12] J. Song, C. Meng, and S. Ermon, "DENOISING DIFFUSION IMPLICIT MODELS," in ICLR 2021 - 9th International Conference on Learning Representations, 2021.

[13] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-Image Diffusion Models," 2022.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 9351.

[15] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," in Advances in Neural Information Processing Systems, 2021, vol. 11.

[16] A. Nichol and P. Dhariwal, "Improved Denoising Diffusion Probabilistic Models," 2 2021.

[17] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image Super-Resolution via Iterative Refinement," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 4, 2023.

[18] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded Diffusion Models for High Fidelity Image Generation," Journal of Machine Learning Research, vol. 23, 2022.

[19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in Advances in Neural Information Processing Systems, 2020, vol. 2020-December.

[20] G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar, "I^2SB: Image-to-Image Schrödinger Bridge," arXiv preprint arXiv:2302.05872, 2023.

[21] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," in NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.