On the Semantic Latent Space of Diffusion-Based Text-to-Speech Models

Miri Varshavsky-Hassid* Roy Hirsch* Regev Cohen Tomer Golany
Daniel Freedman Ehud Rivlin

Verily AI
{mirivar, royhirsch, regevcohen}@google.com

Abstract

The incorporation of Denoising Diffusion Models (DDMs) in the Text-to-Speech (TTS) domain is rising, providing great value in synthesizing high quality speech. Although they exhibit impressive audio quality, the extent of their semantic capabilities is unknown, and controlling their synthesized speech's vocal properties remains a challenge. Inspired by recent advances in image synthesis, we explore the latent space of frozen TTS models, which is composed of the latent bottleneck activations of the DDM's denoiser. We identify that this space contains rich semantic information, and outline several novel methods for finding semantic directions within it, both supervised and unsupervised. We then demonstrate how these enable off-the-shelf audio editing, without any further training, architectural changes or data requirements. We present evidence of the semantic and acoustic qualities of the edited audio, and provide supplemental samples: https://latent-analysis-grad-tts.github.io/speech-samples/.

1 Introduction

Denoising Diffusion Models (DDMs) (Sohl-Dickstein et al., 2015) have emerged as a powerful generative tool across a broad variety of tasks and domains. In particular, Text-to-Speech (TTS) systems based on diffusion have shown high-quality speech generation capabilities (Huang et al., 2022b; Shen et al., 2023). Although these exhibit improved quality, the extent to which they capture semantic information is yet to be uncovered, and the ability to *control* the vocal properties (e.g. volume, pitch, gender) of their generated speech is limited. Uncovering the semantic capabilities of TTS diffusion models will allow editing the properties of synthesized speech, which is essential in real-world applications, such as human-machine interaction.

Diffusion-based TTS methods, such as WaveGrad and Diff-Wave, condition the generation process on mel-spectogram input (Chen et al., 2020; Kong et al., 2020b). More recent advances such as Diff-TTS, WaveGrad2, and Grad-TTS condition the generation process on textual input (Jeong et al., 2021; Chen et al., 2021; Popov et al., 2021), and works like DiffGAN-TTS, FastDiff and ProDiff (Liu et al., 2022; Huang et al., 2022a,b) prioritize generation efficiency and expressiveness.

Beyond efficiency, researchers have explored DDMs for controllable and expressive TTS. PromptTTS (Guo et al., 2023b) and NaturalSpeech 2 (Shen et al., 2023) employ text prompts and speech prompts, respectively, to control speech style and content. In both methods, the conditional denoiser must undergo a specialized training process. Other methods for controlling the vocal characteristics require large quantities of annotated samples (Guo et al., 2023a) or retraining (Kim et al., 2022). We propose a speech editing method that requires no additional data or training and can be applied to any frozen diffusion-based TTS model that incorporates a bottleneck.

In the image synthesis domain, Kwon et al. (2022) recently discovered a semantically meaningful latent space, named *h-space*, providing versatile semantic editing capabilities. This discovery was further explored by Haas et al. (2023), who proposed methods for identifying semantic directions. To the best of our knowledge, despite the widespread adoption of diffusion models for TTS in recent years, the existence of a hidden semantic space has not been examined in the speech synthesis domain. This raises intriguing questions regarding the possibility of facilitating latent space arithmetics for audio editing.

In this work we investigate the existence of a semantic space within diffusion-based TTS systems. We study the properties of *h-space* in pretrained TTS models and uncover its acoustically-

[*] Equal contribution

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 246–255
August 11-16, 2024 ©2024 Association for Computational Linguistics

semantic characteristics. Then, we propose novel methods for semantic speech editing through both supervised and unsupervised latent space arithmetics, inspired by Haas et al. (2023) and adapted to the speech synthesis domain for the first time. Our work offers intuitive and efficient audio editing techniques that require neither classifier guidance (Guo et al., 2023a), model retraining (Kim et al., 2022), optimization, speech prompts nor any architecture modifications. To validate our methods, we present extensive experiments that demonstrate effective and high-quality edited speech synthesis.

