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

Zeqian Ju[12*], **Yuancheng Wang**[3*], **Kai Shen**[41*], **Xu Tan**[1*], **Detai Xin**[15], **Dongchao Yang**[1], **Yanqing Liu**[1], **Yichong Leng**[1], **Kaitao Song**[1], **Siliang Tang**[4], **Zhizheng Wu**[3], **Tao Qin**[1], **Xiang-Yang Li**[2], **Wei Ye**[6], **Shikun Zhang**[6], **Jiang Bian**[1], **Lei He**[1], **Jinyu Li**[1], **Sheng Zhao**[1]
[1]Microsoft Research & Microsoft Azure
[2]University of Science and Technology of China [3]The Chinese University of Hong Kong, Shenzhen
[4]Zhejiang University, [5]The University of Tokyo, [6]Peking University
https://aka.ms/speechresearch

Abstract

While recent large-scale text-to-speech (TTS) models have achieved significant progress, they still fall short in speech quality, similarity, and prosody. Considering 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 to generate attributes in each subspace following its corresponding prompt. With this factorization design, NaturalSpeech 3 can effectively and efficiently model intricate speech with disentangled subspaces in a divide-and-conquer way. Experiments show that NaturalSpeech 3 outperforms the state-of-the-art TTS systems on quality, similarity, prosody, and intelligibility, and achieves on-par quality with human recordings. Furthermore, we achieve better performance by scaling to 1B parameters and 200K hours of training data.

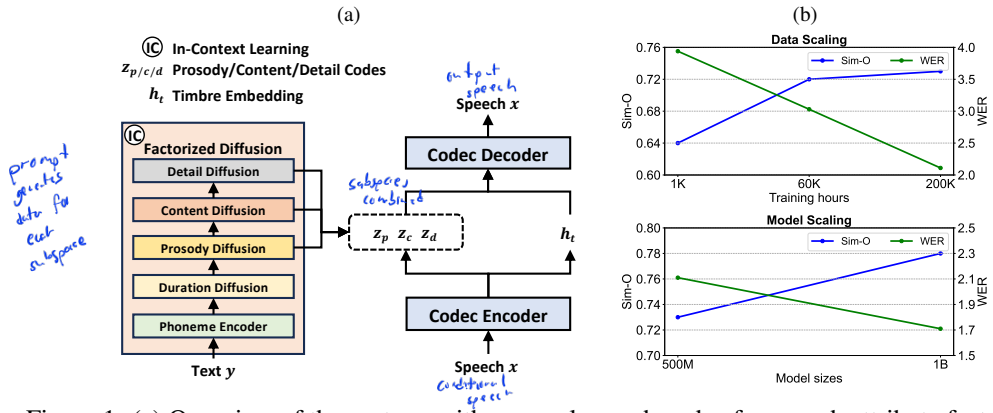


Figure 1: (a) Overview of the system, with a neural speech codec for speech attribute factorization and a factorized diffusion model. (b) Data and model scaling of the system.

*The first four authors contributed equally to this work, and their names are listed in random order. Corresponding author: Xu Tan, xuta@microsoft.com

Preprint. Work in progress.

1 Introduction

In recent years, significant advancements have been achieved in text-to-speech (TTS) synthesis. Traditional TTS systems [1, 2, 3, 4] are typically trained on limited datasets recorded in studios, and thus fail to support high-quality zero-shot speech synthesis. Recent works [5, 6, 7] 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 [8, 9] and mel-spectrogram [1, 2, 10, 7, 11] as data representations suffer from these intricate complexities during speech generation. 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 [12, 13, 6] encode speech into multi-level discrete tokens using a neural audio codec [14, 15] 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 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 waveform into distinct subspaces of content, prosody, timbre, and acoustic details and reconstruct speech waveform with these disentangled representations, leveraging information bottleneck [16, 17], various supervised losses, and adversarial training [18] 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 our method, referred to 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[2]. Furthermore, we demonstrate the scalability of NaturalSpeech 3 by scaling it to 1B parameters and 200K hours of training data. 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 [6, 19, 20]; 2) Discrete Tokens + Non-

[2]While NaturalSpeech 1 [4] 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.

autoregressive [13, 21, 22]; 3) Continuous Vectors + Autoregressive [23]; 4) Continuous Vectors + Non-autoregressive [5, 11, 24, 25]. Discrete tokens are typically derived from neural codec, while continuous vectors are generally obtained from mel-spectrogram or latents from audio autoencoder or 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 representation such as raw waveform [26, 27, 28] or mel-spectrogram [29, 30, 3, 31]. Recently, large-scale TTS systems [6, 13, 5] leverage data-driven representation, i.e., either discrete tokens or continuous vectors form an auto-encoder [14, 15, 32]. 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 [3, 33, 34, 7, 5, 11] 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 [2, 30, 6, 23, 35] have better diversity, prosody, expressiveness, and flexibility than NAR-based models, due to their implicitly 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 [36, 37, 38] utilize disentangled representation for speech generation, such as speech content from self-supervised pre-trained models [39, 40, 41], fundamental frequency, and timbre, but speech quality is not satisfying. Recently, some works explore attribute disentanglement in neural speech codec. SpeechTokenizer [42] uses HuBERT [43] for semantic distillation, aiming to render the first-layer RVQ representation as semantic information. Disen-TF-Codec [44] proposes the disentanglement with content and timbre representation, and applies them for zero-shot voice conversion. In this paper, we achieve better disentanglement with more speech attributes including content, prosody, acoustic details and timbre while ensuring high-quality reconstruction. We validate such disentanglement can bring about significant improvements in zero-shot TTS task.

3 NaturalSpeech 3

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 (i.e., 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.

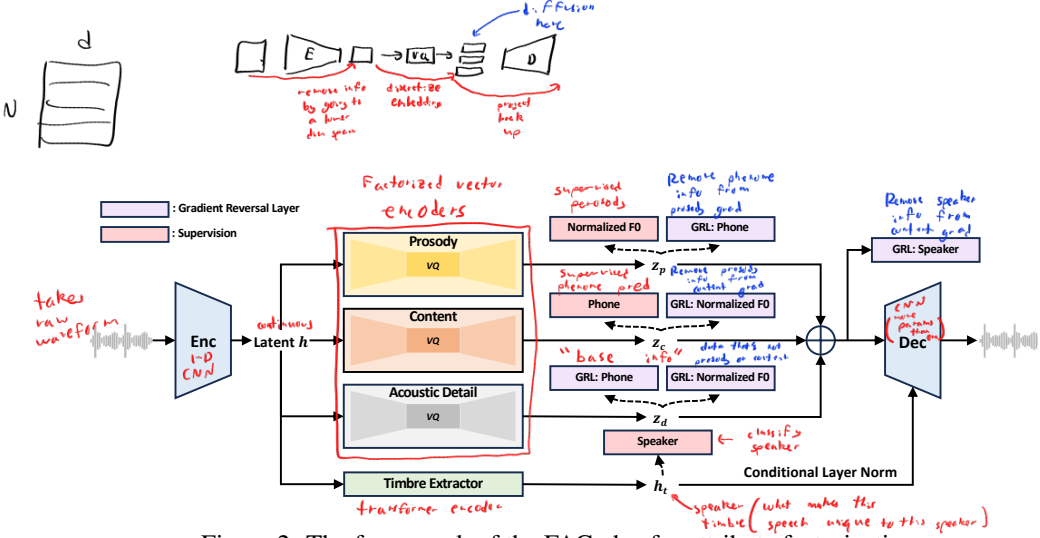


Figure 2: The framework of the FACodec for attribute factorization.

3.2 FACodec for Attribute Factorization

3.2.1 FACodec Model Overview

We propose a factorized neural speech codec (i.e., FACodec[3]) 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.

As shown in Figure 2, our 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 [14, 5], 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 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; 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 [45] 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 [16, 17], 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 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 [32, 46]. After quantization, we will project the quantized vector back to original 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 [37], 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 [47], we adopt adversarial classifier with the gradient reversal layer

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

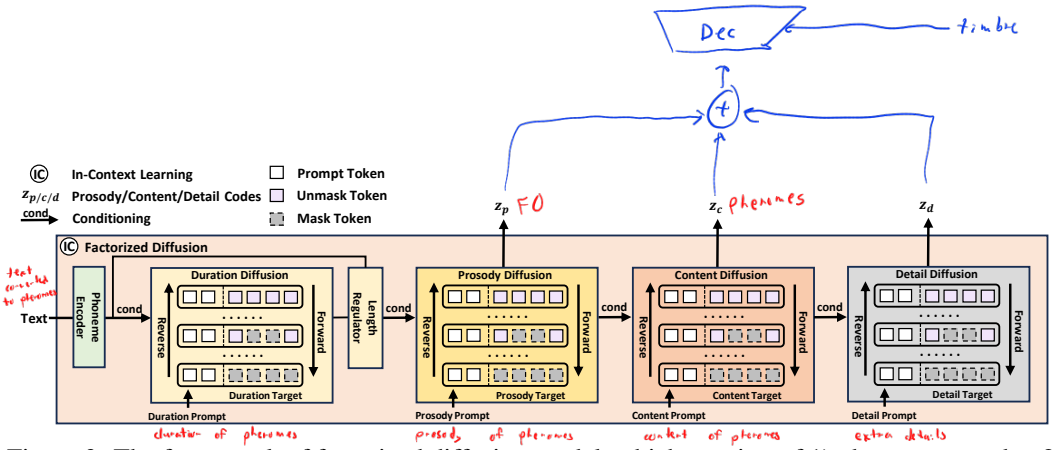


Figure 3: The framework of factorized diffusion model, which consists of 1) phoneme encoder, 2) duration diffusion and length regulator, 3) prosody diffusion, 4) content diffusion, 5) detail (acoustic detail) diffusion. Note that modules 2-5 share the same diffusion formulation.

(GRL) [48] to eliminate undesired information in latent space. Specifically, for prosody, we apply phoneme-GRL (i.e., GRL layer by predicting phoneme labels) to eliminate 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 eliminate both content and prosody information. In addition, we apply speaker-GRL on the sum of z_p, z_c, z_d to eliminate timbre.

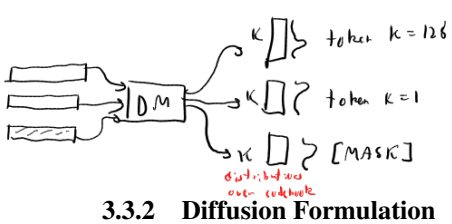
Detail Dropout. We have the following considerations: 1) empirically, we find that the codec tends to preserve undesired information (e.g., content, prosody) in acoustic details subspace since there is no supervision; 2) intuitively, without acoustic details, the decoder should reconstruct speech only 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.

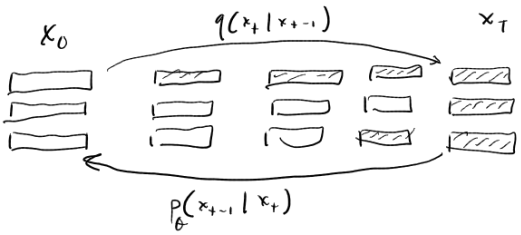
3.3 Factorized Diffusion Model

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 [49, 13]. 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 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 conditions; 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. Specifically, we do not explicitly generate the timbre attribute. Due to the factorization design in our FACodec, we can obtain timbre from the prompt directly and do not need to generate it. 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

Forward Process-
- Randomly mask tokens which the model learns to directly predict in one shot

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-
- Diffusion model is trained to predict if a token is masked or not.
- Obviously at t=0, no tokens are masked and at t=T all tokens are masked
- Model conditioned on $\mathbf{X}^p$ - prompt token sequence
  C - some condition like previous data such as phone

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:

train model to predict the codebook index of masked tokens

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

mask of token i; log probs of token i

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-
1. Start with a fully masked sequence
2. Get preds for all tokens in sequence
3. Remask tokens with lowest confidence score
4. Repeat from 2 till satisfied

w "time step"

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 [50, 51, 52], 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.

pretty standard CFG

Classifier-free Guidance. