# PROSODY-TTS: SELF-SUPERVISED PROSODY PRE-TRAINING WITH LATENT DIFFUSION FOR TEXT-TO-SPEECH

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Expressive text-to-speech aims to generate high-quality samples with rich and diverse prosody, which is hampered by two major challenges: 1) considering the one-to-many mapping problem, prosodic attributes in highly dynamic voices are difficult to capture and model without intonation; 2) the TTS model should learn a diverse latent space and prevent producing dull samples with a collapsed prosodic distribution. This paper proposes Prosody-TTS, a two-stage TTS pipeline that improves prosody modeling and sampling by introducing several components: 1) a self-supervised learning model to derive the prosodic representation without relying on text transcriptions or local prosody attributes, which ensures the model covers diverse speaking voices, preventing sub-optimal solutions and distribution collapse; and 2) a latent diffusion model to sample and produce diverse patterns within the learned prosodic space, which prevents TTS models from generating the dull samples with mean distribution. Prosody-TTS achieves high-fidelity speech synthesis with rich and diverse prosodies. Experiments results demonstrate that it surpasses the state-of-the-art models in terms of audio quality and prosody naturalness. The downstream evaluation and ablation studies further demonstrate the effectiveness of each design.[1]

## 1 INTRODUCTION

Text-to-speech (TTS) (Wang et al., 2017; Ren et al., 2019; Kim et al., 2020; Popov et al., 2021) aims to generate human-like audios using text and auxiliary conditions, which attracts broad interest in the machine learning community. TTS models have been extended to more complex scenarios, requiring more natural and expressive voice generation with improved prosody modeling (Min et al., 2021; Chen et al., 2021; Li et al., 2021). A growing number of applications, such as personalized voice assistants and game commentary, have been actively developed and deployed to real-world applications.

Expressive text-to-speech aims to generate samples with natural, rich, and diverse prosodic attributes, which is challenged by two major obstacles: 1) Emotions or styles (prosody patterns) (Qian et al., 2021; Wang et al., 2018) in human speech are often very sparse, which are difficult to capture and model without supervision signals (i.e., detailed rich transcriptions) in natural voices; 2) machine learning models (Li et al., 2018; Wang et al., 2022) usually learn a mean distribution over input data, resulting a dull prediction with prosody learners which fails to produce desired prosodic styles in the generated speech. Although recent studies (Choi et al., 2021; Kim et al., 2021; Ren et al., 2022) have proposed several ways to enhance prosody modeling for high-fidelity TTS, there still exist some challenges and issues in their methods:

- **Prosody capturing and modeling.** Researchers leverage several designs to capture and model prosodic attributes: 1) Local prosody features. Popular works such as (Ren et al., 2020; Choi et al., 2021) introduce the idea of predicting pitch and energy explicitly, however, those signal processing-based prosodic attributes have inevitable errors, which make the optimization of TTS models difficult, resulting in degraded TTS performance. 2) Variational latent representations.

[1] Audio samples are available at `https://Prosody-TTS.github.io/`.

A series of works (Sun et al., 2020; Kenter et al., 2019; Liu et al., 2022) utilize conditional variational auto-encoder to model prosody in a latent space, where global, local, or hierarchical features are sampled from a prior distribution. Nevertheless, they generally request paired data with transcriptions, which constrained the learned representation to the paired TTS data.

- **Prosody producing and sampling.** Most works (Wang et al., 2017; Min et al., 2021; Yang et al., 2021a) utilize L1 or L2 losses for reconstruction and assume that representations follow a unimodal distribution. However, the highly multimodal prosodic distribution cannot be well modeled by these simple objectives, which causes blurry and over-smoothing predictions in latent space.

To address the above challenges for expressive text-to-speech, we propose Prosody-TTS, a two-stage TTS pipeline that improves prosody modeling and sampling by introducing several novel designs:

- **Self-supervised prosody pre-training.** To handle different acoustic conditions for expressive speech, we propose prosody masked autoencoders (Prosody-MAE), a transformer-based model that captures prosody patterns (e.g., local rises and falls of the pitch and stress). It is trained in a self-supervised manner with only the audio modality, which ensures the model covers diverse speech corpora and explicit better generalization.
- **Generative diffusion modeling in latent space** A latent diffusion model is explored to bridge TTS inputs (i.e., textual input and target speaker) and the prosody representation. Specifically, we formulate the generative process (i.e., Prosody-TTS) with multiple conditional diffusion steps. Therefore, we expect our model to exhibit better diversity and prevent generating dull samples with a mean prosodic distribution.

Experimental results on LJSpeech and LibriTTS benchmarks demonstrate that our proposed Prosody-TTS generates high-fidelity speech with rich and diverse prosodic attributes. Both subjective and objective evaluation metrics demonstrate that Prosody-TTS surpasses the state-of-the-art models in terms of audio quality and naturalness. The downstream evaluations and ablation studies further justify the effectiveness of each module that we propose.

## 2 Related Works

### 2.1 Prosody Modeling in Text-To-Speech

Prosody modeling has been studied for decades in the TTS community. The idea of pitch and energy prediction (Łańcucki, 2021; Ren et al., 2020) represents a popular way to address the one-to-many mapping challenges. Wang et al. (2019) utilize the VQ-VAE framework to learn a latent representation for the F0 contour of each linguistic unit and adopt a second-stage model which maps from linguistic features to the latent features. Choi et al. (2021) further use a new set of analysis features, i.e., the wav2vec and Yingram feature for self-supervised training. However, these signal processing-based prosodic attributes have inevitable errors, which make the optimization of TTS models difficult and result in degraded TTS performance. Instead of relying on local prosody attributes, VITS (Kim et al., 2021) and its derivative (Casanova et al., 2022) utilize a posterior encoder to capture the prosody features and sample them from an enhanced conditional prior distribution. As these methods request a paired corpus with transcriptions during training, they constrain the learned representation to the paired TTS data and explicit poor generalization (Wang et al., 2022). ProsoSpeech (Ren et al., 2022) introduces a prosody encoder to disentangle the prosody to latent vectors, while the requirement of a pre-trained TTS model hurts model generalization. In this work, we propose to effectively learn the prosodic distribution given speech samples without relying on pre-trained TTS models or text transcriptions.

### 2.2 Self-Supervised Learning in Speech

Recently, self-supervised learning (SSL) has emerged as a popular solution to many speech processing problems with a massive amount of unlabeled speech data. HuBERT (Hsu et al., 2021) is trained with a masked prediction with masked continuous audio signals. SS-AST (Gong et al., 2022) is a self-supervised learning method that operates over spectrogram patches. Baade et al. (2022) propose a simple yet powerful improvement over the recent audio spectrogram transformer (SSAST) model. Audio-MAE (Xu et al., 2022) is a simple extension of image-based Masked Autoencoders (MAE) (He

et al., 2022) for SSL from audio spectrograms. While the majority of the SSL models in speech aim to capture linguistic content and learn prosody-agnostic representation, we focus on learning prosodic representation in expressive speech samples in contrast, which is relatively overlooked.