2 Methods

2.1 Denoising Diffusion Models

DDMs generate realistic data by iteratively removing noise, and are applicable to various modalities like images, audio, and text (Ho et al., 2020). Initially formulated as Markov chains, DDMs can be unified under stochastic differential equations (SDEs) (Song et al., 2020) and adapted for TTS (Popov et al., 2021). DDMs consist of two processes: forward diffusion and reverse diffusion. The forward process transforms any data distribution to a Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ via an SDE. The reverse diffusion process is defined by another SDE:

$$d\mathbf{x}_t = \frac{\beta_t}{2}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{x}_t) - s(\mathbf{x}_t)\right) dt + \sqrt{\beta_t} d\mathbf{w}_t$$

where w_t is a Brownian motion, β_t is a predefined noise schedule, and $s(\mathbf{x}_t) = \nabla \log p_t(\mathbf{x}_t)$ is the score function of the probability density function p_t of $\mathbf{x}_t$. The reverse process is typically solved via the Euler-Maruyama scheme (Kloeden et al., 1992), discretizing the time interval $[0, 1]$ into T time-steps. By training a denoising neural network $s_t^\theta(\mathbf{x}_t) \approx s(\mathbf{x}_t)$ to estimate the true score function, we can sample from the target data distribution. Within TTS systems, DDMs are utilized as acoustic models, vocoders, or as end-to-end solutions.

2.2 Semantic Audio Editing via Latent Space Manipulation

We aim to discover a semantic latent space within frozen diffusion-based TTS models.We build upon the work of Kwon et al. (2022) who introduced a semantic latent space in image diffusion models. Leveraging the standard implementation of the denoising network, $s_t^\theta(\cdot)$, as a U-Net architecture (Ronneberger et al., 2015) in state-of-the-art models, Kwon et al. (2022) examined the deepest feature maps, residing at the bottleneck of the network (visualized in Figure 1). These features are subsequently concatenated across all T time-steps to construct the following latent code:

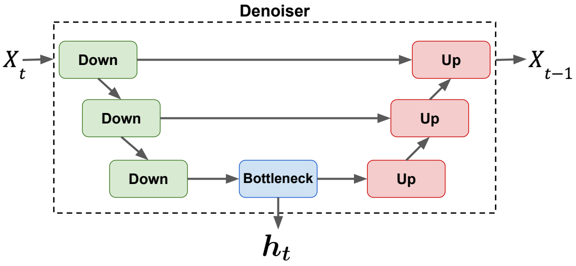


Figure 1: The *h-space* of a diffusion model is defined as the concatenation of the bottleneck activations of the U-Net architecture.

$$\mathbf{h} \triangleq \mathbf{h}_{T:1} = \texttt{concat}(\mathbf{h}_T, \mathbf{h}_{T-1}, \ldots, \mathbf{h}_1) \quad (1)$$

This approach yields the *h-space*: a latent space exhibiting favorable properties for versatile semantic editing and quality enhancement of images (Kwon et al., 2022; Haas et al., 2023).

We adapt the concept of *h-space* to the domain of TTS, demonstrating it encapsulates semantic information and performing semantic editing of synthesized speech through simple latent space arithmetics. Specifically, given a speech sample whose features are $\mathbf{h} \triangleq \mathbf{h}_{T:1}$ and a direction $\mathbf{v} \triangleq \mathbf{v}_{T:1}$, associated with desired acoustic attributes, we propose the following editing process:

$$\mathbf{h}^{edit} \triangleq \mathbf{h}_{T:1}^{edit} = \mathbf{h}_{T:1} + \lambda \cdot \mathbf{v}_{T:1} \quad (2)$$

where λ controls edit intensity, and both addition and scaling are element-wise. Replacing the latent code $\mathbf{h}$ with $\mathbf{h}^{edit}$ during the generation process embodies the synthesized speech with the acoustic attributes related to the chosen editing direction.

Having established the editing framework, we next derive editing directions via the following (illustrated in Figure 2):

Supervised Approach. Given a pre-trained TTS model and a specific text prompt, we generate m paired samples $\{(\mathbf{x}_{(k)}^+, \mathbf{x}_{(k)}^-)\}_{k=1}^m$ characterized by the presence or absence of a desired attribute. Denoting their matching latent codes by $\{(\mathbf{h}_{(k)}^+, \mathbf{h}_{(k)}^-)\}_{k=1}^m$, we define a semantic direction towards this attribute as

$$\mathbf{v} \triangleq \Delta\mathbf{h} = \frac{1}{m}\sum_{k=1}^{m}(\mathbf{h}_{(k)}^+ - \mathbf{h}_{(k)}^-) \quad (3)$$

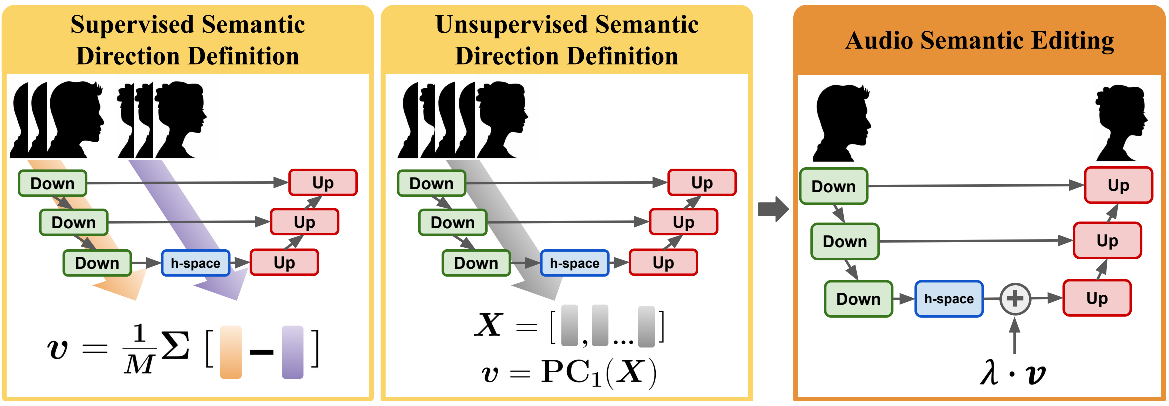


Figure 2: We propose a simple yet effective semantic audio-editing method. A latent semantic direction is defined either in a supervised or an unsupervised manner, and the corresponding speech attribute is edited by applying that direction to the latent space during the generation process of a new speech sample. The method is demonstrated with the male-to-female editing direction.

Unsupervised Approach. For a given text input, we generate speech samples and extract their bottleneck features $\{\mathbf{h}_t^{(i)}\}_{i=1}^n$ for each time-step $t \in [1, T]$. Applying PCA per time-step, we define the editing direction $\mathbf{v}^{(j)}$ as a concatenation of the jth principal components across time-steps. Surprisingly, the main principle components display clear semantic attributes as gender and intensity. The above framework unlocks semantic editing in diffusion-based TTS models, facilitating expressive and diverse speech synthesis.

3 Experimental Results

3.1 Implementation Details

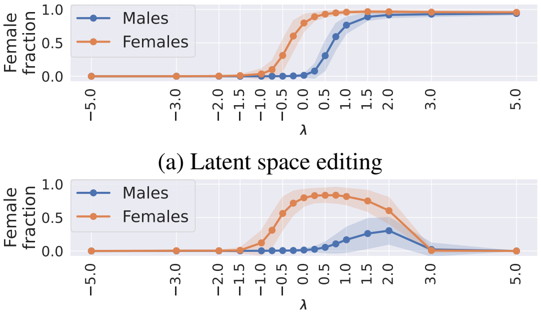
For demonstration, we use Grad-TTS (Popov et al., 2021), a recently published publicly available diffusion-based TTS model, trained on LibriTTS (Zen et al., 2019). However, our method can also be applied to any other unguided diffusion-based TTS model that contains a bottleneck. Grad-TTS takes a text and a speaker embedding as input, and generates a clean mel-spectogram through a U-Net-based denoiser. We use 10 diffusion timesteps for mel-spectogram generation, as suggested by Grad-TTS authors, followed by the Universal HifiGan vocoder (Kong et al., 2020a) for waveform generation.

3.2 Supervised Latent Space Editing

We begin our analysis by exploring the semantic-capturing capabilities of *h-space* using the per-speaker gender annotations available for LibriTTS. Capturing the latent code during all timesteps of the generation process and following Equation 3, we calculate the male-to-female latent direction, and utilize it for audio editing as outlined in Equation 2. As the latent vectors' lengths vary with the input texts, editing direction is defined per text. For a comparable baseline, we use another, simpler, approach for gender-editing: manipulating the speaker embedding, which is provided to the model as an input. We calculate the male-to-female direction in the speaker embedding space in a similar manner by averaging the differences of speaker embeddings between pairs of male and female speakers. The input speaker embedding is modified by adding this direction with different scales (λ). We provide supplemental samples, demonstrating the suggested audio editing methods: https://latent-analysis-grad-tts.github.io/speech-samples/.

Semantic properties evaluation. We fine-tuned a speech gender classifier (Bhamidipati, 2023) on Grad-TTS outputs, acknowledging the different quality of synthesized speech compared to human-recorded samples. Then, we applied gender editing via both latent space and speaker embedding editing using varying λ values, across the first 50 texts of the LibriTTS test set and all 247 speakers. In Figure 3 we report the fraction of samples classified as female for each λ value, averaged across input male and female speakers separately. Latent space editing exhibits a monotonic behavior with more samples classified as female as λ increases. On the contrary, speaker embedding editing fails to transform male voices to female ones, and when $\lambda \geq 3$ even originally female voices are not classified as such.

Additionally, 10 human evaluators classified speech samples as male or female. Analyzing sam-



(a) Latent space editing

(b) Speaker embedding editing

Figure 3: Supervised latent space editing allows gender manipulation, while speaker embedding editing does not. The percentage of samples classified as female is reported separately for male and female input speakers, averaged across 50 texts and all speakers (standard deviation, STD, is shaded).

Method		Gender acc. ↑		MOS ↑
Grad-TTS		0.82 ± 0.14		3.95 ± 0.15
Speaker Editing	***	0.76 ± 0.24	***	3.19 ± 0.17
Latent Editing		0.94 ± 0.07		3.59 ± 0.24

***** p-value < 0.001**

Table 1: Supervised latent space editing generates intelligible samples where the perceived speaker's gender is correctly classified, while speaker embedding editing does not. Average gender accuracy and MOS (mean ± STD) are reported. Latent-editing results compared to speaker-editing results are statistically significant (using Wilcoxon (1945) rank sum test).

ples from 20 different speakers, we compared the unedited Grad-TTS outputs to the gender-edited samples. For an effective gender alteration as shown in Figure 3, we used $\lambda = 2$ and -2 for male-to-female and female-to-male editing, respectively. Table 1 presents the accuracy of predicting the expected gender (original gender for original samples, and contrasting gender for edited samples). Comparing to speaker editing, latent space editing achieves a classification accuracy that is higher by 24%, with statistical significance (p-value < 0.001).

