NaturalSpeech 3: Zero-Shot Speech Synthesis with Factorized Codec and Diffusion Models

Zeqian Ju [*1,2] **Yuancheng Wang** [*3] **Kai Shen** [*4,2] **Xu Tan** [*2] **Detai Xin** [2,5] **Dongchao Yang** [2] **Yanqing Liu** [2] **Yichong Leng** [2] **Kaitao Song** [2] **Siliang Tang** [4] **Zhizheng Wu** [3] **Tao Qin** [2] **Xiang-Yang Li** [1] **Wei Ye** [6] **Shikun Zhang** [6] **Jiang Bian** [2] **Lei He** [2] **Jinyu Li** [2] **Sheng Zhao** [2]

Abstract

While recent large-scale text-to-speech (TTS) models have achieved significant progress, they still fall shorts in speech quality, similarity, and prosody. Considering that speech intricately encompasses various attributes (e.g., content, prosody, timbre, and acoustic details) that pose significant challenges for generation, a natural idea is to factorize speech into individual subspaces representing different attributes and generate them individually. Motivated by it, we propose *NaturalSpeech 3*, a TTS system with novel factorized diffusion models to generate natural speech in a zero-shot way. Specifically, 1) we design a neural codec with factorized vector quantization (FVQ) to disentangle speech waveform into subspaces of content, prosody, timbre, and acoustic details; 2) we propose a factorized diffusion model, which generates attributes in each subspace following its corresponding prompt. With this factorization design, NaturalSpeech 3 can effectively and efficiently model the intricate speech with disentangled subspaces in a divide-and-conquer way. Experimental results show that NaturalSpeech 3 outperforms the state-of-the-art TTS systems on quality, similarity, prosody, and intelligibility.

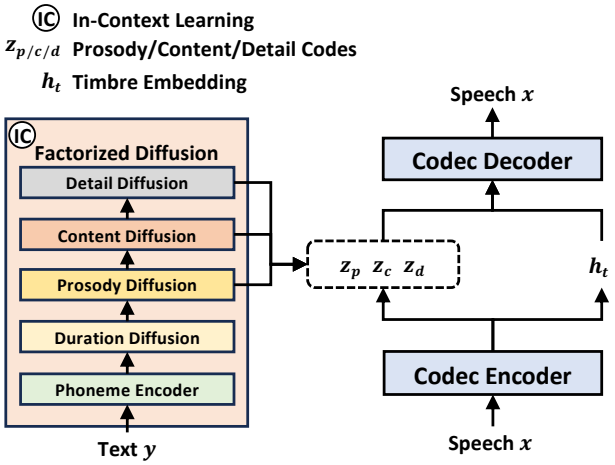

Figure 1: The overview of NaturalSpeech 3, with a neural speech codec for attribute factorization and a factorized diffusion model.

1. Introduction

In recent years, significant advancements have been achieved in text-to-speech (TTS) synthesis. Traditional TTS systems (Wang et al., 2017; Shen et al., 2018; Ren et al., 2019; Tan et al., 2022) are typically trained on limited datasets recorded in studios, and thus fail to support high-quality zero-shot speech synthesis. Recent works (Shen et al., 2023; Wang et al., 2023a; Jiang et al., 2023c) have made considerable progress for zero-shot TTS by largely scaling up both the corpus and the model sizes. However, the synthesis results of these large-scale TTS systems are not satisfactory in terms of voice quality, similarity, and prosody.

The challenges of inferior results stem from the intricate information embedded in speech, since speech encompasses numerous attributes, such as content, prosody, timbre, and acoustic detail. Previous works using raw waveform (Kim et al., 2021; Lim et al., 2022) and mel-spectrogram (Wang et al., 2017; Shen et al., 2018; Popov et al., 2021; Jiang et al., 2023c; Le et al., 2023) as data representations suffer from these intricate complexities during speech generation.

[*]Equal contribution [1]University of Science and Technology of China [2]Microsoft Research & Microsoft Azure [3]The Chinese University of Hong Kong, Shenzhen [4]Zhejiang University [5]The University of Tokyo [6]Peking University. Correspondence to: Xu Tan <xuta@microsoft.com>, Zhizheng Wu <wuzhizheng@cuhk.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

A natural idea is to factorize speech into disentangled subspaces representing different attributes and generate them individually. However, achieving this kind of disentangled factorization is non-trivial. Previous works (Borsos et al., 2022; 2023; Wang et al., 2023a) encode speech into multi-level discrete tokens using a neural audio codec (Zeghidour et al., 2021; Défossez et al., 2022) based on residual vector quantization (RVQ). Although this approach decomposes speech into different hierarchical representations, it does not effectively disentangle the information of different attributes of speech across different RVQ levels and still suffers from modeling complex coupled information.

To effectively generate speech with better quality, similarity and prosody, we propose NaturalSpeech 3, a TTS system with novel factorized diffusion models to generate natural speech in a zero-shot way. Specifically, 1) we introduce a novel neural speech codec with factorized vector quantization (FVQ), named FACodec, to decompose speech waveforms into distinct subspaces of content, prosody, timbre, and acoustic details and reconstruct speech waveforms with these disentangled representations, leveraging information bottleneck (Qian et al., 2020; 2019), various supervised losses, and adversarial training (Kong et al., 2020) to enhance disentanglement; 2) we propose a factorized diffusion model, which generates the factorized speech representations of duration, content, prosody, and acoustic detail, based on their corresponding prompts. This design allows us to use different prompts to control different attributes. The overview of NaturalSpeech 3 is shown in Figure 1.

We decompose complex speech into subspaces representing different attributes, thus simplifying the modeling of speech representation. This approach offers several advantages: 1) our factorized diffusion model is able to learn these disentangled representations efficiently, resulting in higher quality speech generation; 2) by disentangling timbre information in our FACodec, we enable our factorized diffusion model to avoid directly modeling timbre. This reduces learning complexity and leads to improved zero-shot speech synthesis; 3) we can use different prompts to control different attributes, enhancing the controllability of NaturalSpeech 3.

Benefiting from these designs, NaturalSpeech 3 has achieved significant improvements in speech quality, similarity, prosody, and intelligibility. Specifically, 1) it achieves comparable or better speech quality than the ground-truth speech on the LibriSpeech test set in terms of CMOS; 2) it achieves a new SOTA on the similarity between the synthesized speech and the prompt speech ($0.64 \rightarrow 0.67$ on Sim-O, $3.69 \rightarrow 4.01$ on SMOS); 3) it shows a significant improvement in prosody compared to other TTS systems with -0.16 average MCD (lower is better), $+0.21$ SMOS; 4) it achieves a SOTA on intelligibility ($1.94 \rightarrow 1.81$ on WER).5) it achieves human-level naturalness on multi-speaker datasets (e.g., LibriSpeech), another breakthrough after NaturalSpeech[1]. Audio samples can be found in https://speechresearch.github.io/naturalspeech3.

2. Background

In this section, we discuss the recent progress in TTS including: 1) zero-shot TTS; 2) speech representations in TTS; 3) generation methods in TTS; 4) speech attribute disentanglement.

Zero-shot TTS. Zero-shot TTS aims to synthesize speech for unseen speakers with speech prompts. We can systematically categorize these systems into four groups based on data representation and modelling methods: 1) Discrete Tokens + Autoregressive (Wang et al., 2023a; Kharitonov et al., 2023; Huang et al., 2023); 2) Discrete Tokens + Non-autoregressive (Borsos et al., 2023; Yang et al., 2023a; Du et al., 2023); 3) Continuous Vectors + Autoregressive (Nachmani et al., 2023); 4) Continuous Vectors + Non-autoregressive (Shen et al., 2023; Le et al., 2023; Li et al., 2023; Lee et al., 2023). Discrete tokens are typically derived from a neural codec, while continuous vectors are generally obtained from mel-spectrogram or latents from an audio autoencoder or a codec. In addition to the aforementioned perspectives, we disentangle speech waveforms into subspaces based on attribute disentanglement and propose a factorized diffusion model to generate attributes within each subspace, motivated by the principle of divide-and-conquer. Meanwhile, we can reuse previous methods, employing discrete tokens along with autoregressive models.

Speech Representations in TTS. Traditional works propose using prior-based speech representations such as raw waveforms (Oord et al., 2016; 2018; Sotelo et al., 2017) or mel-spectrogram (Ping et al., 2018; Li et al., 2019; Ren et al., 2019; Kim et al., 2020). Recently, large-scale TTS systems (Wang et al., 2023a; Borsos et al., 2023; Shen et al., 2023) leverage data-driven representations, i.e., either discrete tokens or continuous vectors form an auto-encoder (Zeghidour et al., 2021; Défossez et al., 2022; Kumar et al., 2023). However, these methods ignore that speech contains various complex attributes and encounter intricate complexities during speech generation. In this paper, we factorize speech into individual subspaces representing different attributes which can be effectively and efficiently modeled.

Generation Methods in TTS. Previous works have demonstrated that NAR-based models (Ren et al., 2019; Elias et al., 2020; Jiang et al., 2023c; Shen et al., 2023; Le et al., 2023)

[1]While NaturalSpeech 1 (Tan et al., 2024) achieved human-level quality on the single-speaker LJSpeech dataset, NaturalSpeech 3 achieved human-level quality on the diverse multi-speaker LibriSpeech dataset for the first time.

enjoy better robustness and generation speed than AR-based models, because they explicitly model the duration and predict all features simultaneously. Instead, AR-based models (Shen et al., 2018; Li et al., 2019; Wang et al., 2023a; Nachmani et al., 2023; Yang et al., 2023c) have better diversity, prosody, expressiveness, and flexibility than NAR-based models, due to their implicit duration modeling and token sampling strategy. In this study, we adopt the NAR modeling approach and propose a factorized diffusion model to support our disentangled speech representations and also extend it to AR modeling approaches. This allows NaturalSpeech 3 to achieve better expressiveness while maintaining stability and generation speed.

Speech Attribute Disentanglement. Prior works (Choi et al., 2021; 2022; Polyak et al., 2021) utilize disentangled representation for speech generation, such as speech content from self-supervised pre-trained models (Chung et al., 2021; Baevski et al., 2020; Schneider et al., 2019), fundamental frequency, and timbre, but their speech quality is not satisfying. Recently, some works explore attribute disentanglement in neural speech codec. SpeechTokenizer (Zhang et al., 2023) uses HuBERT (Hsu et al., 2021) for semantic distillation, aiming to render the first-layer RVQ representation as semantic information. Disen-TF-Codec (Jiang et al., 2023a) leverages a global and a local encoder to separate speaker identity from speech content, and applies the disentangled representations for zero-shot voice conversion. In this paper, we consider more speech attributes (e.g., content, prosody, acoustic details and timbre), and employ a series of robust decoupling techniques (e.g., information bottleneck, supervision, gradient reversal, and detail dropout), thus achieving better disentanglement. We validate such disentanglement can bring about significant improvements in zero-shot TTS task.

3. Method

3.1. Overall Architecture

In this section, we present NaturalSpeech 3, a cutting-edge system for natural and zero-shot text-to-speech synthesis with better speech quality, similarity and controllability. As shown in Figure 1, NaturalSpeech 3 consists of 1) a neural speech codec (FACodec) for attribute disentanglement; 2) a factorized diffusion model which generates factorized speech attributes. Since the speech waveform is complex and intricately encompasses various attributes, we factorize speech into five attributes including: duration, prosody, content, acoustic details, and timbre. Specifically, although the duration can be regarded as an aspect of prosody, we choose to model it explicitly due to our non-autoregressive speech generation design. We use our internal alignment tool to alignment speech and phoneme and obtain phoneme-level duration. For other attributes, we implicitly utilize the factorized neural speech codec to learn disentangled speech attribute subspaces (i.e., content, prosody, acoustic details, and timbre). Then, we use the factorized diffusion model to generate each speech attribute representation. Finally, we employ the codec decoder to reconstruct the waveform with the generated speech attributes. We introduce the FACodec in Section 3.2 and the factorized diffusion model in Section 3.3.

3.2. FACodec for Attribute Factorization

3.2.1. Model Overview

We propose a factorized neural speech codec (i.e., FACodec[2]) to convert complex speech waveform into disentangled subspaces representing speech attributes of content, prosody, timbre, and acoustic details and reconstruct high-quality speech waveform from these attributes.

As shown in Figure 2, FACodec consists of a speech encoder, a timbre extractor, three factorized vector quantizers (FVQ) for content, prosody, acoustic detail, and a speech decoder. Given a speech x, 1) following Zeghidour et al. (2021); Shen et al. (2023), we adopt several convolutional blocks for the speech encoder with a downsample rate of 200 for 16KHz speech data (i.e., each frame corresponding to a 12.5ms speech segment) to obtain pre-quantization latent h; 2) the timbre extractor is a Transformer encoder and a temporal pooling layer which converts the output of the speech encoder h into a global vector h_t representing the timbre attributes; 3) for other attribute i ($i = p, c, d$ for prosody, content, and acoustic detail, respectively), we use a factorized vector quantizer (FVQ_i) to capture fine-grained speech attribute representation and obtain corresponding discrete tokens z_i; 4) the speech decoder mirrors the structure of speech encoder but with much larger parameter amount to ensure high-quality speech reconstruction. We first add the representation of prosody, content, and acoustic details together and then fuse the timbre information by conditional layer normalization (Chen et al., 2021) to obtain the input z for the speech decoder. We discuss how to achieve better speech attribute disentanglement in the next section.

3.2.2. Attribute Disentanglement

Directly factorizing speech into different subspaces does not guarantee the disentanglement of speech. In this section, we introduce some techniques to achieve better speech attribute disentanglement: 1) information bottleneck, 2) supervision, 3) gradient reverse, and 4) detail dropout. Please refer to Appendix B.1 for more training details.

Information Bottleneck. Inspired by Qian et al. (2020;

[2]We release the code and pre-trained checkpoint of FACodec at https://huggingface.co/spaces/amphion/naturalspeech3_facodec.

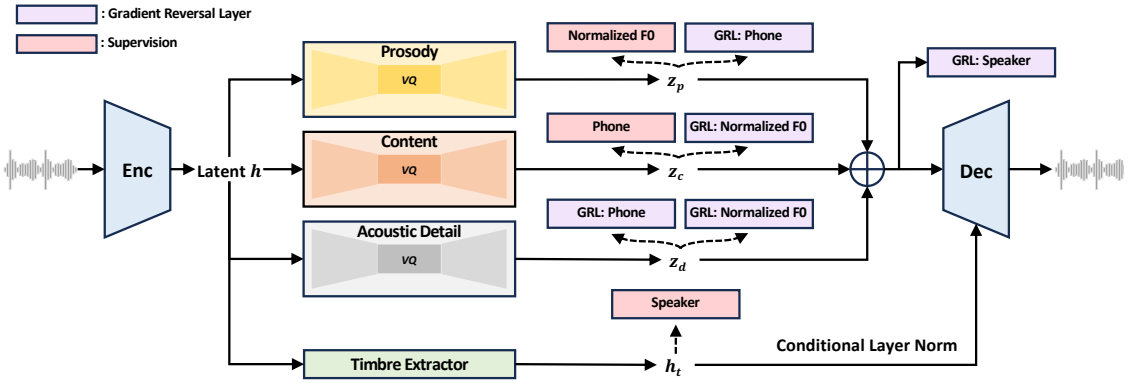

Figure 2: The framework of the speech codec for attribute factorization.

2019), to force the model to remove unnecessary information (such as prosody in content subspace), we construct the information bottleneck in prosody, content, and acoustic details FVQ by projecting the encoder output h into a low-dimensional space (i.e., 8-dimension) and subsequently quantize within this low-dimensional space. This technique ensures that each code embedding contains less information, facilitating information disentanglement (Kumar et al., 2023; Yu et al., 2021). After quantization, we will project the quantized vector back to the dimension of latent h (i.e., 256-dimension).

Supervision. To achieve high-quality speech disentanglement, we introduce supervision as auxiliary task for each attribute. For prosody, since pitch is an important part of prosody (Choi et al., 2022), we take the post-quantization latent z_p to predict pitch information. We extract the F0 for each frame and use normalized F0 (z-score) as the target. For content, we directly use the phoneme labels as the target (we use our internal alignment tool to get the frame-level phoneme labels). For timbre, we apply speaker classification on h_t by predicting the speaker ID.

Gradient Reversal. Avoiding the information leak (such as the prosody leak in content) can enhance disentanglement. Inspired by Yang et al. (2022), we adopt adversarial classifier with the gradient reversal layer (GRL) (Ganin & Lempitsky, 2015) to reduce undesired information in latent space. Specifically, for prosody, we apply phoneme-GRL (i.e., GRL layer by predicting phoneme labels) to reduce content information; for content, since the pitch is an important aspect of prosody, we apply F0-GRL to reduce the prosody information for simplicity; for acoustic details, we apply both phoneme-GRL and F0-GRL to reduce both content and prosody information. In addition, we apply speaker-GRL on the sum of z_p, z_c, z_d to reduce timbre.

Detail Dropout. We have the following considerations: 1) empirically, we find that the codec tends to preserve undesired information (e.g., content, prosody) in acoustic details subspace since there is no supervision; 2) intuitively, without acoustic details, the decoder should reconstruct speech only with prosody, content and timbre, although in low-quality. Motivated by them, we design the detail dropout by randomly masking out z_d during the training process with probability p. With detail dropout, we achieve the trade-off of disentanglement and reconstruction quality: 1) the codec can fully utilize the prosody, content and timbre information to reconstruct the speech to ensure the decouple ability, although in low-quality; 2) we can obtain high-quality speech when the acoustic details are given.

3.3. Factorized Diffusion Model

3.3.1. Model Overview

We generate speech with discrete diffusion for better generation quality. We have the following considerations: 1) we factorize speech into the following attributes: duration, prosody, content, and acoustic details, and generate them in sequential with specific conditions. Firstly, as we mentioned in Section 3.1, due to our non-autoregressive generation design, we first generate duration. Secondly, intuitively, the acoustic details should be generated at last; 2) following the speech factorization design, we only provide the generative model with the corresponding attribute prompt and apply discrete diffusion in its subspace; 3) to facilitate in-context learning in diffusion model, we utilize the codec to factorize speech prompt into attribute prompts (i.e., content, prosody and acoustic details prompt) and generate the target speech attribute with partial noising mechanism following Gong et al. (2022); Borsos et al. (2023). For example, for prosody generation, we directly concatenate prosody prompt (without noise) and target sequence (with noise) and gradually remove noise from target sequence with prosody prompt.

With these thoughts, as shown in Figure 3, we present our factorized diffusion model, which consists of a phoneme encoder and speech attribute (i.e., duration, prosody, content, and acoustic details) diffusion modules with the same discrete diffusion formulation: 1) we generate the speech duration by applying duration diffusion with duration prompt

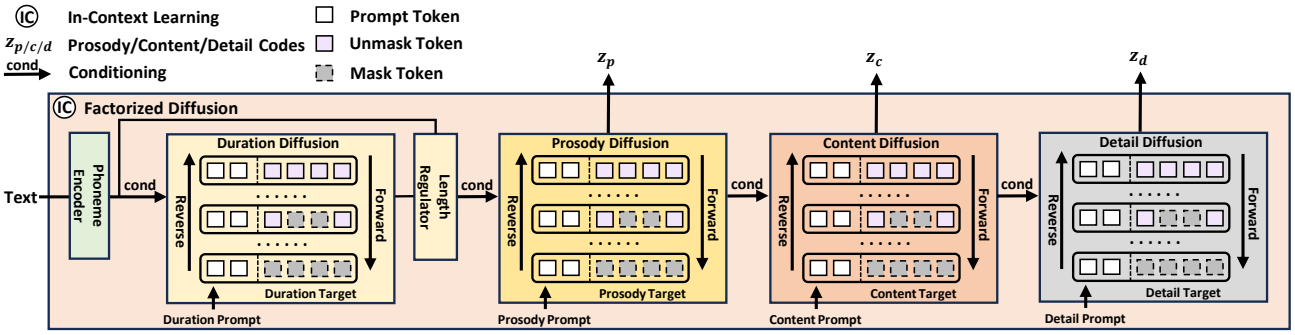

Figure 3: The framework of factorized diffusion model, which consists of 1) phoneme encoder, 2) duration diffusion and length regulator, 3) prosody diffusion, 4) content diffusion, 5) detail (acoustic detail) diffusion. Note that modules 2-5 shares the same diffusion formulation.

and phoneme-level textural condition encoded by phoneme encoder. Then we apply the length regulator to obtain frame-level phoneme condition c_{ph}; 2) we generate prosody z_p with prosody prompt and phoneme condition c_{ph}; 3) we generate content prosody z_c with content prompt and use generated prosody z_p and phoneme c_{ph} as condition; 4) we generate acoustic details z_d with acoustic details prompt and use generated prosody ,content and phoneme z_p, z_c, c_{ph} as conditions. Architecturally, we sum up the target sequence with the condition. Specifically, we do not explicitly generate the timbre attribute. Due to the factorization design in FACodec, we can obtain timbre from the prompt directly and do not need to generate it. We additionally leverage an auxiliary phoneme-level prosody diffusion model to generate phoneme-level prosody condition to facilitate accurate duration prediction. Please refer to Section 4.1 for more details. Finally, we synthesize the target speech by combining attributes z_p, z_c, z_d and h_t and decoding it with codec decoder. We discuss the diffusion formulation in Section 3.3.2.

3.3.2. Diffusion Formulation

Inspired by the notable achievements in text-to-image generation (Chang et al., 2022; Gu et al., 2022), discrete diffusion is being increasingly applied to the generation of speech and audio (Wu et al., 2024; Yang et al., 2023d). This subsection describes the forward and reverse process, and then details the inference method and classifier-free guidance.

Forward Process. Denote $\mathbf{X} = [x_i]_{i=1}^N$ the target discrete token sequence, where N is the sequence length, $\mathbf{X}^p$ is the prompt discrete token sequence, and $\mathbf{C}$ is the condition. The forward process at time t is defined as masking a subset of tokens in $\mathbf{X}$ with the corresponding binary mask $\mathbf{M}_t = [m_{t,i}]_{i=1}^N$, formulated as $\mathbf{X}_t = \mathbf{X} \odot \mathbf{M}_t$, by replacing x_i with [MASK] token if $m_{t,i} = 1$, and otherwise leaving x_i unmasked if $m_{t,i} = 0$. $m_{t,i} \overset{iid}{\sim} \text{Bernoulli}(\sigma(t))$ and $\sigma(t) \in (0, 1]$ is a monotonically increasing function. In this paper, $\sigma(t) = \sin(\frac{\pi t}{2T}), t \in (0, T]$. Specially, we denote $\mathbf{X}_0 = \mathbf{X}$ for the original token sequence and $\mathbf{X}_T$ for the fully masked sequence.

Reverse Process. The reverse process gradually restores $\mathbf{X}_0$ by sampling from reverse distribution $q(\mathbf{X}_{t-\Delta t}|\mathbf{X}_0, \mathbf{X}_t)$, starting from full masked sequence $\mathbf{X}_T$. Since $\mathbf{X}_0$ is unavailable in inference, we use the diffusion model p_θ, parameterized by θ, to predict the masked tokens conditioned on $\mathbf{X}^p$ and $\mathbf{C}$, denoted as $p_\theta(\mathbf{X}_0|\mathbf{X}_t, \mathbf{X}^p, \mathbf{C})$. The parameters θ are optimized to minimize the negative log-likelihood of the masked tokens:

$$\mathcal{L}_{\text{mask}} = \mathop{\mathbb{E}}_{\mathbf{X}\in\mathcal{D}, t\in[0,T]} -\sum_{i=1}^{N} m_{t,i} \cdot \log(p_\theta(x_i|\mathbf{X}_t, \mathbf{X}^p, \mathbf{C})).$$

Then we can get the reverse transition distribution:

$$p(\mathbf{X}_{t-\Delta t}|\mathbf{X}_t, \mathbf{X}^p, \mathbf{C}) = \mathop{\mathbb{E}}_{\hat{\mathbf{X}}_0 \sim p_\theta(\mathbf{X}_0|\mathbf{X}_t, \mathbf{X}^p, \mathbf{C})} q(\mathbf{X}_{t-\Delta t}|\hat{\mathbf{X}}_0, \mathbf{X}_t).$$

Inference. During inference, we progressively replace masked tokens, starting from the fully masked sequence $\mathbf{X}_T$, by iteratively sampling from $p(\mathbf{X}_{t-\Delta t}|\mathbf{X}_t, \mathbf{X}^p, \mathbf{C})$. Inspire by Chung et al. (2022); Gu et al. (2022); Lezama et al. (2022), we first sample $\hat{\mathbf{X}}_0$ from $p_\theta(\mathbf{X}_0|\mathbf{X}_t, \mathbf{X}^p, \mathbf{C})$, and then sample $\mathbf{X}_{t-\Delta t}$ from $q(\mathbf{X}_{t-\Delta t}|\hat{\mathbf{X}}_0, \mathbf{X}_t)$, which involves remask $\lfloor N \cdot \sigma(t - \Delta t) \rfloor$ tokens in $\hat{\mathbf{X}}_0$ with the lowest confidence score, where we define the confidence score of $\hat{x}_i$ in $\hat{\mathbf{X}}_0$ to $p_\theta(\hat{x}_i|\mathbf{X}_t, \mathbf{X}^p, \mathbf{C})$ if $m_{t,i} = 1$, otherwise, we set confidence score of x_i to 1, which means that tokens already unmasked in $\mathcal{X}_t$ will not be remasked.

Classifier-free Guidance. Moreover, we adapt the classifier-free guidance technique (Nichol et al., 2021; Ho & Salimans, 2022). Specifically, in training, we do not use the prompt with a probability of $p_{\text{cfg}} = 0.15$. In inference, we extrapolate the logit output of the model towards the conditional generation guided by the prompt $g_{\text{cond}} = g(\mathbf{X}|\mathbf{X}^p)$ and away from the unconditional generation $g_{\text{uncond}} = g(\mathbf{X})$, i.e., $g_{\text{cfg}} = g_{\text{cond}} + \alpha \cdot (g_{\text{cond}} - g_{\text{uncond}})$, with a guidance scale α selected based on experimental results. We then

rescale it through $g_{\text{final}} = \text{std}(g_{\text{cond}}) \times g_{\text{cfg}}/\text{std}(g_{\text{cfg}})$, following Lin et al. (2024).

4. Experiments and Results

4.1. Experimental Settings

In this subsection, we introduce the training, inference and evaluation for the Factorized Diffusion Model. Please refer to Appendix A.1 for model configuration, and Appendix B.1 for implementation details of the FACodec.

Implementation Details. We use Librilight (Kahn et al., 2020), which contains 60K hours of 16KHz unlabeled speech data and around 7000 distinct speakers from LibriVox audiobooks, as the training set. In duration diffusion, each token represents the duration (the number of frames) of each corresponding phoneme. We further improve the performance by conditioning phoneme-level prosody codes, as shown in Figure 4. Specifically, we perform phoneme-level pooling according to duration on the pre-quantized vectors, and then feed these phoneme-level representations into the prosody quantizer in our codec to obtain the phoneme-level prosody codes. We employ an additional discrete diffusion to generate these in inference. We perform 4 iterations in each diffusion process. We generate duration without classifier-free guidance and generate others with a classifier-free guidance scale of 1.0. This strategy results in 4×2 for phoneme-level prosody, 4 for duration, 4×2 for each token sequence of prosody, content, and acoustic details, totaling 60 forward passes due to the double computation with classifier-free guidance. As shown in Figure 1, these generated speech attribute codes, along with the timbre representation which is separately derived from the prompt via timbre extractor in FACodec, are passed to the codec decoder for speech synthesis. Please refer to Appendix A.2 for more details of our factorization diffusion model.

Evaluation Dataset. We employ two benchmark datasets: 1) LibriSpeech (Panayotov et al., 2015) test-clean, a widely-used testset for zero-shot TTS task. It contains 40 distinct speakers and 5.4-hour speech. Following Shen et al. (2023), we randomly select one sentence for each speaker for LibriSpeech test-clean benchmark. Specifically, we randomly select 3-second clips as prompts from the same speaker's speech. 2) RAVDESS (Livingstone & Russo, 2018), an emotional TTS dataset featuring 24 professional actors (12 female, 12 male) across 8 emotions (neutral, calm, happy, sad, angry, fearful, surprise, and disgust) in 2 emotional intensity (normal and strong). We use strong-intensity samples for RAVDESS benchmark. We adopt this benchmark for prosody evaluation, considering 1) for the same speaker, speech with the same emotion shares similar prosody, while speech with different emotions displays varied prosodies; 2) the benchmark provides speech samples with the same text from the same speaker across eight different emotions.

Evaluation Metrics. Objective Metrics: In the Librispeech test-clean benchmark, we evaluate both speaker-similarity (SIM-O and SIM-R) and robustness (WER). In specific, 1) for SIM-O and SIM-R, we employ the WavLM-TDCNN[3] speaker embedding model to assess speaker similarity between generated samples and the prompt. Results are reported for both similarity to original prompt (SIM-O) and reconstructed prompt (SIM-R). 2) for speech quality, we employ UTMOS (Saeki et al., 2022) which is a surrogate objective metric of MOS; 3) for Word Error Rate (WER), we use an ASR model[4] to transcribe generated speech. The model is a CTC-based HuBERT pre-trained on Librilight and fine-tuned on the 960 hours training set of LibriSpeech. We also use an advanced ASR model based on transducer (He et al., 2019)[5]. We additionally present the WER results on more samples of LibriSpeech test-clean dataset. Following (Wang et al., 2023a), we select utterances with a length of 4-second to 10-second, which results in 1205 utterances. In the RAVDESS benchmark, we evaluate the prosody similariy (MCD and MCD-Acc). In specific, 1) following Al-Radhi et al. (2023), we adopt Mel-Ceptral Distortion (MCD) for prosody evaluation by measuring the differences between generated samples and ground truth samples. We employ dynamic time warping (DTW) to align the generated speech with the ground truth. We report the results for eight emotions, along with the average result. 2) for MCD-Acc, we evaluate the top-1 emotion accuracy of the generated speech on the RAVDESS benchmark for prosodic similarity measures. Specifically, we adopt a K-Nearest-Neighbors (KNN) model as emotion classifier. We compare MCD distances between the generated speech and the ground-truth speech from the same speaker, across eight different emotions. Subjective Metrics: We employ comparative mean option score (CMOS) and similarity mean option score (SMOS) in both two benchmarks to evaluate naturalness and similarity, respectively.

Evaluation Baselines. We compare NaturalSpeech 3 with baselines: 1) VALL-E (Wang et al., 2023a). 2) NaturalSpeech 2 (Shen et al., 2023). 3) Voicebox (Le et al., 2023). 4) Mega-TTS 2 (Jiang et al., 2023b). 5) UniAudio (Yang et al., 2023c). 6) StyleTTS 2 (Li et al., 2023). 7) HierSpeech++ (Lee et al., 2023). Please refer to Appendix A.3 for details.

[3] https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification

[4] https://huggingface.co/facebook/hubert-large-ls960-ft

[5] https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge

Table 1: The evaluation results for NaturalSpeech 3 and the baseline methods on LibriSpeech test-clean. ♠ means the results are obtained from the authors. ♥ means the results directly obtained from the paper. ♣ means the results are infered from offical checkpoints. ♦ means the reproduced results. Abbreviation: LT (LibriTTS), V (VCTK), LJ (LJSpeech), LL* (Librilight Small, Medium), EX (Expresso), MS (MSSS Kor), NI (NIKL Kor).The '(WER)' results shown in parenthesis are evaluated across 1205 utterances ranging from 4 to 10 seconds in LibriSpeech test-clean dataset. Please refer to Appendix A.4 for more results on 1) WER inferred by an advanced ASR system, and 2) UTMOS, an automatic metric for MOS.

	Model Size	Training Data	Sim-O ↑	Sim-R ↑	WER↓	CMOS↑	SMOS↑
Ground Truth	-	-	0.68	-	0.34 (2.14)	+0.08	3.85
VALL-E ♥	0.4B	Librilight	-	0.58	- (5.90)	-	-
VALL-E ♦	0.4B	Librilight	0.47	0.51	6.11 (6.22)	-0.60	3.46
NaturalSpeech 2♠	0.4B	Librilight	0.55	0.62	1.94 (2.60)	-0.18	3.65
Voicebox♠	0.4B	Self-Collected (60kh)	0.64	0.67	2.03 (-)	-0.23	3.69
Voicebox♦	0.4B	Librilight	0.48	0.50	2.14 (2.89)	-0.32	3.52
Mega-TTS 2♠	0.4B	Librilight	0.53	-	2.32 (-)	-0.20	3.63
UniAudio♠	1.0B	Mixed (165kh)	0.57	0.68	2.49 (-)	-0.25	3.71
StyleTTS 2♣	0.2B	LT+V+LJ	0.38	-	2.49 (2.75)	-0.21	3.07
HierSpeech++♣	0.1B	LT+ LL*+EX+MS+NI	0.51	-	6.33 (3.85)	-0.41	3.50
NaturalSpeech 3	0.5B	Librilight	**0.67**	**0.76**	**1.81** (**2.41**)	**0.00**	**4.01**

4.2. Experimental Results on Zero-shot TTS

In this subsection, we compare NaturalSpeech 3 with baselines in terms of: 1) generation quality in Section 4.2.1; 2) generation similarity in Section 4.2.2; 3) robustness in Section 4.2.3. Specifically, for generation similarity, we evaluate in two aspects: 1) speaker similarity; 2) prosody similarity. Please refer to Appendix A.5 for latency analysis.

4.2.1. Generation Quality

To evaluate speech quality, we conduct CMOS test, with 12 native as the judges. We randomly select 20 utterances from both LibriSpeech test-clean and RAVDESS benchmarks. As shown in Table 1, we find that 1) NaturalSpeech 3 is close to the ground-truth recording (-0.08 on Librispeech test-clean, and -0.17 on RAVDESS), which demonstrates NaturalSpeech 3 can generate high-quality and natural speech; 2) NaturalSpeech 3 outperforms baselines by a substantial margin, verifying the effectiveness of NaturalSpeech 3 with factorization.

4.2.2. Generation Similarity

Speaker Similarity. We evaluate the speech similarity with both objective metrics (Sim-O and Sim-R) and subjective metrics (SMOS), with 12 natives as the judges. We randomly select 10 utterances for SMOS test. As shown in Table 1, we find that 1) NaturalSpeech 3 achieves parity in Sim-O and a 0.16 increase in SMOS with ground truth, which indicates great speaker similarity achieved by our proposed method; 2) NaturalSpeech 3 outperforms all baselines on both objective and subjective metrics, highlighting the superiority of NaturalSpeech 3 with factorization in terms of speaker similarity. Additionally, we notice certain discrepancy between Sim-O and SMOS. For instance, the SMOS is not as competitive as SIM-O for Voicebox model, likely due to some unnatural prosody.

Table 2: The evaluation results for NaturalSpeech 3 and the baseline methods on RAVDESS. ♠ means the results are obtained from the authors. ♣ means the results are inferred from official checkpoints. ♦ means the reproduced results. Abbreviation: Avg (average MCD), Acc (MCD-Acc).

	Avg↓	Acc↑	CMOS↑	SMOS↑
Ground Truth	0.00	1.00	+0.17	4.42
VALL-E ♦	5.03	0.34	-0.55	3.80
NaturalSpeech 2♠	4.56	0.25	-0.22	4.04
Voicebox♦	4.88	0.34	-0.34	3.92
Mega-TTS 2♠	4.44	0.39	-0.20	4.51
StyleTTS 2♣	4.50	0.40	-0.25	3.98
HierSpeech++♣	6.08	0.30	-0.37	3.87
NaturalSpeech 3	**4.28**	**0.52**	**0.00**	**4.72**

Prosody Similarity. We evaluate prosody similarity with both objective metrics (MCD and MCD-Acc) and subjective metrics (SMOS) on the RAVDESS benchmark. We randomly select 10 utterances for SMOS test. As shown in Table 2, NaturalSpeech 3 consistently surpasses baselines by a remarkable margin in MCD avg, MCD-Acc, and SMOS. It reveals that NaturalSpeech 3 achieves a significant improvement in terms of prosodic similarity. Please refer to Appendix A.7 for the MCD scores across 8 emotions.

4.2.3. Robustness

We assess the robustness of our zero-shot TTS by measuring the word error rate (WER) of generated speech on the LibriSpeech test-clean benchmark. In addition, we evaluate the WER results across 1205 utterances ranging from 4 to 10 seconds in LibriSpeech test-clean dataset. The results in Table 1 indicate that 1) NaturalSpeech 3 achieves a comparable WER with ground truth evaluated across 1205 utterances, proving the high intelligibility; 2) NaturalSpeech 3 outperforms other baselines by a considerable margin, which demonstrates the superior robustness of NaturalSpeech 3.

4.2.4. Human-Level Naturalness on LibriSpeech Testset

We compare the speech synthesized by NaturalSpeech 3 with human recordings (Ground Truth) in Table 1 (more results can be found in Table 7 in Appendix A.4). We have the following observations: 1) NaturalSpeech 3 achieves -0.01 Sim-O and +0.16 SMOS compared to human recordings, which demonstrates that our method is on par or better on speaker similarity; 2) NaturalSpeech 3 achieves -0.08 CMOS and +0.16 UTMOS compared with recording, which demonstrates that our method can generate on-par or better voice quality; 3) Our method also achieves close WER with human recordings, which demonstrates the robustness of NaturalSpeech 3. Therefore, we can conclude that for the first time, NaturalSpeech 3 has achieved human-level quality and naturalness on the multi-speaker LibriSpeech test set in a zero-shot way. It is another great milestone after NaturalSpeech 1 (Tan et al., 2024) has achieved human-level quality on the single-speaker LJSpeech dataset.

4.3. Ablation Study

In this subsection, we conduct ablation studies to verify the effectiveness of 1) factorization; 2) classifier-free guidance; 3) prosody representation. We also conduct ablation study to compare our duration diffusion model with traditional duration predictor in Appendix A.6.

Factorization. To verify the proposed factorization method, we ablate it by removing factorization in both codec and factorized diffusion model. Specifically, we 1) use the discrete tokens from SoundStream, a neural codec which does not consider factorization, and 2) do not consider factorization in generation. For fair comparison, we use Soundstream with double bandwidth (i.e., a codebook number of 12) as the baseline, since it achieves a comparable reconstruction performance, detailed in Table 11. Specifically, the forward and reverse processes are similar to those of the factorized diffusion model. We generate residual codes sequentially in a coarse-to-fine manner (starting from RVQ level 1 up to N). During this process, all previously generated coarser codes act as conditions when generating the finer codes. As shown in Table 3, we could find a significant performance degradation without the factorization, a drop of 0.12 in Sim-O, 0.15 in Sim-R, 0.68 in WER, 0.25 in CMOS and 0.42 in SMOS. This indicates the proposed factorized method can consistently improve the performance in terms of speaker similarity, robustness, and quality.

Table 3: The ablation study of factorization and classifier-free guidance (cfg) on LibriSpeech test-clean.

	Sim-O / Sim-R ↑	WER↓	CMOS↑	SMOS↑
NaturalSpeech 3	**0.67 / 0.76**	**1.81**	**0.00**	**4.01**
- factorization	0.55 / 0.61	2.49	-0.25	3.59
- cfg	0.64 / 0.72	**1.81**	-0.06	3.80

Classifier-Free Guidance. We conduct an ablation study by dropping the classifier-free guidance in inference to validate its effectiveness. We double the iterations to ensure the same 60 forward passes for fair comparison. Table 3 illustrates a significant degradation without classifier-free guidance, a decrease of 0.03 in Sim-O, 0.04 in Sim-R, 0.06 in CMOS and 0.21 in SMOS, proving that classifier-free guidance can greatly help the speaker similarity and quality.

Prosody Representation. We compare different prosody representations on zero-shot TTS task. In specific, we select handcrafted prosody features (e.g., the first 20 bins of mel-spectrogram (Jiang et al., 2023c; Oh et al., 2023; Ren et al., 2022)) as the baseline. We drop the prosody FVQ module and directly quantize the first 20 bins of the mel-spectrogram, without the normalized F0 loss. Table 5 shows that using "Mel 20 Bins" as prosody representation demonstrates inferiority in terms of prosody similarity compared to the prosody representations learned from codec (4.34 vs 4.28 in average MCD, 0.46 vs 0.52 in MCD-Acc).

4.4. Method Analyses

In this subsection, we first discuss the extensibility of our factorization in Section 4.4.1. We then introduce the application of speech attributes manipulation in a zero-shot way in Section 4.4.2.

4.4.1. Extensibility

NaturalSpeech 3 utilizes a non-autoregressive model for discrete token generation with factorization design. To validate the extensibility of our proposed factorization method, we further explore the autoregressive generative model for discrete token generation under our factorization framework. We utilize VALL-E for verification. We first employ an autoregressive language model to generate prosody codes, followed by a non-autoregressive model to generate the remaining content and acoustic details codes. This approach maintains a consistent order of attribute generation, allow-

Table 4: The reconstruction quality evaluation of codecs. ♣ means results are infered from offical checkpoints. ★ means the reproduced checkpoint. ♦ means the reproduced model following the original paper's implementation and experimental setup. All models use a codebook size of 1024. **Bold** for the best result and underline for the second-best result.

Models	Sampling Rate	Hop Size	Codebook Number	Bandwidth	PESQ ↑	STOI ↑	MSTFT ↓	MCD ↓
EnCodec♣	24kHz	320	8	6.0 kbps	3.28	0.94	0.99	2.70
HiFi-Codec♣	16kHz	320	4	2.0 kbps	3.17	0.93	0.98	3.05
DAC♣	16kHz	320	9	4.5 kbps	**3.52**	**0.95**	0.97	2.65
SoundStream♦	16kHz	200	6	4.8 kbps	3.03	0.90	1.07	3.38
FACodec	16kHz	200	6	4.8 kbps	3.47	**0.95**	**0.93**	**2.59**

Table 5: The ablation study of prosody representation on RAVDESS. Denote "Mel 20 Bins" using the first 20 bins in the mel-spectrogram as the prosody representation.

	MCD Avg↓	MCD-Acc↑
NaturalSpeech 3	**4.28**	**0.52**
Mel 20 Bins	4.34	0.46

Table 6: The comparison between autoregressive approach with (VALL-E + F) and without (VALL-E) our proposed factorization on LibriSpeech test-clean. ♦ means the reproduced results. Abbreviation: Sim-O/R (Sim-O / Sim-R).

	Sim-O / R ↑	WER↓	CMOS↑	SMOS↑
VALL-E + F	**0.57 / 0.65**	**5.60**	**+0.24**	**3.61**
VALL-E♦	0.47 / 0.51	6.11	0.00	3.46

ing for a fair comparison. We name it VALL-E + F. As shown in Table 6, VALL-E + F consistently outperforms VALL-E by a considerable margin in all objective and subjective metrics, demonstrating the factorization design can enhance VALL-E in speech similarity, quality and generation robustness. It further shows our factorization paradigm is not limited in the proposed factorization diffusion model and has a large potential in other generative models. We leave it for future work.

4.4.2. Speech Attribute Manipulation

As discussed in Section 3.3, our factorized diffusion model enables attribute manipulation by selecting different attributes prompts from different speech. We mainly focus on manipulating duration, prosody, and timbre, since the content codes are dictated by the text in TTS, and the acoustic details do not carry semantic information. Leveraging the strong in-context capability of NaturalSpeech 3, the generated speech effectively mirrors the corresponding speech attributes. For instance, 1) we can utilize the timbre prompt from a different speech to control the timbre while keeping other attributes unchanged; 2) despite the correlation between duration and prosody, we can still solely adjust duration prompt to regulate the speed; 3) moreover, we can combine different speech attributes from disparate samples as desired. This allow us to mimic the timbre while using different prosody and speech speed. Samples are available on our demo page[6].

[6] https://speechresearch.github.io/naturalspeech3

4.5. Experimental Results on FACodec

We compare proposed FACodec in terms of the reconstruction quality with strong baselines, such as EnCodec (Défossez et al., 2022), HiFi-Codec (Yang et al., 2023b), Descript-Audio-Codec (DAC) (Kumar et al., 2023), and our reproduced SoundStream (Zeghidour et al., 2021). Table 4 shows that our codec significantly surpasses SoundStream in the same bandwidth setting (0.44 in PESQ, 0.05 in STOI, 0.14 in MSTFT and 0.79 in MCD, respectively). Check more details in Appendix B.2. Compared with other baselines, FACodec also get comparable performance. Additionally, since our codec decouples timbre information, it can enable zero-shot voice conversion easily, we provide the details and experiment results in Appendix B.3. Appendix B.4 shows some ablation studies about our FACodec.

5. Conclusion

In this paper, we develop a TTS system which consists of 1) novel neural speech codec (i.e., FACodec) with factorized vector quantization to decompose speech waveform into distinct subspaces of content, prosody, acoustic details and timbre and 2) novel factorized diffusion model to synthesize speech by generating attributes in subspaces with discrete diffusion. NaturalSpeech 3 outperforms the state-of-art TTS system on speech quality, similarity, prosody, and intelligibility. We also show that NaturalSpeech 3 can enable speech attribute manipulation, by customizing speech attribute prompts. We include our limitation and future works in Appendix C.

Impact Statement

Since our model could synthesize speech with great speaker similarity, it may carry potential risks in misuse of the model, such as spoofing voice identification or impersonating a specific speaker. We conducted the experiments under the assumption that the user agree to be the target speaker in speech synthesis. To prevent misuse, it is crucial to develop a robust synthesized speech detection model and establish a system for individuals to report any suspected misuse.

References

Al-Radhi, M. S., Csapó, T. G., and Németh, G. Nonparallel expressive tts for unseen target speaker using style-controlled adaptive layer and optimized pitch embedding. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 176–181. IEEE, 2023.

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

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

Borsos, Z., Sharifi, M., Vincent, D., Kharitonov, E., Zeghidour, N., and Tagliasacchi, M. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023.

Casanova, E., Weber, J., Shulby, C. D., Junior, A. C., Gölge, E., and Ponti, M. A. 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.

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Chen, M., Tan, X., Li, B., Liu, Y., Qin, T., sheng zhao, and Liu, T.-Y. AdaSpeech: Adaptive text to speech for custom voice. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Drynvt7gg4L.

Choi, H.-S., Lee, J., Kim, W., Lee, J., Heo, H., and Lee, K. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265, 2021.

Choi, H.-S., Yang, J., Lee, J., and Kim, H. Nansy++: Unified voice synthesis with neural analysis and synthesis. *arXiv preprint arXiv:2211.09407*, 2022.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., and Wu, Y. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250. IEEE, 2021.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Du, C., Guo, Y., Shen, F., Liu, Z., Liang, Z., Chen, X., Wang, S., Zhang, H., and Yu, K. Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding. *arXiv preprint arXiv:2306.07547*, 2023.

Elias, I., Zen, H., Shen, J., Zhang, Y., Jia, Y., Weiss, R., and Wu, Y. Parallel Tacotron: Non-autoregressive and controllable TTS. *arXiv preprint arXiv:2010.11439*, 2020.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Gong, S., Li, M., Feng, J., Wu, Z., and Kong, L. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., and Guo, B. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

He, Y., Sainath, T. N., Prabhavalkar, R., McGraw, I., Alvarez, R., Zhao, D., Rybach, D., Kannan, A., Wu, Y., Pang, R., et al. Streaming end-to-end speech recognition for mobile devices. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6381–6385. IEEE, 2019.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. 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.

Huang, R., Zhang, C., Wang, Y., Yang, D., Liu, L., Ye, Z., Jiang, Z., Weng, C., Zhao, Z., and Yu, D. Make-a-voice: Unified voice synthesis with discrete representation. *arXiv preprint arXiv:2305.19269*, 2023.

Jiang, X., Peng, X., Zhang, Y., and Lu, Y. Disentangled feature learning for real-time neural speech coding. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Jiang, Z., Liu, J., Ren, Y., He, J., Zhang, C., Ye, Z., Wei, P., Wang, C., Yin, X., Ma, Z., et al. Mega-tts 2: Zero-shot text-to-speech with arbitrary length speech prompts. *arXiv preprint arXiv:2307.07218*, 2023b.

Jiang, Z., Ren, Y., Ye, Z., Liu, J., Zhang, C., Yang, Q., Ji, S., Huang, R., Wang, C., Yin, X., et al. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*, 2023c.

Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673. IEEE, 2020.

Kharitonov, E., Vincent, D., Borsos, Z., Marinier, R., Girgin, S., Pietquin, O., Sharifi, M., Tagliasacchi, M., and Zeghidour, N. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*, 2023.

Kim, J., Kim, S., Kong, J., and Yoon, S. Glow-TTS: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33, 2020.

Kim, J., Kong, J., and Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv preprint arXiv:2106.06103*, 2021.

Kong, J., Kim, J., and Bae, J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33, 2020.

Kumar, R., Seetharaman, P., Luebs, A., Kumar, I., and Kumar, K. High-fidelity audio compression with improved rvqgan. *arXiv preprint arXiv:2306.06546*, 2023.

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., et al. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*, 2023.

Lee, S.-g., Ping, W., Ginsburg, B., Catanzaro, B., and Yoon, S. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

Lee, S.-H., Choi, H.-Y., Kim, S.-B., and Lee, S.-W. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. *arXiv preprint arXiv:2311.12454*, 2023.

Lezama, J., Chang, H., Jiang, L., and Essa, I. Improved masked image generation with token-critic. In *European Conference on Computer Vision*, pp. 70–86. Springer, 2022.

Li, N., Liu, S., Liu, Y., Zhao, S., and Liu, M. Neural speech synthesis with Transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6706–6713, 2019.

Li, Y. A., Han, C., Raghavan, V. S., Mischler, G., and Mesgarani, N. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *arXiv preprint arXiv:2306.07691*, 2023.

Lim, D., Jung, S., and Kim, E. Jets: Jointly training fastspeech2 and hifi-gan for end to end text to speech. *arXiv preprint arXiv:2203.16852*, 2022.

Lin, S., Liu, B., Li, J., and Yang, X. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5404–5411, 2024.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Nachmani, E., Levkovitch, A., Salazar, J., Asawaroengchai, C., Mariooryad, S., Skerry-Ryan, R., and Ramanovich, M. T. Lms with a voice: Spoken language modeling beyond speech tokens. *arXiv preprint arXiv:2305.15255*, 2023.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Oh, H.-S., Lee, S.-H., and Lee, S.-W. Diffprosody: Diffusion-based latent prosody generation for expressive speech synthesis with prosody conditional adversarial training. *arXiv preprint arXiv:2307.16549*, 2023.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Oord, A. v. d., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., et al. Parallel WaveNet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pp. 3918–3926. PMLR, 2018.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. 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.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. Deep Voice 3: 2000-speaker neural text-to-speech. *Proc. ICLR*, pp. 214–217, 2018.

Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.-N., Mohamed, A., and Dupoux, E. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*, 2021.

Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., and Kudinov, M. Grad-TTS: A diffusion probabilistic model for text-to-speech. *arXiv preprint arXiv:2105.06337*, 2021.

Qian, K., Zhang, Y., Chang, S., Yang, X., and Hasegawa-Johnson, M. AutoVC: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pp. 5210–5219. PMLR, 2019.

Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., and Cox, D. Unsupervised speech decomposition via triple information bottleneck. In *International Conference on Machine Learning*, pp. 7836–7846. PMLR, 2020.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. FastSpeech: Fast, robust and controllable text to speech. In *NeurIPS*, 2019.

Ren, Y., Lei, M., Huang, Z., Zhang, S., Chen, Q., Yan, Z., and Zhao, Z. 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.

Saeki, T., Xin, D., Nakata, W., Koriyama, T., Takamichi, S., and Saruwatari, H. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. wav2vec: Unsupervised pre-training for speech recognition. *Proc. Interspeech 2019*, pp. 3465–3469, 2019.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Shen, K., Ju, Z., Tan, X., Liu, Y., Leng, Y., He, L., Qin, T., Zhao, S., and Bian, J. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., and Bengio, Y. Char2wav: End-to-end speech synthesis. 2017.

Sun, H., Tan, X., Gan, J.-W., Liu, H., Zhao, S., Qin, T., and Liu, T.-Y. Token-level ensemble distillation for grapheme-to-phoneme conversion. In *INTERSPEECH*, 2019.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., et al. NaturalSpeech: End-to-end text to speech synthesis with human-level quality. *arXiv preprint arXiv:2205.04421*, 2022.

Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., Wang, X., Leng, Y., Yi, Y., He, L., et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6309–6318, 2017.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023a.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., et al. Tacotron: Towards end-to-end speech synthesis. *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

Wang, Z., Chen, Y., Xie, L., Tian, Q., and Wang, Y. Lm-vc: Zero-shot voice conversion via speech generation based on language models. *arXiv preprint arXiv:2306.10521*, 2023b.

Wu, Z., Li, Q., Liu, S., and Yang, Q. Dctts: Discrete diffusion model with contrastive learning for text-to-speech generation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11336–11340. IEEE, 2024.

Yang, D., Liu, S., Huang, R., Lei, G., Weng, C., Meng, H., and Yu, D. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *arXiv preprint arXiv:2301.13662*, 2023a.

Yang, D., Liu, S., Huang, R., Tian, J., Weng, C., and Zou, Y. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023b.

Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., Shi, J., Zhao, S., Bian, J., Wu, X., et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023c.

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., and Yu, D. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023d.

Yang, S., Tantrawenith, M., Zhuang, H., Wu, Z., Sun, A., Wang, J., Cheng, N., Tang, H., Zhao, X., Wang, J., et al. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. *arXiv preprint arXiv:2208.08757*, 2022.

Yu, J., Li, X., Koh, J. Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., and Wu, Y. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

Zhang, X., Zhang, D., Li, S., Zhou, Y., and Qiu, X. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023.

A. Details of Factorization Diffusion Model

A.1. Model Configuration

The phoneme encoder uses a similar architecture as Shen et al. (2023) and comprises a 6-layer Transformer with 8 attention heads, 512 embedding dimensions, filter size 2048 and kernel size 9 for 1D convolution, and a dropout of 0.1. In prosody, content and acoustic details diffusion, we adopt a 12-layer Transformer, with 8 attention heads, 1024 embedding dimensions, filter size 2048 and kernel size 3 for 1D convolution, and a dropout of 0.1. The weights of Transformers in prosody, content and acoustic details diffusion are shared for training efficiency. We additionally use conditional layer normalization in each Transformer block to support diffusion time input. In phoneme-level prosody and duration diffusion, we adopt a 6-layer Transformer with 8 attention heads, 1024 embedding dimensions, filter size 2048 and kernel size 3 for 1D convolution, and a dropout of 0.1. We also use conditional layer normalization in the model to support diffusion time input.

A.2. Training and Inference Details

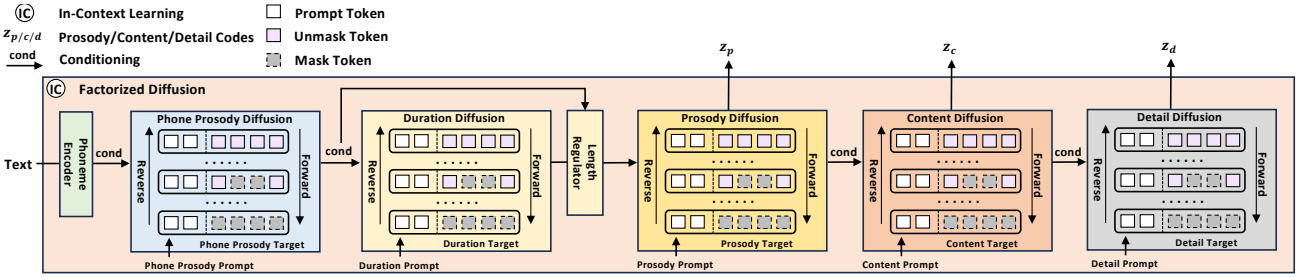

Figure 4: The detailed framework of factorized diffusion model, which consists of 1) phoneme encoder, 2) duration diffusion and length regulator, 3) prosody diffusion, 4) content diffusion, 5) detail (acoustic detail) diffusion. We additionally integrate 6) a phoneme-level prosody diffusion, thereby providing conditions to facilitate accurate duration prediction. Note that modules 2-6 shares the same diffusion formulation.

We use Librilight (Kahn et al., 2020), which contains 60K hours of 16KHz unlabeled speech data and around 7000 distinct speakers from LibriVox audiobooks, as the training set. We transcribe using an internal ASR system, convert transcriptions to phonemes via grapheme-to-phoneme conversion (Sun et al., 2019), and obtain duration with an internal alignment tool. We use 8 A100 80GB GPUs with a batch size of 10K frames of latent vectors per GPU for 1M steps. We use the AdamW optimizer with a learning rate of $1e-4$, $\beta_1 = 0.9$, and $\beta_2 = 0.98$, 5K warmup steps following the inverse square root learning schedule.

During inference, we perform 4 iterations in each diffusion process, including phoneme-level prosody, duration, prosody, content and acoustic details diffusion, as shown in Figure 4. We generate duration without classifier-free guidance, and generate others with a classifier-free guidance scale of 1.0. This strategy results a 4×2 for phoneme-level prosody, 4 for duration, 4×2 for each token sequence of prosody, content and acoustic details, totaling 60 forward passes due to the double computation with classifier-free guidance. We use a top-k of 20, with sampling temperature annealing from 1.5 to 0. Following Chang et al. (2022), Gumbel noises are added to token confidences when determining which positions to re-mask in $q(\mathbf{X}_{t-\Delta t}|\hat{\mathbf{X}}_0, \mathbf{X}_t)$, mentioned in Section 3.3.2.

A.3. Evaluation Baselines

We compare NaturalSpeech 3 with following strong zero-shot TTS baselines:

- VALL-E (Wang et al., 2023a). It use an autoregressive and an additional non-autoregressive model for discrete token generation. We report the scores directly obtained from the paper. We additionally reproduce it using discrete tokens from SoundStream on Librilight.
- NaturalSpeech 2 (Shen et al., 2023). It use a non-autoregressive model for continuous vectors generation. We obtain samples through communication with the authors.
- Voicebox (Le et al., 2023). It use a non-autoregressive model for continuous vectors generation. We obtain samples through communication with the authors. We additionally reproduce it using mel-spectrogram on Librilight.

Table 7: The evaluation results for NaturalSpeech 3 and the baseline methods on LibriSpeech test-clean. ♠ means the results are obtained from the authors. ♥ means the results directly obtained from the paper. ♣ means the results are inferred from offical checkpoints. ♦ means the reproduced results. WER* means the word error rate calculated by an advanced ASR system mentioned in A.4. The '(WER)' results shown in parenthesis are evaluated across 1205 utterances ranging from 4 to 10 seconds in LibriSpeech test-clean dataset.

	Sim-O ↑	Sim-R ↑	WER ↓	WER*↓	UTMOS ↑	CMOS↑	SMOS↑
Ground Truth	0.68	-	0.34 (2.14)	0.68 (1.78)	4.14	+0.08	3.85
VALL-E ♥	-	0.58	- (5.90)	- (-)	-	-	-
VALL-E ♦	0.47	0.51	6.11 (6.22)	4.87 (5.46)	3.68	-0.60	3.46
NaturalSpeech 2♠	0.55	0.62	1.94 (2.60)	1.24 (2.24)	3.88	-0.18	3.65
Voicebox♠	0.64	0.67	2.03 (-)	1.81 (-)	3.82	-0.23	3.69
Voicebox♦	0.48	0.50	2.14 (2.89)	1.24 (2.31)	3.73	-0.32	3.52
Mega-TTS 2♠	0.53	-	2.32 (-)	2.17 (-)	4.02	-0.20	3.63
UniAudio♠	0.57	0.68	2.49 (-)	1.81 (-)	3.79	-0.25	3.71
StyleTTS 2♣	0.38	-	2.49 (2.75)	1.58 (2.27)	3.94	-0.21	3.07
HierSpeech++♣	0.51	-	6.33 (3.85)	4.97 (2.70)	3.80	-0.41	3.50
NaturalSpeech 3	**0.67**	**0.76**	**1.81** (**2.41**)	**1.13** (**1.99**)	**4.30**	**0.00**	**4.01**

- Mega-TTS 2 (Jiang et al., 2023b). It use a non-autoregressive model for continuous vectors generation. We obtain samples through communication with the authors.
- UniAudio (Yang et al., 2023c). It use an autoregressive model for discrete token generation. We obtain samples through communication with the authors.
- StyleTTS 2 (Li et al., 2023). It use a non-autoregressive model for continuous vectors generation. We use official code and checkpoint[7].
- HierSpeech++ (Lee et al., 2023). It use a non-autoregressive model for continuous vectors generation. We use official code and checkpoint[8]. We do not use its super resolution model for fair comparison.

A.4. More Experimental Results on Zero-shot TTS

In this section, we report more evaluation results for NaturalSpeech 3 and other baselines on: 1) WER, inferred by an advanced ASR system[9]; 2) UTMOS (Saeki et al., 2022), which is a surrogate objective metric of MOS. The results are shown in Table 7.

A.5. Latency Analysis

In this subsection, we compare the inference latency of NaturalSpeech 3 with an autoregressive method (VALL-E) and a non-autoregressive method (NaturalSpeech 2). We also investigate the effect of reducing the number of iterations in each diffusion from 4 to 1, resulting in a total of 15 forward passes. We call this variant NaturalSpeech 3 one-step. We evaluate the performance on Librispeech test-clean in terms of speaker similarity (Sim-O/Sim-R) and quality (UTMOS (Saeki et al., 2022) [10], a surrogate objective metric of CMOS). The latency tests are conducted on a server with E5-2690 Intel Xeon CPU, 512GB memory, and one NVIDIA V100 GPU. The results are shown in Table 8. From the results, we have several observations. 1) NaturalSpeech 3 achieves a $15.27\times$ speedup over VALL-E and $1.24\times$ speedup over NaturalSpeech 2, while consistently surpasses these baselines on all metrics. This demonstrate NaturalSpeech 3 is both effective and efficient. 2) when using fewer diffusion steps, NaturalSpeech 3 can still maintain robust performance (-0.01 in Sim-O, -0.01 in Sim-R, and -0.29 in UTMOS) with a $4.41\times$ faster speed, proving the robustness of diffusion steps.

[7] https://github.com/yl4579/StyleTTS2
[8] https://github.com/sh-lee-prml/HierSpeechpp
[9] https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge
[10] https://github.com/tarepan/SpeechMOS

Table 8: The latency study on LibriSpeech test-clean. NaturalSpeech 3 one-step denotes using only 1 iteration in each diffusion process instead of original 4. Abbreviation: NFE (number of function evaluation).

Models	NFE	RTF ↓	Sim-O ↑	Sim-R ↑	UTMOS ↑
NaturalSpeech 2	150	0.366	0.55	0.62	3.87
VALL-E	-	4.520	0.47	0.51	3.67
NaturalSpeech 3	60	0.296	**0.67**	**0.76**	**4.30**
NaturalSpeech 3 one-step	15	**0.067**	0.66	0.75	4.01

Table 9: The ablation results of the design of the duration predictor on LibriSpeech test-clean.

	Sim-O ↑	Sim-R ↑	WER↓	UTMOS↑
NaturalSpeech 3	**0.67**	**0.76**	**1.94**	**4.30**
Generation ablation	0.62	0.73	**1.94**	4.18
Objective ablation	0.62	0.72	2.38	4.13
Conditioning ablation	0.62	0.72	2.49	4.11
Prompting ablation	0.61	0.71	2.83	4.08

A.6. Ablation Study on Duration Diffusion Model

In this subsection, we conduct an ablation study to compare our duration discrete diffusion model with the traditional duration predictor, which regresses the duration in logarithmic domain. The ablation study focus on 1) Generation: multi-step generation vs. one-step generation. 2) Objective: classification-based cross-entropy loss vs. regression-based L2 loss. 3) Conditioning: with vs. without phoneme-level prosody conditioning. 4) Prompting: with vs. without duration prompting. We evaluate them on Librispeech test-clean in terms of speaker similarity (Sim-O/Sim-R), robustness (WER) and qualtiy (UTMOS). As shown in Table 9, we can find that 1) without multi-step generation, there's a significant drop in performance (-0.05 in Sim-O, -0.03 in Sim-R, and -0.12 in UTMOS). 2) replacing cross-entropy loss with l2 loss affects the performance, causing a decrease of -0.05 in Sim-O, -0.04 in Sim-R, 0.44 in WER and -0.17 in UTMOS. 3) dropping phoneme-level prosody conditioning will affect both speaker similarity (-0.05 in Sim-O and -0.04 in Sim-R), robustness (0.55 in WER) and quality (-0.19 in UTMOS) 4) the duration prompting mechanism is crucial for speaker similarity, robustness and quality, with changes of -0.06 in Sim-O, -0.05 in Sim-R, 0.89 in WER and -0.22 in UTMOS. These results confirm that each design aspect of our duration predictor contributes to performance improvement.

A.7. Details of Prosody Similarity Evaluation

In Table 10, we present MCD on 8 different emotions, comparing NaturalSpeech 3 with the baseline methods on the RAVDESS benchmark. NaturalSpeech 3 demonstrates robust performance across 8 emotions, verifying the effectiveness and robustness in terms of prosody similarity.

B. Details of FACodec

B.1. Implementation Details

Model Architecture. The basic architecture of our codec encoder and decoder follows Kumar et al. (2023) and employs the SnakeBeta activation function (Lee et al., 2022). The timbre extractor consists of several conformer (Gulati et al., 2020) blocks. We use $N_{q_c} = 2, N_{q_p} = 1, N_{q_d} = 3$ as the number of quantizers for each of the three FVQ $\mathcal{Q}^c$, $\mathcal{Q}^p$, $\mathcal{Q}^d$, the codebook size for all the quantizers is 1024.

Loss Functions. We utilize the multi-scale mel-reconstruction loss $\mathcal{L}_{\text{rec}}$ as detailed in Kumar et al. (2023). For the adversarial loss $\mathcal{L}_{\text{adv}}$, we employ both the multi-period discriminator (MPD) and the multi-band multi-scale STFT discriminator, as proposed by Kumar et al. (2023). Additionally, we incorporate the relative feature matching loss $\mathcal{L}_{\text{feat}}$. For codebook learning, we use the codebook loss $\mathcal{L}_{\text{codebook}}$ and the commitment loss $\mathcal{L}_{\text{commit}}$ from VQ-VAE (van den Oord et al., 2017). The training loss also includes the phone prediction loss $\mathcal{L}_{\text{ph}}$, the normalized F0 prediction loss $\mathcal{L}_{\text{f0}}$, and the gradient reverse

Table 10: The MCD scores on 8 different emotions of NaturalSpeech 3 and the baseline methods on RAVDESS. ♠ means the results are obtained from the authors. ♣ means the results are inferred from official checkpoints. ♦ means the reproduced results. We use **bold** to indicate the best result and underline to indicate the second-best result.

	MCD↓							
	neutral	calm	happy	sad	angry	fearful	disgust	surprised
Ground Truth	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VALL-E ♦	3.97	4.75	4.83	5.51	5.19	5.29	5.45	5.29
Voicebox ♦	3.93	4.90	4.96	4.93	5.01	5.03	5.34	4.89
NaturalSpeech 2 ♠	**2.77**	**3.51**	4.85	4.88	5.42	5.23	5.31	4.52
Mega-TTS 2 ♠	3.28	4.39	4.44	4.67	**4.21**	5.00	5.42	**4.14**
StyleTTS 2 ♣	3.41	4.38	4.40	4.64	4.80	4.69	5.10	4.57
HierSpeech++ ♣	5.54	6.55	5.78	5.84	6.37	6.17	6.74	5.62
NaturalSpeech 3	3.23	4.32	**4.26**	**4.41**	4.64	**4.25**	**4.80**	4.45

losses of phone prediction $\mathcal{L}_{\text{gr-ph}}$, normalized F0 prediction $\mathcal{L}_{\text{gr-f0}}$, and speaker classification $\mathcal{L}_{\text{gr-spk}}$ for disentanglement learning. The total training loss for the generator can be formulated as: $\lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}} + \lambda_{\text{codebook}}\mathcal{L}_{\text{codebook}} + \lambda_{\text{commit}}\mathcal{L}_{\text{commit}} + \lambda_{\text{ph}}\mathcal{L}_{\text{ph}} + \lambda_{\text{f0}}\mathcal{L}_{\text{f0}} + \lambda_{\text{gr-ph}}\mathcal{L}_{\text{gr-ph}} + \lambda_{\text{gr-f0}}\mathcal{L}_{\text{gr-f0}} + \lambda_{\text{gr-spk}}\mathcal{L}_{\text{gr-spk}}$, where λ_{rec}, λ_{adv}, λ_{feat}, $\lambda_{\text{codebook}}$, λ_{commit},λ_{f0}, λ_{ph}, $\lambda_{\text{gr-f0}}$, $\lambda_{\text{gr-ph}}$ and $\lambda_{\text{gr-spk}}$ are coefficients for balancing each loss terms. In our paper, we set these coefficients as follows: $\lambda_{\text{rec}} = 10.0$, $\lambda_{\text{adv}} = 2.0$, $\lambda_{\text{feat}} = 2.0$, $\lambda_{\text{codebook}} = 1.0$, $\lambda_{\text{commit}} = 0.25$, $\lambda_{\text{f0}} = 5.0$, $\lambda_{\text{ph}} = 5.0$, $\lambda_{\text{gr-f0}} = 5.0$, $\lambda_{\text{gr-ph}} = 5.0$ and $\lambda_{\text{gr-spk}} = 1.0$.

Training Details. We use Librilight as the training set. We train the codec using 8 NVIDIA TESLA V100 32GB GPUs with a batch size of 32 speech clips of 16000 frames each per GPU for 800K steps. We use the Adam optimizer with a learning rate of $2e-4$, $\beta_1 = 0.5$, and $\beta_2 = 0.9$.

B.2. Reconstruction Performance Comparison

We evaluate the reconstruction performance with the following objective metrics: Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), Multi-Resolution STFT Distance (MSTFT), and Mel-Cepstral Distortion (MCD). These metrics collectively measure the difference between the original and the reconstructed samples. We select the following open-source codec models as baselines: EnCodec (Défossez et al., 2022)[11], HiFi-Codec (Yang et al., 2023b)[12], and Descript-Audio-Codec (DAC) (Kumar et al., 2023)[13]. We additionally reproduce SoundStream (Zeghidour et al., 2021) following the original paper's implementation and experimental setup. Table 11 shows that 1) FACodec significantly surpasses SoundStream in the same bandwidth setting (0.44 in PESQ, 0.05 in STOI, 0.14 in MSTFT and 0.79 in MCD, respectively). Moreover, FACodec achieves on-par performance with SoundStream even when its bandwidth is doubled (0.02 in PESQ, 0.01 in STOI, -0.01 in MSTFT and 0.17 in MCD, respectively). 2) For a fair comparison, we compare FACodec with other baselines in a similar bandwidth. FACodec achieve comparable or better result on most metrics than these strong baselines, which means that we can still achieve excellent reconstruction speech quality when disentangling speech attributes.

B.3. Zero-shot Voice Conversion

Voice conversion aims to transform speech from a source speaker into that of a target speaker, preserving content while altering timbre. Zero-shot voice conversion achieves this by utilizing a prompt speech sample from the target speaker to convert the source speaker's speech. FACodec achieves zero-shot voice conversion by extracting the speaker embedding h_t^{prompt} from the prompt speech to replace the speaker embedding h_t^{source} from the source speech, and utilizing content codes z_c^{source}, prosody codes z_p^{source}, and detail codes z_d^{source} from the source speaker to reconstruct the target speech $\mathcal{D}(z_c^{source}, z_p^{source}, z_d^{source}, h_t^{prompt})$. We compare FACodec with some previous SOTA models: YourTTS (Casanova et al.,

[11] https://github.com/facebookresearch/encodec
[12] https://github.com/yangdongchao/AcademiCodec
[13] https://github.com/descriptinc/descript-audio-codec

Table 11: The reconstruction quality evaluation of codecs. ♣ means results are infered from offical checkpoints. ★ means the reproduced checkpoint. ♦ means the reproduced model following the original paper's implementation and experimental setup. All models use a codebook size of 1024. We use **bold** to indicate the best result and underline to indicate the second-best result.

Models	Sampling Rate	Hop Size	Codebook Number	Bandwidth	PESQ ↑	STOI ↑	MSTFT ↓	MCD ↓
EnCodec♣	24kHz	320	8	6.0 kbps	3.28	0.94	0.99	2.70
EnCodec★	16kHz	320	10	5.0 kbps	3.10	0.92	0.97	3.10
HiFi-Codec♣	16kHz	320	4	2.0 kbps	3.17	0.93	0.98	3.05
DAC♣	16kHz	320	9	4.5 kbps	**3.52**	**0.95**	0.97	2.65
SoundStream♦	16kHz	200	6	4.8 kbps	3.03	0.90	1.07	3.38
SoundStream♦	16kHz	200	12	9.6 kbps	3.45	0.94	**0.92**	2.76
FACodec	16kHz	200	6	4.8 kbps	3.47	**0.95**	0.93	**2.59**

2022), Make-A-Voice (VC) (Huang et al., 2023), LM-VC (Wang et al., 2023b), and UniAudio (Yang et al., 2023c). We use VCTK dataset for comparison. We use Sim-O[14] to compare speaker similarity to baselines and WER to evaluate speech quality. Table 12 shows the evaluation results. The experimental results demonstrate that FACodec solely achieves comparable similarity and superior intelligence compared to the state-of-the-art zero-shot VC models, which need additional training on this task. This implies that FACodec achieves superior disentanglement, especially in timbre.

Table 12: The zero-shot voice conversion evaluation results for FACodec with previous SOTA methods. We use **bold** to indicate the best result and underline to indicate the second-best result.

Models	Sim-O ↑	WER ↓
Ground Truth	-	3.25
YourTTS	0.72	10.1
Make-A-Voice (VC)	0.68	6.20
LM-VC	0.82	4.91
UniAudio	**0.87**	4.80
FACodec	0.86	**3.46**

B.4. Ablation Study

In this subsection, we study 1) the impact of the information bottleneck on the disentanglement of FACodec; 2) the effect of gradient reversal on the disentanglement of FACodec; 3) the role of the acoustic details quantizers; 4) the effects of different prosody representations for TTS generation.

Information Bottleneck for Disentanglement. We investigate the impact of the information bottleneck on speech disentanglement through qualitative analysis. We find that without using information bottleneck (quantize in original dimensional space rather than low dimensional space) can lead to incomplete disentanglement. For example, we conduct zero-shot voice conversion in the same experimental setting using the FACodec without information bottleneck, as mentioned in Appendix B.3. We observe that the timbre of the converted speech is the interpolation between the source and target, indicating its poor timbre disentanglement. Table 13 demonstrates that without the information bottleneck, the speaker similarity of zero-shot voice conversion decreases by 0.13.

Gradient Reversal for Disentanglement. We investigate the impact of gradient reversal on the disentanglement of the FACodec through qualitative analysis. We observe that not using gradient reversal diminishes the disentangling ability of FACodec. For instance, removing the content and prosody gradient reversal from the acoustic detail module results in some content and prosody information leaking into the detail acoustic. We can confirm this by solely reconstructing the speech using detail codes and timbre embedding, where partial content and pitch variations can be heard.

Role of Acoustic Details Quantizer. Although content, prosody, and timbre information already encompass the majority of

[14] https://huggingface.co/microsoft/wavlm-base-plus-sv

Table 13: Comparison of zero-shot voice conversion evaluation results for FACodec with and without using information bottleneck.

	Sim-O ↑
w. information bottleneck	**0.86**
w.o. information bottleneck	0.73

speech information, Table 14 demonstrates that employing acoustic details quantizers enhances the speech reconstruction quality of FACodec. We find 1) without using acoustic details quantizers (only utilizing three codebooks), FACodec achieves comparable or better results compared to SoundStream with using three codebooks, which means that content codes, prosody codes, and timbre embedding already contain most of the necessary information for speech reconstruction; 2) adding acoustic details achieves better reconstruction quality, which suggests that acoustic details codes primarily serve to supplement high-frequency details.

Table 14: The reconstruction quality comparison between our FACodec with and without using acoustic details quantizers.

	Codebook Number	PESQ ↑	STOI ↑	MSTFT ↓	MCD ↓
FACodec	6	**3.47**	**0.95**	**0.93**	**2.59**
- acoustic details quantizers	3	<u>3.09</u>	<u>0.92</u>	1.08	<u>3.12</u>
SoundStream	6	3.03	0.90	<u>1.07</u>	3.38

C. Limitation and Future Works

Despite our proposed TTS system has achieved great progress, we still have the following limitations:

Attribute Coverage. In this work, we propose the factorization design for speech representation and generation, and have achieved significant improvement by factorizing speech into content, prosody, duration, acoustic details and timbre. However, these attributes can not coverage all speech aspects. For example, we can not extract the background sounds, which is a common challenge for speech disentanglement. In the future, we will explore more attributes including: 1. energy, 2. background sounds, and etc.

Data Coverage. Although we have achieved remarkable improvement on zero-shot speech synthesis on speech quality, similarity and robustness, NaturalSpeech 3 is trained on English corpus from LibriVox audiobooks. Thus, it can not coverage real word people's diverse voice and can not support multilingual TTS. In the future, we will address this limitation by collecting more speech data with larger diversity.

Neural Speech Codec. Although our FACodec can factorize speech into attributes and reconstruct with high quality, it still has the following limitations: 1) we need phoneme transcription for content supervision, which limits the scalability; 2) we only verified the disentanglement in zero-shot TTS task. In the future, firstly, we will explore more general methods to achieve better disentanglement, especially without supervision. Secondly, we would like to explore more tasks with the FACodec, such as zero-shot voice conversion and automatic speech recognition.