### 2.3 DIFFUSION PROBABILISTIC MODEL

Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020; Song et al., 2020a) are likelihood-based generative models that have recently succeeded in advancing the SOTA results in several important domains, including image (Dhariwal & Nichol, 2021; Song et al., 2020a), audio (Huang et al., 2022b; Liu et al., 2021), and 3D point cloud generation (Luo & Hu, 2021). In this work, we investigate generative modeling for latent representations with a conditional diffusion model. The latent diffusion model generates realistic results that match the ground-truth distribution and avoid over-smoothing predictions.

## 3 PROSODY-TTS

In this section, we first overview the Prosody-TTS framework, following which we introduce several critical designs including prosody masked autoencoder (Prosody-MAE), latent diffusion model, and the vector quantization layer. Finally, we present the pre-training, training, and inference pipeline, which supports high-fidelity speech synthesis with natural, rich, and diverse prosodic attributes.

### 3.1 OVERVIEW

We adopt one of the most popular non-autoregressive text-to-speech models FastSpeech 2 (Ren et al., 2020) as the model backbone. As illustrated in Figure 1(b), to address the aforementioned challenges of modeling prosody from expressive voices, we introduce a multi-stage pipeline with several novel designs: 1) In the pre-training stage, the Prosody-MAE captures prosodic information from large-scale unpaired speech data without relying on transcriptions or local prosody attributes. The self-supervised training manner ensures Prosody-MAE covers diverse speaking styles; 2) In training Prosody-TTS, the converged prosody encoder derives style representations for optimizing the latent diffusion model (LDM), which bridges the TTS conditions (i.e., textual features and target speaker) and prosody-MAE representations via forward/backward diffusion/denoising process; 3) In inference time, the LDM samples diverse latent representations within the prosodic space through reverse denoising, which is conditioned on textual information. It breaks the generation process into several conditional diffusion steps, thus avoiding generating dull samples with the mean prosodic distribution. We describe these designs in detail in the following subsections.

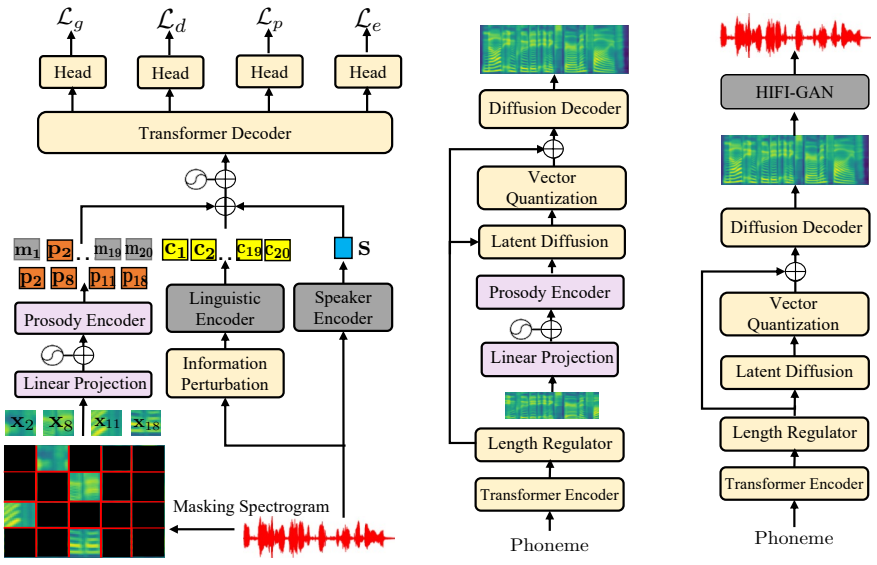
### 3.2 SELF-SUPERVISED PROSODY PRE-TRAINING

In this part, we propose Prosody-MAE, a self-supervised learning model that can effectively capture and model prosodic style given speech samples without relying on text annotations. Autoencoders (AE) (Kingma & Welling, 2013) which consist of an encoder and decoder have played an essential role in learning distributed latent representations of sensory data. We select the autoencoder as the backbone and design several techniques to learn prosodic representation in a self-supervised manner:

- Information flow. The Prosody-MAE enjoys a carefully-crafted information bottleneck design. By introducing pre-trained speech encoders to restrict the information flow, the model could capture the prosodic representation more efficiently by residual branch;
- Masking strategy. Instead of considering the time-align masking operation, the hierarchal masking learns both the temporal and frequency structure. Thus, we expect the model to exhibit better abilities in learning style attributes;
- Multi-task learning. The style (i.e., pitch and energy) classifications have been investigated as the auxiliary objectives, which guarantees the model to discover discriminative prosodic representation.

#### 3.2.1 INFORMATION FLOW

In this section, we decompose the speech into linguistic content, speaker, and prosody variations and provide a brief primer on the carefully-crafted information flow.




(a) Prosody-MAE Pretraining (b) Prosody-TTS Training (c) Prosody-TTS Inference

Figure 1: Pre-training, training and inference stage of Prosody-TTS. We use the sinusoidal-like symbol to denote the positional encoding, and the publicly-available pre-trained blocks are printed in black color.

**Linguistic Encoder.** Learning the linguistic content $\mathcal{C}$ from the speech signal is crucial to construct an intelligible speech signal. We propose to obtain linguistic representation by 1) utilizing the commonly-used 12th encoder layer in pre-trained XLSR-53 (Conneau et al., 2020), which is pre-trained on 56k hours of speech in 53 languages, to provide linguistic information; 2) adopting information perturbation to disentangle acoustic information from noisy input. Since SSL features (Choi et al., 2021) contain both linguistic and acoustic information, we perturb the speaker and prosody patterns in audios by randomly shifting pitch and shaping energy values, ensuring it only provides the linguistic-related (i.e., prosodic-agnostic) information. Detailed information on the perturbation functions has been included in Appendix D.

**Speaker Encoder.** Speaker $\mathcal{S}$ is perceived as the timbre characteristic of a voice. It has been reported that (Choi et al., 2021) the features from the first layer of XLSR-53 perform as clusters representation for each speaker.

**Prosody Encoder.** Prosody is an important part of the domain style, where different emotions or styles have distinctive prosody patterns. In the multi-layer transformer prosody encoder, 1) speech is first transformed and embedded into spectrogram patches, and 2) we add positional embeddings to these features. 3) The encoders $f : \mathcal{X} \mapsto \mathcal{P}$ takes patches $\mathcal{X}$ as input and effectively capture prosodic latent representations $\mathbf{p}_1, \ldots, \mathbf{p}_T$ for $T$ time-steps.

**Transformer Decoder.** The decoder has a series of transformer blocks, which is used only during pre-training and discarded in the downstream text-to-speech. As illustrated in Figure 1(a), we constrain the information flow by conducting the element-wise addition operation between the linguistic content $\mathcal{C}$, speaker $\mathcal{S}$ and the prosody $\mathcal{P}$ representations before passing through the transformer decoder. Furthermore, we include the positional embeddings to all tokens in this full set.

### 3.2.2 Masking Strategy

Masked reconstruction is largely inspired by the masked language model (MLM) task from BERT (Devlin et al., 2018) and MAE (He et al., 2022). During pre-training, some tokens in the input sentences are masked by randomly replacing them with a learned masking token ($\mathbf{m}_i$ illustrated in Figure 1(a)). In practice, we mask by shuffling the input patches and keeping the first $1 - p$ proportion of tokens.

After padding encoded patches with learnable embeddings to represent masked patches, it restores the order of these patches in frequency and time and propagates through a decoder to reconstruct the spectrogram. In summary, the hierarchal masking operation in the spectrogram helps learn both the temporal and frequency structure, and we expect the model to exhibit better learning style attributes.

#### 3.2.3 MULTI-TASK LEARNING

For training Prosody-MAE, we adopt discriminative and generative objectives with four linear layer heads for our final output projection. Reconstruction loss $\mathcal{L}_g$ is calculated as a mean squared error between the output of the linear reconstruction head and the input patches. Contrastive head (Gong et al., 2022) intends to create an output vector $\mathbf{v}_i$ similar to the masked input patch $\mathbf{x}_i$ but dissimilar to other masked inputs. Therefore we consider different masked inputs as negative samples and implement the InfoNCE (Oord et al., 2018) as a criterion.

$$\mathcal{L}_d = -\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(\mathbf{v}_i^T\mathbf{x}_i\right)}{\sum_{j=1}^{N}\exp\left(\mathbf{v}_i^T\mathbf{x}_j\right)}\right) \tag{1}$$

To further enhance prosody learning in latent space, we investigate the frame-level style (i.e., pitch $\mathcal{L}_p$, energy $\mathcal{L}_e$) classification as the auxiliary tasks and employ the cross-entropy (Oord et al., 2018) as a criterion, which guarantees the model to discover style representation. To formulate the classification target, we respectively 1) quantize the fundamental frequency (f0) of each frame to 256 possible values $\mathbf{p}_i$ in log-scale; and 2) compute the L2-norm of the amplitude of each short-time Fourier transform (STFT) and then quantize to 256 possible values $\mathbf{e}_i$ uniformly.

### 3.3 GENERATIVE MODELING OF PROSODIC REPRESENTATIONS

To sample and interpolate latent representation within the learned prosodic space, we implement our method over Latent Diffusion Models (LDMs) (Rombach et al., 2022; Gal et al., 2022), which is a recently introduced class of Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) that operate in the latent representation space. As illustrated in Figure 1(c), the proposed latent diffusion model conditions on linguistic representation, breaking the generation process into several conditional diffusion steps. The training loss is defined as the mean squared error in the noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ space, and efficient training is optimizing a random term of $t$ with stochastic gradient descent:

$$\mathcal{L}_\theta = \left\|\boldsymbol{\epsilon}_\theta\left(\alpha_t\mathbf{x}_0 + \sqrt{1-\alpha_t^2}\boldsymbol{\epsilon}\right) - \boldsymbol{\epsilon}\right\|_2^2, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{2}$$

To conclude, our approach ensures extremely faithful reconstructions and requires little regularization of the latent space. The latent diffusion model can be efficiently trained by optimizing ELBO without adversarial feedback, which generates realistic prosodic representation strongly matching the ground-truth distribution and thus prevents collapsing with mean prosodic distribution.

### 3.4 VECTOR QUANTIZATION

It has been reported (Rombach et al., 2022) that due to the expressiveness of diffusion models, the produced latent spaces could be highly variant and diverse. To avoid instability, we impose a vector quantization (VQ) layer after the latent diffusion for regularization. Specifically, it receives prosodic representations from different origins in training or inference stages: 1) When training the TTS model, the VQ layer receives clean features derived from the prosody encoder in Prosody-MAE, instead of noisy diffusion output. 2) During inference, the VQ layer receives prosodic representation via LDM's backward latent denoising, which bridges the TTS conditions and prosody features.

Denote the latent space $e \in R^{K\times D}$ where $K$ is the size of the discrete latent space (i.e., a $K$-way categorical), and $D$ is the dimensionality of each latent embedding vector $e_i$. Note that there are $K$ embedding vectors $e_i \in \mathbb{R}^D, i \in 1, 2, \ldots, K$. To make sure the representation sequence commits to an embedding and its output does not grow, we add a commitment loss following previous work (van den Oord et al., 2017):

$$\mathcal{L}_c = \|q_p(x) - \mathrm{sg}[e]\|_2^2, \tag{3}$$

Where $q_p(x)$ is the output of the vector quantization block, and sg stands for stop gradient.

### 3.5 PRE-TRAINING, TRAINING AND INFERENCE PROCEDURES

#### 3.5.1 PRE-TRAINING AND TRAINING

In the pre-training stage, we train the Prosody-MAE to learn the latent prosodic representation in a self-supervised manner using the following loss objective: 1) reconstruction loss $\mathcal{L}_g$: the mean squared

error between the estimated and ground-truth sample; 2) contrastive loss $\mathcal{L}_d$: the discriminative gradient to pick the correct patch for each masked position from all patches being masked, and 3) frame-level style (i.e., pitch, energy) classification loss $\mathcal{L}_p, \mathcal{L}_e$: the cross entropy error between the estimated and ground-truth values.

In training Prosody-TTS, the final loss terms consist of the following parts: 1) duration prediction loss $\mathcal{L}_{dur}$: MSE between the predicted and the GT phoneme-level duration in log scale; 2) latent diffusion loss $\mathcal{L}_{ldm}$ and decoder diffusion loss $\mathcal{L}_{dec}$: two diffusion losses between the estimated and gaussian noise according to Equation 2; 3) commitment loss $\mathcal{L}_c$: the objective to constrain vector quantization layer according to Equation 3.

#### 3.5.2 INFERENCE

As illustrated in Figure 1, Prosody-TTS generates expressive speech with natural, rich, and diverse prosody in the following pipeline: 1) The text encoder encodes the phoneme sequence, and the representations could be expanded according to the inference duration; 2) conditioning on linguistic information, the latent diffusion model randomly samples a noise latent and iteratively denoises to produce a new prosodic representation in latent space, and 3) the mel decoder converts randomly samples noise latent and iteratively decodes to expressive mel-spectrogram predictions.

## 4 EXPERIMENTS

### 4.1 EXPERIMENTAL SETUP

#### 4.1.1 PRE-TRAINING PROSODY-MAE

In the pre-training stage, we utilize the commonly-used LibriSpeech (Panayotov et al., 2015) dataset with labels discarded, which provides 960 hours of audiobook data in English, read by over 1,000 speakers. We convert the 16kHz waveforms into 128-dimensional log-Mel filterbank features with a frame length of 25 ms and frame shift of 10 ms. The spectrogram is then split into $16\times16$ patches.

By default, we use an encoder with 6 layers and a decoder of 2 layers, both using 12 heads and a width of 768. We train Prosody-MAE for up to 400k iterations on 8 NVIDIA V100 GPUs using the publicly-available *fairseq* framework (Ott et al., 2019), and the pre-training takes about 5 days. For downstream evaluation, we use the standard SUPERB (Yang et al., 2021b) training and testing framework. More detailed information has been attached in Appendix B.

#### 4.1.2 TRAINING PROSODY-TTS

**Dataset.** For a fair and reproducible comparison against other competing methods, we use the benchmark LJSpeech dataset (Ito, 2017), which consists of 13,100 audio clips from a female speaker for about 24 hours in total. For the multi-speaker scenario, we utilize the 300-hour LibriTTS dataset derived from LibriSpeech. We convert the text sequence into the phoneme sequence with an open-source grapheme-to-phoneme conversion tool (Sun et al., 2019) [2].

Following the common practice (Chen et al., 2021; Min et al., 2021), we conduct preprocessing on the speech and text data: 1) convert the sampling rate of all speech data to 16kHz; 2) extract the spectrogram with the FFT size of 1024, hop size of 256, and window size of 1024 samples; 3) convert it to a mel-spectrogram with 80 frequency bins.

**Model Configurations.** Prosody-TTS consists of 4 feed-forward transformer blocks for the phoneme encoder. We add a linear layer to transform the 768-dimension prosody latent representation from Prosody-MAE to 256 dimensions. The default size of the codebook in the vector quantization layer is set to 1000. The diffusion model comprises a 1x1 convolution layer and $N$ convolution blocks with residual connections to project the input hidden sequence with 256 channels. For any step $t$, we use the cosine schedule $\beta_t = \cos(0.5\pi t)$. More detailed information has been attached in Appendix A.

**Training and Evaluation.** We train Prosody-TTS for 200,000 steps using 4 NVIDIA V100 GPUs with a batch size of 64 sentences. Adam optimizer is used with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$. We

[2] https://github.com/Kyubyong/g2p

Table 1: **Performance (audio quality and prosody naturalness) comparison with other models.** We report the evaluation metrics including MOS(↑), FFE(↓), and MCD(↓). The mel-spectrograms are converted to waveforms using Hifi-GAN (V1).

| Method | LJSpeech | | | | LibriTTS | | | |
|---|---|---|---|---|---|---|---|---|
| | MOS-P | MOS-Q | MCD | FFE | MOS-P | MOS-Q | MCD | FFE |
| GT | 4.36±0.05 | 4.39±0.06 | / | / | 4.38±0.05 | 4.42±0.06 | / | / |
| GT(voc.) | 4.31±0.06 | 4.25±0.06 | 1.67 | 0.08 | 4.35±0.04 | 4.22±0.05 | 1.52 | 0.06 |
| FastSpeech 2 | 3.92±0.07 | 3.84±0.06 | 3.88 | 0.43 | 3.89±0.06 | 3.81±0.07 | 4.35 | 0.37 |
| Meta-StyleSpeech | 3.94±0.06 | 3.88±0.05 | 5.54 | 0.46 | 3.95±0.07 | 3.91±0.08 | 3.78 | 0.33 |
| Glow-TTS | 3.88±0.06 | 3.91±0.06 | 3.54 | 0.48 | 3.91±0.08 | 3.86±0.08 | 5.38 | 0.38 |
| Grad-TTS | 3.91±0.07 | 3.92±0.06 | 5.01 | 0.44 | 3.96±0.06 | 3.97±0.05 | 3.93 | 0.37 |
| YourTTS | 3.97±0.06 | 3.96±0.06 | 5.09 | 0.48 | 3.99±0.07 | 3.99±0.06 | 4.61 | 0.35 |
| Prosody-TTS | **4.10±0.06** | **4.03±0.05** | **3.52** | **0.35** | **4.12±0.07** | **4.09±0.06** | **3.39** | **0.29** |

utilize HiFi-GAN (Kong et al., 2020) as the vocoder to synthesize waveform from the mel-spectrogram in our experiments. To evaluate the perceptual quality, we conduct crowd-sourced human evaluations on the testing set via Amazon Mechanical Turk, which is reported with 95% confidence intervals (CI). We analyze the MOS/CMOS in two aspects: prosody (naturalness of pitch, energy, and duration) and audio quality (clarity, high-frequency and original timbre reconstruction), respectively scoring MOS-P/CMOS-P and MOS-Q/CMOS-Q. For objective evaluation, we include MCD and FFE to measure the audio and prosody quality. More details have been attached in Appendix E.

### 4.2 Comparision with other models

We compare the quality of generated audio samples with other systems, including 1) GT, the ground-truth audio; 2) GT (voc.), we first convert the ground-truth audio into mel-spectrograms and then convert them back to audio using HiFi-GAN (V1) (Kong et al., 2020); 3) FastSpeech 2 (Ren et al., 2020): a model that predicts local prosody attributes; 4) Meta-StyleSpeech (Kim et al., 2020): the finetuned multi-speaker model with meta-learning; 5) Glow-TTS (Kim et al., 2020): a flow-based TTS model trained with monotonic alignment search; 6) Grad-TTS (Popov et al., 2021): a denoising diffusion probabilistic models for speech synthesis. 7) YourTTS (Casanova et al., 2022): an expressive model for zero-shot multi-speaker synthesis. The results are compiled and presented in Table 1, and we have the following observations:

1) In terms of audio quality, Prosody-TTS achieves the highest perceptual quality with MOS-Q of 4.03 (LJSpeech) and 4.09 (LibriTTS). 2) For prosody naturalness, Prosody-TTS scores the highest overall MOS-P with a gap of 0.21 (LJSpeech) and 0.23 (LibriTTS) compared to the ground truth audio. For objective evaluation, Prosody-TTS also demonstrates the outperformed performance in MCD and FFE, superior to all baseline models. Without relying on text transcriptions or local prosody ground truth, our model is demonstrated to cover diverse speaking styles and avoid sub-optimal predictions. For prosody diversity, we have attached objective evaluation in Appendix F. We include an additional AXY evaluation in Appendix G and discuss computational cost in Appendix H

We further plot the mel-spectrograms and corresponding pitch tracks generated by the TTS systems in Figure 2, and have the following observations: 1) Prosody-TTS can generate mel-spectrograms with rich details in frequency bins between two adjacent harmonics, unvoiced frames, and high-frequency parts, which results in more natural sounds. However, some baseline models fail to synthesize high-fidelity mel-spectrograms; 2) Prosody-TTS demonstrates its ability to generate samples with diverse prosodic styles. Informally, by breaking the generation process into several conditional diffusion steps, generative diffusion models prevent TTS from learning the collapsed prosodic distribution. In contrast, most baseline models learn a mean distribution over their input data, leading to less expressive synthesis with dull samples, especially for long-form phrases.

### 4.3 Downstream Evaluation on Model Properties

To demonstrate several critical designs in the SSL models, we introduce style-aware downstream challenges including the frame-level pitch and energy recognition. In the fine-tuning phase, we remove the decoder and only fine-tune the encoder on the commonly-used dataset IEMOCAP (Busso et al., 2008) that contains about 12 hours of emotional speech. We use a fixed learning rate of 1e-4

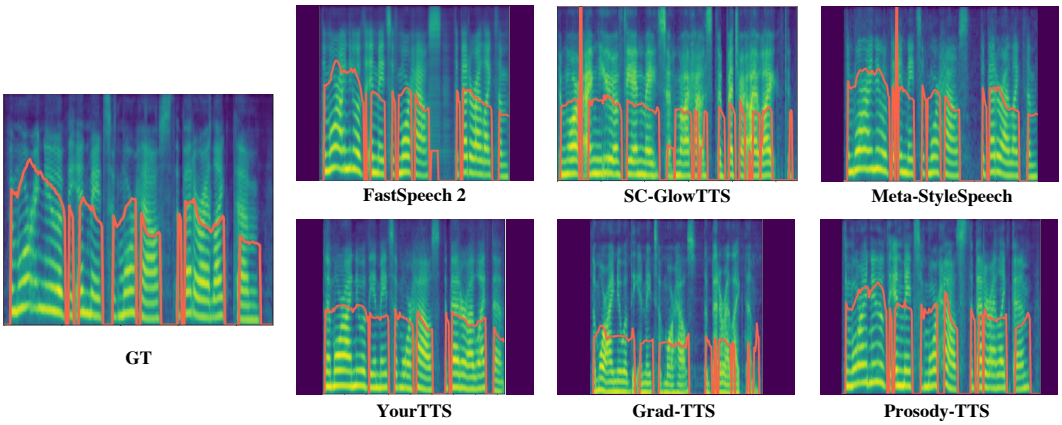



Figure 2: Visualizations of the generated mel-spectrograms. The corresponding text of generated speech samples is "*there was not a worse vagabond in Shrewsbury than old Barney the piper.*".

and max iteration of 10k and fine-tune on 4 V100 GPUs for 60 epochs using the SUPERB (Yang et al., 2021b) framework. More detailed information on downstream fine-tuning is available in Appendix B. The results are compiled and presented in Table 2, and we have the following observations:

**Pretext task.** We investigate the impact of different pretext tasks for pre-training the SSL model, and find that the additional contrastive objective leads to better performance for all tasks. Furthermore, the joint multi-task learning with frame-level style classification has witnessed a distinct promotion of downstream accuracy, demonstrating the efficiency of the auxiliary tasks in enhancing prosody modeling.

**Information flow.** We conduct ablation studies to demonstrate the effectiveness of the carefully-crafted information flow in learning prosodic style attributes: 1) Dropping the linguistic and speaker encoder has witnessed a distinct degradation of downstream performance, proving that through restricting the information flow of speech variations, the Prosody-MAE encoder could disentangle prosody variations; and 2) utilizing the raw wav2vec feature as a linguistic representation by removing the information perturbation also decreases accuracy, demonstrating that the perturbation assists to selectively provide only the linguistic (i.e., prosodic-agnostic) information.

**Network architecture.** Similar to the MAE paper demonstrated for the visual domain, increasing the decoder depth only provides minor improvements if any, indicating that the decoder depth can be small relative to the encoder.

**Masking strategies.** We compare different masking ratios for pre-training Prosody-MAE, and observe that a high masking ratio (70% in our case) is optimal for audio spectrograms. Due to the fact that audio spectrograms and images are continuous signals with significant redundancy, and thus SSL models still could reconstruct results given most tokens dropped, which is consistent with the masked autoencoders (He et al., 2022) in the visual domain.

**Comparision with other state-of-the-art.** We compare our proposed Prosody-MAE with prior state-of-the-art SSL models, including: 1) wav2vec 2.0 (Baevski et al., 2020), 2) hubert (Hsu et al., 2021), 3) robust hubert (Huang et al., 2022a), and 4) mae-ast (Baade et al., 2022) and find that our proposed Prosody-MAE achieves the best performance across all tasks compared to other systems. Specifically, the majority of the speech SSL models focus on learning the linguistic content information, which try to disentangle unwanted variations (e.g. acoustic variations) from the content. In contrast, we hope to capture prosodic information from speech, and thus Prosody-MAE exhibits outperformed capability in capturing style attributes.

### 4.4 Ablation Studies

We conduct ablation studies to verify the effectiveness of several designs in Prosody-TTS, including the latent diffusion model and vector quantization layer. The CMOS evaluation results have been presented in Table 3, and we have the following observations: 1) Replacing the latent diffusion model with the regression-based style predictor results in decreased prosody naturalness and expressiveness, demonstrating that generative diffusion models avoid producing blurry and over-smoothing results.

Table 2: **Ablations and model properties**. We report the evaluation metrics including accuracy (PA↑), mean absolute error (PM↓) in pitch recognition, and mean absolute error (EM↓) in energy recognition to evaluate model properties. In table (b), we use IF and IP to denote the carefully-crafted information flow design and the perturbation.

(a) **Pretext Task**

| Objective | PA | PM | EM |
|---|---|---|---|
| Generative | 70.0 | 8.35 | 4.63 |
| + Contrastive | 73.1 | 7.50 | 3.13 |
| + Contrastive + Auxiliary | **75.2** | **7.22** | **1.76** |

(b) **Information Flow**

| IF | IP | PA | PM | EM |
|---|---|---|---|---|
| ✗ | ✗ | 73.0 | 7.50 | 3.13 |
| ✓ | ✗ | 74.9 | 7.76 | 6.26 |
| ✓ | ✓ | **75.2** | **7.22** | **1.76** |

(c) **Network Architecture**

| Layers | PA | PM | EM |
|---|---|---|---|
| 2 | 75.2 | **7.22** | **1.76** |
| 4 | 75.3 | 7.41 | 2.01 |
| 6 | **75.5** | 7.73 | 2.25 |
| 8 | 74.6 | 7.85 | 2.52 |

(d) **Masking Strategies**

| Mask Ratio | PA | PM | EM |
|---|---|---|---|
| 80% | **75.2** | 7.22 | 1.76 |
| 70% | **75.2** | 7.11 | **1.65** |
| 60% | 74.9 | **7.05** | 2.11 |
| 50% | 74.6 | 7.34 | 2.82 |

(e) **Comparision with other state-of-the-art**

| Model | PA | PM | EM |
|---|---|---|---|
| wav2vec 2.0 | 70.7 | 7.34 | 3.21 |
| HuBERT | 69.9 | 8.00 | 5.63 |
| Robust HuBERT | 69.5 | 7.95 | 5.37 |
| MAE-AST | 73.1 | 8.17 | 5.43 |
| Prosody-MAE | **75.2** | **7.22** | **1.76** |

Table 3: **Performance (audio quality and prosody naturalness) comparison for ablation study.**

| Model | CMOS-P | CMOS-Q |
|---|---|---|
| Prosody-MAE | 0.00 | 0.00 |
| w/o LDM | -0.11 | -0.04 |
| w/o VQ | -0.04 | -0.08 |
| Local Prosody | -0.12 | -0.02 |
| Variational Inference | -0.10 | -0.03 |

2) Removing the vector quantization layer has witnessed a distinct degradation of audio quality, verifying the significance of VQ compression layer in constraining latent spaces and preventing arbitrarily high-variance out-of-distribution predictions.

We further compare different prosody modeling techniques in TTS models to check which demonstrates better performance. The side-by-side subjective test indicates that raters prefer speech synthesis with prosodic features derived from Prosody-MAE, for the reason that: 1) Local prosody ground truth with inevitable errors cannot be accurately estimated, and 2) variational autoencoder requests transcripted corpus for conditions, which constrains the distribution of learned representation to TTS corpora. To conclude, Prosody-MAE significantly improves TTS model by effectively capturing the prosodic style and producing latent in a diverse space, enabling high-fidelity speech synthesis with natural and prosperous prosody.

## 5 Conclusion

In this work, we propose Prosody-TTS, an expressive text-to-speech model with self-supervised prosody pre-training. To enhance high-quality synthesis with prosperous and diverse prosody, we design a two-stage pipeline to model and control the prosody variations in speech: 1) Prosody-MAE was pre-trained on large-scale unpaired speech datasets to capture prosodic information without relying on text transcriptions or local ground truth. This ensured the model covered diverse speaking voices and prevented sub-optimal and distribution collapse. 2) The latent diffusion model was further adopted to produce diverse patterns within the learned prosody space. It broke the generation process into several conditional diffusion steps, which prevented TTS models from generating the dull sample with mean prosodic distribution. Experimental results demonstrated that Prosody-TTS promoted prosody modeling and synthesized high-fidelity speech samples, achieving new state-of-the-art results with outperformed audio quality and prosody expressiveness. For future work, we will further verify the effectiveness of Prosody-TTS in more general scenarios such as multilingual prosody learning. We envisage that our work could serve as a basis for future text-to-speech synthesis studies.

## REFERENCES


Alan Baade, Puyuan Peng, and David Harwath. Mae-ast: Masked autoencoding audio spectrogram transformer. *arXiv preprint arXiv:2203.16691*, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pp. 2709–2720. PMLR, 2022.

Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu. Adaspeech: Adaptive text to speech for custom voice. *arXiv preprint arXiv:2103.00993*, 2021.

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265, 2021.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Yuan Gong, Cheng-I Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10699–10709, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. 2020.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Kuan Po Huang, Yu-Kuan Fu, Yu Zhang, and Hung-yi Lee. Improving distortion robustness of self-supervised speech processing tasks with domain adaptation. *arXiv preprint arXiv:2203.16104*, 2022a.

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

Keith Ito. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`, 2017.

Tom Kenter, Vincent Wan, Chun-An Chan, Rob Clark, and Jakub Vit. Chive: Varying prosody in speech synthesis with a linguistically driven dynamic hierarchical conditional variational network. In *International Conference on Machine Learning*, pp. 3331–3340. PMLR, 2019.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pp. 5530–5540. PMLR, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. of NeurIPS*, 2020.

Robert Kubichek. Mel-cepstral distance measure for objective speech quality assessment. 1:125–128, 1993.

Max WY Lam, Jun Wang, Dan Su, and Dong Yu. Bddm: Bilateral denoising diffusion models for fast and high-quality speech synthesis. In *Proc. of ICLR*, 2022.

Adrian Łańcucki. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6588–6592. IEEE, 2021.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5400–5409, 2018.

Xiang Li, Changhe Song, Jingbei Li, Zhiyong Wu, Jia Jia, and Helen Meng. Towards multi-scale style control for expressive speech synthesis. *arXiv preprint arXiv:2104.03521*, 2021.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, Peng Liu, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. *arXiv preprint arXiv:2105.02446*, 2, 2021.

Zhengxi Liu, Qiao Tian, Chenxu Hu, Xudong Liu, Menglin Wu, Yuping Wang, Hang Zhao, and Yuxuan Wang. Controllable and lossless non-autoregressive end-to-end text-to-speech. *arXiv preprint arXiv:2207.06088*, 2022.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2837–2845, 2021.

Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang. Meta-stylespeech: Multi-speaker adaptive text-to-speech generation. pp. 7748–7759, 2021.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*, 2019.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pp. 8599–8608. PMLR, 2021.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Mark Hasegawa-Johnson, and David Cox. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pp. 7836–7846. PMLR, 2020.

Kaizhi Qian, Yang Zhang, Shiyu Chang, Jinjun Xiong, Chuang Gan, David Cox, and Mark Hasegawa-Johnson. Global rhythm style transfer without text transcriptions. *arXiv preprint arXiv:2106.08519*, 2021.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32, 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020.

Yi Ren, Ming Lei, Zhiying Huang, Shiliang Zhang, Qian Chen, Zhijie Yan, and Zhou Zhao. Prosospeech: Enhancing prosody with quantized vector pre-training in text-to-speech. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7577–7581. IEEE, 2022.

Eitan Richardson and Yair Weiss. On gans and gmms. In *Proc. of ICONIP*, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pp. 4693–4702. PMLR, 2018.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proc. of ICLR*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. of ICLR*, 2020b.

Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen, and Yonghui Wu. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6264–6268. IEEE, 2020.

Hao Sun, Xu Tan, Jun-Wei Gan, Hongzhi Liu, Sheng Zhao, Tao Qin, and Tie-Yan Liu. Token-level ensemble distillation for grapheme-to-phoneme conversion. *arXiv preprint arXiv:1904.03446*, 2019.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pp. 6309–6318, 2017.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Xin Wang, Shinji Takaki, Junichi Yamagishi, Simon King, and Keiichi Tokuda. A vector quantized variational autoencoder (vq-vae) autoregressive neural $f\_0$ model for statistical parametric speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:157–170, 2019.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*, pp. 5180–5189. PMLR, 2018.

Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, Christoph Feichtenhofer, et al. Masked autoencoders that listen. *arXiv preprint arXiv:2207.06405*, 2022.

Jinhyeok Yang, Jae-Sung Bae, Taejun Bak, Youngik Kim, and Hoon-Young Cho. Ganspeech: Adversarial training for high-fidelity multi-speaker speech synthesis. *arXiv preprint arXiv:2106.15153*, 2021a.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021b.

# Appendices

## Prosody-TTS: Self-Supervised Prosody Pre-training with Latent Diffusion For Text-to-Speech

## A Details of Models

In this section, we describe hyper-parameters and details of several modules.

### A.1 Model Configurations

We list the model hyper-parameters of Prosody-TTS in Table 4.

| Hyperparameter | | Prosody-TTS |
|---|---|---|
| Text Encoder | Phoneme Embedding | 192 |
| | Encoder Layers | 4 |
| | Encoder Hidden | 256 |
| | Encoder Conv1D Kernel | 9 |
| | Encoder Conv1D Filter Size | 1024 |
| | Encoder Attention Heads | 2 |
| | Encoder Dropout | 0.1 |
| Duration Predictor | Duration Predictor Conv1D Kernel | 3 |
| | Duration Predictor Conv1D Filter Size | 256 |
| | Duration Predictor Dropout | 0.5 |
| Prosody Generator | VQ Codebook Size | 1000 |
| | Latent Diffusion Residual Layers | 30 |
| | Latent Diffusion Residual Channels | 256 |
| | Latent Diffusion WaveNet Conv1d Kernel | 3 |
| | Latent Diffusion WaveNet Conv1d Filter | 512 |
| Diffusion Decoder | Diffusion Embedding | 256 |
| | Residual Layers | 20 |
| | Residual Channels | 256 |
| | WaveNet Conv1d Kernel | 3 |
| | WaveNet Conv1d Filter | 512 |
| Total Number of Parameters | | 53M |

Table 4: Hyperparameters of Prosody-TTS models.

### A.2 Duration prediction

Rhythm is a major component of prosody. In practice, however, baseline models demonstrate their superiority in generative modeling with inherited duration prediction (regression duration predictor (Ren et al., 2020; Min et al., 2021; Popov et al., 2021) or stochastic duration predictor (Kim et al., 2021; Casanova et al., 2022)).

Though the duration predictor can include textual and prosody-MAE features as joint input to improve rhythm modeling, we inherit the original architecture from baselines and leave it unchanged for a fair comparison. As such, we attribute the success of Prosody-TTS to better capturing and producing natural prosody patterns (e.g., local rises and falls of the pitch and stress).

For a better understanding of rhythm in comprising prosody, we conduct an extensive experiment to utilize the oracle duration derived by forced alignment in our backbone (FastSpeech 2) for comparison. The MOS-P evaluation procedure stays consistent with the manuscript, where we explicitly instruct the raters to "*focus on the naturalness of the prosody and style and ignore the differences of content, grammar, or audio quality.*".

As illustrated in Table 5, the optimized rhythm leads to the MOS-P gain with a score of 0.05 compared to the original model, while Prosody-TTS still demonstrates the leading performance. The perfect

| Method | MOS-P | MCD | FFE |
|---|---|---|---|
| Backbone | 3.92±0.07 | 3.88 | 0.43 |
| Backbone + GT duration | 3.97±0.06 | 3.79 | 0.40 |
| Prosody-TTS | 4.10±0.06 | 3.52 | 0.35 |

Table 5: Performance (audio quality and prosody naturalness) comparison with other models.

duration may still suffer some issues in prosody patterns (e.g., local rises and falls of the pitch and stress), indicating the necessity of explicitly adding a prosody branch for expressive text-to-speech.

### A.3 DIFFUSION MECHANISM

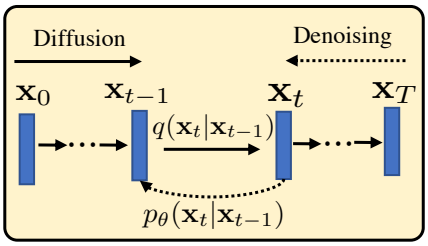



Figure 3: Graph for Diffusion.

For the training prosody latent diffusion model, the clean prosodic representation derived by Prosody-MAE passes through the vector quantification layer, which is also adopted to optimize the latent diffusion model (LDM) via the forward diffusion process. In inference time, the LDM samples diverse latent representations within the prosodic space through reverse backward denoising. According to the spectrogram denoiser, sampling from the Gaussian prior distribution is regarded as a common assumption. The diffusion decoder receives the textual hidden representation as a conditional signal and iteratively denoises Gaussian noise to reconstruct the target distribution by reverse sampling.

## B DETAILS OF PRE-TRAINING AND FINE-TUNING

We list the pre-training and fine-tuning settings in Table 6.

| Settings | | Values |
|---|---|---|
| Pre-training | Optimizer | Adam |
| | Base Learning Rate | 0.0001 |
| | Batch Size | 900 |
| | Optimizer Momentum | 0.9,0.98 |
| | Weight Decay | 0.01 |
| | Warmup Updates | 32000 |
| Fine-tuning | Optimizer | Adam |
| | Base Learning Rate | 0.0001 |
| | Batch Size | 4 |

Table 6: Pre-training and fine-tuning settings.

## C DIFFUSION PROBABILISTIC MODELS

Given i.i.d. samples $\{\mathbf{x}_0 \in \mathbb{R}^D\}$ from an unknown data distribution $p_{data}(\mathbf{x}_0)$. In this section, we introduce the theory of diffusion probabilistic model (Ho et al., 2020; Lam et al., 2022; Song et al., 2020a;b), and present diffusion and reverse process given by denoising diffusion probabilistic models (DDPMs), which could be used to learn a model distribution $p_\theta(\mathbf{x}_0)$ that approximates $p_{data}(\mathbf{x}_0)$.

**Diffusion process** Similar as previous work (Ho et al., 2020; Lam et al., 2022; Song et al., 2020a), we define the data distribution as $q(\mathbf{x}_0)$. The diffusion process is defined by a fixed Markov chain from data $x_0$ to the latent variable $x_T$:

$$q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T|x_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \tag{4}$$

For a small positive constant $\beta_t$, a small Gaussian noise is added from $x_t$ to the distribution of $x_{t-1}$ under the function of $q(x_t|x_{t-1})$.

The whole process gradually converts data $x_0$ to whitened latent $x_T$ according to the fixed noise schedule $\beta_1, \cdots, \beta_T$.

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t I) \tag{5}$$

Efficient training is optimizing a random term of $t$ with stochastic gradient descent:

$$\mathcal{L}_\theta = \left\| \boldsymbol{\epsilon}_\theta \left( \alpha_t \mathbf{x}_0 + \sqrt{1-\alpha_t^2}\boldsymbol{\epsilon} \right) - \boldsymbol{\epsilon} \right\|_2^2, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{6}$$

**Reverse process** Unlike the diffusion process, reverse process is to recover samples from Gaussian noises. The reverse process is a Markov chain from $x_T$ to $x_0$ parameterized by shared $\theta$:

$$p_\theta(\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{T-1}|x_T) = \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \tag{7}$$

where each iteration eliminates the Gaussian noise added in the diffusion process:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) := \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t), \sigma_\theta(\boldsymbol{x}_t, t)^2 I) \tag{8}$$

# D INFORMATION PERTURBATION

We apply the following functions (Qian et al., 2020; Choi et al., 2021) on acoustic features (i.e., pitch, and energy) to create acoustic-perturbed speech samples $\hat{S}$, while the linguistic content remains unchanged, including 1) formant shifting $fs$, 2) pitch randomization $pr$, and 3) random frequency shaping using a parametric equalizer $peq$.

- For $fs$, a formant shifting ratio is sampled uniformly from $\mathrm{Unif}(1, 1.4)$. After sampling the ratio, we again randomly decided whether to take the reciprocal of the sampled ratio or not.
- In $pr$, a pitch shift ratio and pitch range ratio are sampled uniformly from $\mathrm{Unif}(1, 2)$ and $\mathrm{Unif}(1, 1.5)$, respectively. Again, we randomly decide whether to take the reciprocal of the sampled ratios or not. For more details for formant shifting and pitch randomization, please refer to Parselmouth `https://github.com/YannickJadoul/Parselmouth`.
- $peq$ represents a serial composition of low-shelving, peaking, and high-shelving filters. We use one low-shelving HLS, one high-shelving HHS, and eight peaking filters HPeak.

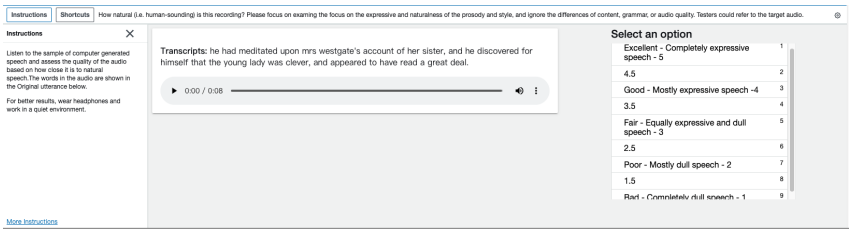
# E EVALUATION

## E.1 SUBJECTIVE EVALUATION

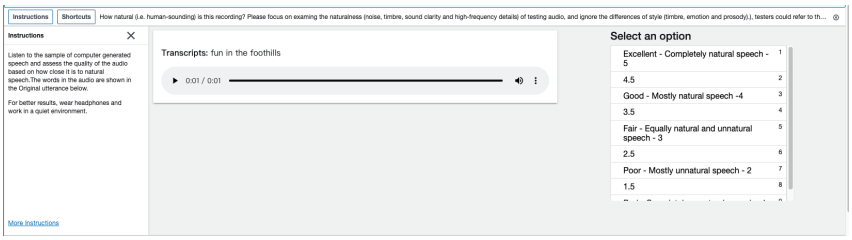
For MOS tests, the testers present and rate the samples, and each tester is asked to evaluate the subjective naturalness on a 1-5 Likert scale. For CMOS, listeners are asked to compare pairs of audio generated by systems A and B and indicate which of the two audio they prefer, and choose one of the following scores: 0 indicating no difference, 1 indicating a small difference, 2 indicating a large difference and 3 indicating a very large difference.

For quality evaluation, we explicitly instruct the raters to "*(focus on examining the audio quality and naturalness, and ignore the differences of style (timbre, emotion and prosody).)*". For prosody evaluation, we explicitly instruct the raters to "*(focus on the naturalness of the prosody and style, and ignore the differences of content, grammar, or audio quality.)*".

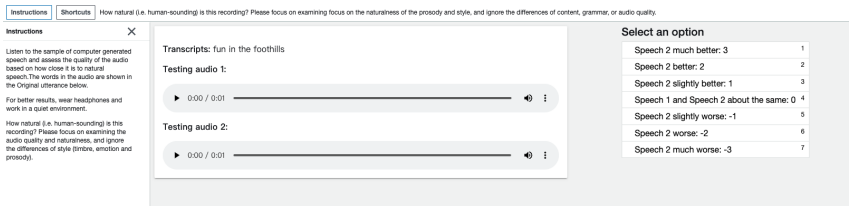
Our subjective evaluation tests are crowd-sourced and conducted by 25 native speakers via Amazon Mechanical Turk. The screenshots of instructions for testers have been shown in Figure 4. We paid $8 to participants hourly and totally spent about $800 on participant compensation. A small subset of speech samples used in the test is available at `https://Prosody-TTS.github.io/`.



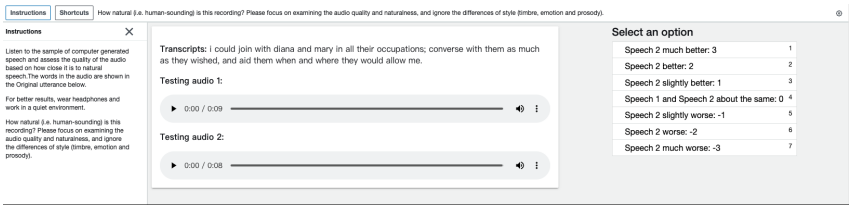

(a) Screenshot of MOS-P testing.




(b) Screenshot of MOS-Q testing.




(c) Screenshot of CMOS-P testing.




(d) Screenshot of CMOS-Q testing.

Figure 4: Screenshots of subjective evaluations.

### E.2 Objective Evaluation

Mel-cepstral distortion (MCD) (Kubichek, 1993) measures the spectral distance between the synthesized and reference mel-spectrum features.

F0 Frame Error (FFE) combines voicing decision error and F0 error metrics to capture F0 information.

Number of Statistically-Different Bins (NDB) and Jensen-Shannon divergence (JSD) (Richardson & Weiss, 2018). They measure diversity by 1) clustering the training data into several clusters, and 2) measuring how well the generated samples fit into those clusters.

## F PROSODY DIVERSITY

We employ two common metrics, including the number of Statistically-Different Bins (NDB) and Jensen-Shannon divergence (JSD), to explore the diversity of generated mel-spectrograms. Specifically, we 1) cluster the training data into several clusters, and 2) measure how well the generated samples fit into those clusters. For easy comparison, the results are presented in the following table.

| **Method** | **LJSpeech** | | **LibriTTS** | |
|---|---|---|---|---|
| | NDB | JS | NDB | JS |
| GT | / | / | / | / |
| GT(voc.) | 19 | 0.02 | 41 | 0.01 |
| FastSpeech 2 | 45 | 0.05 | 74 | 0.04 |
| Meta-StyleSpeech | 41 | 0.07 | 58 | 0.01 |
| Glow-TTS | 34 | 0.03 | 61 | 0.03 |
| Grad-TTS | 49 | 0.13 | 71 | 0.05 |
| YourTTS | 47 | 0.08 | 73 | 0.06 |
| Prosody-TTS | 30 | 0.03 | 52 | 0.01 |

Table 7: Diversity comparison with other models.

We can see that Prosody-TTS scores the superior NDB with scores of 30 (LJSpeech) and 52 (LibriTTS), demonstrating that Prosody-TTS surpasses the baseline models in generating samples with diverse prosody patterns (e.g., local rises and falls of the pitch and stress). Informally, by breaking the generation process into several conditional diffusion steps, generative latent diffusion prevents TTS from learning collapsed prosodic distribution. In contrast, most baseline models learn a mean distribution over their input data which leads to dull speech synthesis with similar patterns.

## G AXY EVALUATION

The evaluation of the TTS models is very challenging due to its subjective nature in the evaluation of the perceptual quality of generated speech. To further demonstrate the superiority of Prosody-TTS, we randomly choose 20 reference signals from the testing set and include an AXY test to assess the prosody expressiveness and naturalness following previous literature (Skerry-Ryan et al., 2018).

For each sentence (A), the listeners are asked to choose a preferred one among the samples synthesized by baseline models (X) and proposed Prosody-TTS (Y), from which AXY preference rates are calculated. The scale ranges of 7-point are from “X is much closer" to “Both are about the same distance" to “Y is much closer", and can naturally be mapped on the integers from -3 to 3. We present results in the following tables:

| Baseline | 7-point score | X | Neutral | Y |
|---|---|---|---|---|
| FastSpeech 2 | 1.13 $\pm$0.19 | 21% | 10% | 69% |
| Meta-StyleSpeech | 1.50$\pm$0.11 | 33% | 12% | 55% |
| Glow-TTS | 1.11$\pm$0.11 | 13% | 22% | 65% |
| Grad-TTS | 1.20$\pm$0.08 | 19 % | 21% | 60% |
| YourTTS | 1.42$\pm$0.10 | 28% | 13% | 59% |

The side-by-side subjective test indicates that raters prefer our model synthesis against baselines in terms of prosody naturalness and expressiveness. Without relying on text transcriptions or local prosody attributes, Prosody-TTS covers diverse speaking styles superior to all baseline models and avoid sub-optimal predictions.

## H COMPUTATIONAL COST

As DDPMs (Ho et al., 2020) are gradient-based models, a guarantee of high sample quality typically comes at the cost of hundreds to thousands of de-noising steps. Prosody-TTS adopts generative diffusion models for high-quality synthesis, and thus it inherently requires multiple iterative refinement for better results.

Though the denoising sampling could be accelerated by several techniques (e.g., scheduler (DDIM (Song et al., 2020a)), or training diagram (Progressive Distill (Salimans & Ho, 2022))), it is beyond our focus in this paper, and thus we leave it for future work.

## I REPRODUCIBILITY STATEMENT

We will release our code in the future. The Prosody-MAE model that we build upon is publicly available through the fairseq code repository (Ott et al., 2019). To aid reproducibility, we have included a schematic overview of the algorithm in Figure 1, and hyperparameters in Appendix 4.