Acoustic properties evaluation. To assess the perceived naturalness of the generated speech we measure the Mean Opinion Score (MOS), as quantified by 10 experienced evaluators on a scale of 1 to 5, across the same set of samples reported before. Table 1 shows that the perceived naturalness of latent space editing, compared to speaker editing, is higher by 12%, a statistically significant difference (p-value < 0.001). This, combined with the superior perceived gender editing quality, reinforces the latent space's capability to encapsulate non-trivial semantic information.

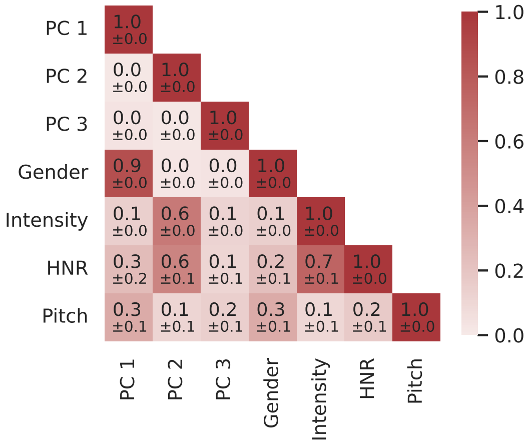


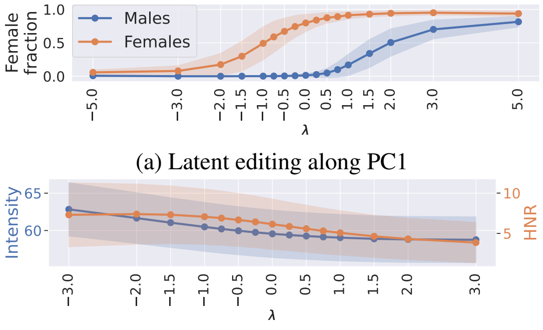
Figure 4: Absolute values of the Spearman correlation between the latent space PC-projections and the vocal attributes of the generated speech. We report mean and STD across all speakers, timesteps, and 50 texts.

3.3 Unsupervised Latent Space Editing

Next, we investigate semantically meaningful directions in *h-space* without prior annotations. First, we generated speech samples for the first 50 test texts of LibriTTS and across all 247 speakers, and recorded the latent vectors $\mathbf{h}_{T:1}$. Then, following the unsupervised process defined in Section 2.2, PCA of the latent space was performed for each text across all samples, calculating the first 3 principal components (PCs). As vocal attributes, for each speech sample we extracted its speaker's gender from the metadata, and measured its intensity, Harmonics-to-Noise Ratio (HNR), and pitch using the Parselmouth Python package (Jadoul et al., 2018).

The latent vectors of each sample were projected onto each PC. Next, we calculated the absolute value Spearman correlation between each vocal attribute and PC-projection vector, averaging across texts and timesteps. As Figure 4 shows, PC1 strongly correlates ($\rho = 0.9 \pm 0.0$) with speaker's gender (also see Figure 6 in Appendix A), while PC2 correlates ($\rho = 0.6 \pm 0.1$) with intensity and HNR. Other PCs and vocal attributes show no significant correlation and neither did random projections in the latent space (see Figure 8 in Appendix A).

Semantic properties evaluation. Using PCs as editing directions in *h-space*, we explore speech editing capabilities. Since the PCs are unitary vec-



(a) Latent editing along PC1

(b) Latent editing along PC2

Figure 5: Interpolation along the semantic directions revealed by PCA changes the vocal attributes accordingly. The reported values are averaged over 50 texts and all speakers. Shaded area is the STD.

Method		Gender acc. ↑	MOS ↑
Grad-TTS	*	0.82 ± 0.14	3.95 ± 0.15
PC1 Editing		0.88 ± 0.14	3.86 ± 0.20
PC2 Editing		0.82 ± 0.16	3.98 ± 0.17

*** p-value < 0.05**

Table 2: Gender accuracy and MOS results (mean ± STD) for unsupervised latent space editing.

tors, the editing directions were normalized to the norm of the latent vectors. Intriguingly, our experiments indicate that decreasing the editing norm at later timesteps improves acoustic quality. As can be seen in Figure 5a, interpolation along PC1 exhibits a smooth transition between male and female voices. Simialrly, intensity and HNR decrease when interpolating along PC2 (see Figure 5b). Importantly, no gender-editing occurs when interpolating along PC2 (see Figure 7 in Appendix A).

Additionally, we measured the accuracy of gender classification as evaluated by human annotators on the same 20 speakers. Following the analysis in Figure 5, to ensure effective gender alteration, we used $\lambda = 3$ or -3 for originally male or originally female speakers, respectively, while editing along PC1. For PC2, $\lambda = -2$ was used to maximize HNR. PC1-edited samples were successfully classified as the contrasting gender with an even higher accuracy than un-edited ones (Table 2).

Acoustic properties evaluation. Using the same setup, we assessed speech naturalness using MOS. Table 2 compares the perceived naturalness of samples with and without latent editing, presenting similar scores between the groups. The Wilcoxon rank sum test indicated no statistically significant difference in the MOS between groups (p-value $\gg 0.05$). Thus, we conclude that speech editing through unsupervised latent space manipulation does not compromise the acoustic quality.

4 Conclusions

In this paper, we identify the semantic properties of the latent space of diffusion-based TTS models, referred to as *h-space*. We develop supervised and unsupervised methods for finding interpretable directions in that space, and provide empirical qualitative evidence for their semantic quality. Moreover, the proposed latent space editing methods preserve and even enhance the acoustic quality of the generated samples. This study presents evidence regarding specific vocal attribute manipulation, such as gender or intensity. However, the presented method can be applied to any vocal attribute present in the data.

Limitations and Ethics

This study is subject to several limitations. We demonstrated our analysis on the Grad-TTS model (Popov et al., 2021) (trained on LibriTTS dataset (Zen et al., 2019)), and used the Universal HifiGAN (Kong et al., 2020a) for waveform generation. These are all publicly available for our research purposes. We do not develop novel TTS models from scratch, and focus on analysing existing ones. Under these settings, several limitations apply to our analysis:

1. LibriTTS is an English-only dataset, hence other languages are not supported by Grad-TTS, and were not analyzed.

2. LibriTTS is an audio-book reading dataset, and besides the speaker's gender no vocal attributes are provided. Therefore, we were limited to use the speaker's gender and the statistical audio attributes that we measure directly from the waveform. Properties such as emotion could not be analysed under these settings. We only refer to "male" or "female" voices to align with the original metadata.

3. Our method is general and can be applied to any frozen unguided diffusion-based TTS model that contains a bottleneck. However, since we were limited to publicly available models, we chose to focus on analysing the Grad-TTS model.

4. The acoustic quality of generated samples is bounded by the quality of the TTS system, including the Grad-TTS spectogram denoiser and the Universal HifiGAN vocoder quality.

5. The system cannot generate speech with a custom voice, as it does not take a voice-prompt as input. Thus, our edited audios are limited to the given subspace of speaker voices. This also points to the fact that our work does not pose risks regarding deep-fake or identity theft.

Acknowledgements

We thank Michael Hassid for the great feedback and moral support.

References

Sai Satya Vamsi Karthik Bhamidipati. 2023. multi-task-speech-classification. https://github.com/karthikbhamidipati/multi-task-speech-classification.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, Najim Dehak, and William Chan. 2021. Wavegrad 2: Iterative refinement for text-to-speech synthesis. *arXiv preprint arXiv:2106.09660*.

Yiwei Guo, Chenpeng Du, Xie Chen, and Kai Yu. 2023a. Emodiff: Intensity controllable emotional text-to-speech with soft-label guidance. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023b. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

René Haas, Inbar Huberman-Spiegelglas, Rotem Mulayoff, and Tomer Michaeli. 2023. Discovering interpretable directions in the semantic latent space of diffusion models.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022a. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*.

Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022b. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. 2021. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*.

Sungwon Kim, Heeseung Kim, and Sungroh Yoon. 2022. Guided-tts 2: A diffusion model for high-quality adaptive text-to-speech with untranscribed data. *arXiv preprint arXiv:2205.15370*.

Peter E Kloeden, Eckhard Platen, Peter E Kloeden, and Eckhard Platen. 1992. *Stochastic differential equations*. Springer.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020a. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020b. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. 2022. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*.

Songxiang Liu, Dan Su, and Dong Yu. 2022. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv preprint arXiv:2201.11972*.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8599–8608. PMLR.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Heiga Zen, Rob Clark, Ron J. Weiss, Viet Dang, Ye Jia, Yonghui Wu, Yu Zhang, and Zhifeng Chen. 2019. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech*.

A Additional Results

Further results supporting our main claims are presented in the following section.

An analysis of the PC1 and PC2 components of all the male and female speakers from LibriTTS is shown in Figure 6. It can be seen that PC1 provides an excellent separation between male and female voices. In contrast, PC2 does not provide such a separation.

Figure 7 presents the interpolation across PC2 for different λ values while monitoring the perceived speaker's gender. In line with expectations, interpolating across this editing direction does not affect the perceived speaker's gender, and it remains relatively unchanged. This is another indication of the disentanglement between the different editing directions found in the latent space by using our method.

A more detailed version of Figure 4 is presented in Figure 8, with random latent space projections and additional PC directions. As can be seen, only PC1 and PC2 exhibit significant correlations with the vocal attributes that were tested. Contrary to PCs, random projections do not correlate with any vocal attribute. This observation supports our claim that the latent space is capturing unique semantic properties.

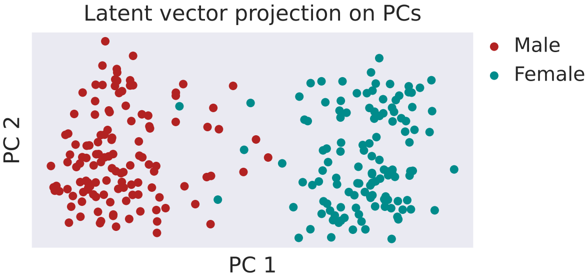


Figure 6: PC1 separates male from female speakers. Shown are the projection of latent spaces of samples generated with male and female speaker IDs onto PC1 and PC2.

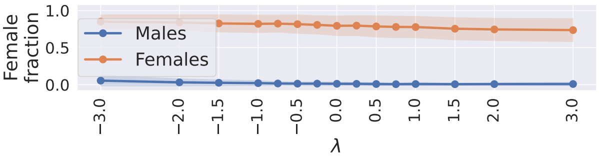


Figure 7: Interpolation along PC2 does not edit the perceived speaker's gender, indicating disentanglement of editing directions.

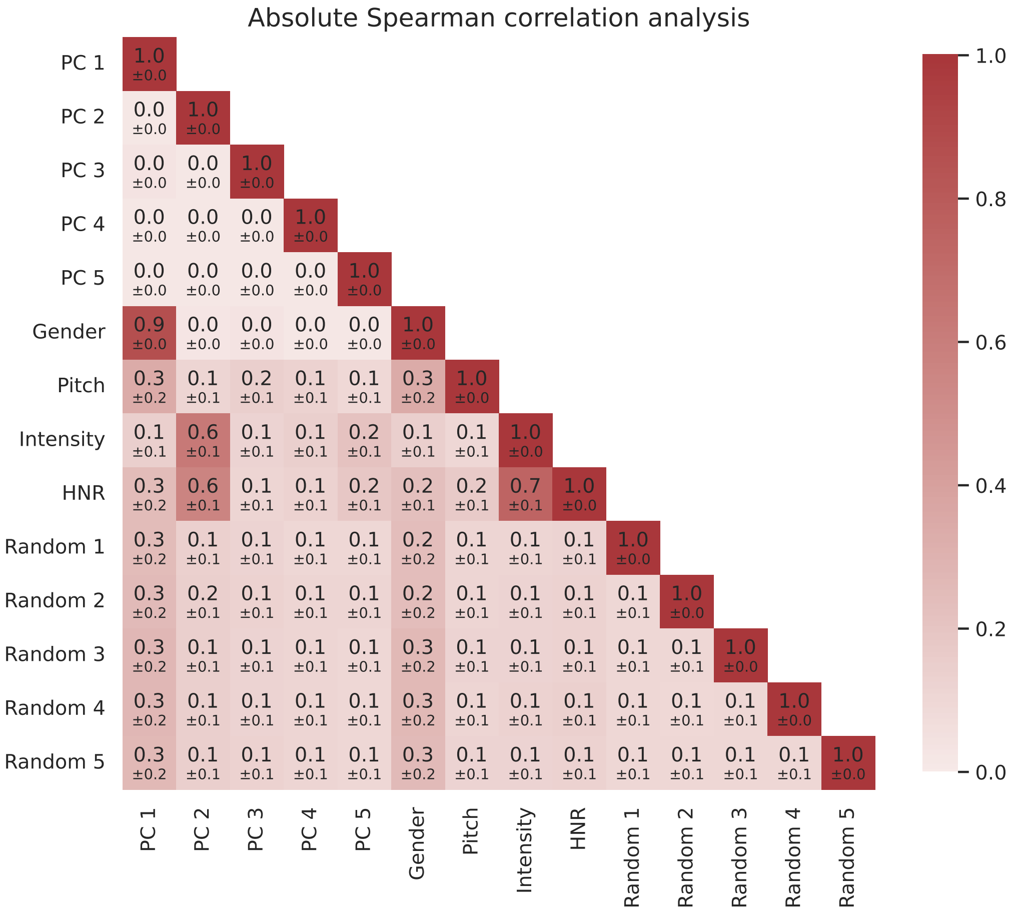


Figure 8: Principal components of latent space correlate with attributes of the generated audio. Shown are the mean and STD of the absolute value Spearman correlation of the PCs of the latent space, vocal attributes of the generated audios, and random projections of the latent space, averaged across all speakers, timesteps and 50 texts.

B Human Annotators

B.1 Human Evaluation of Perceived Speaker's Gender

To evaluate the perceived speaker's gender of generated samples, we used the Amazon Mechanical Turk (MTurk) crowd-sourcing platform. The MTurk workers we recruited and filtered had an approval rate above 50% and were located in the USA. The workers were instructed to classify the gender of each sample (binary classification). Each crowd worker was given the following instruction: "You are given an audio sample generated from a Text-To-Speech computer program. To the best of your ability, please classify the gender of the speaker in each audio sample. For better results, wear headphones and work in a quiet environment". We paid 0.02$ per Human Intelligence Task (HIT), and each worker was paid 4$ on average.

B.2 Mean Opinion Score Evaluation

To evaluate the quality of the generated speech, we utilized an internal annotation system. 34 experienced workers from the USA, who are native English speakers, have been assigned to assess the Mean Opinion Score (MOS) of the generated speech. Each worker was paid 0.34$ per-task (annotating a 3-second audio file) and each worker was paid an average of 51$ in total. The workers have been instructed to rate each speech sample quality based on the acceptable 5-point MOS score, Table 3 provides details regarding the scoring methodology used.

Score	Quality
5.0	Excellent (Completely defined)
4.5	
4.0	Good (Mostly defined)
3.5	
3.0	Fair (Equally defined and undefined)
2.5	
2.0	Poor (Mostly undefined)
1.5	
1.0	Bad (Completely undefined)

Table 3: Mean Opinion Score (MOS) scoring schema.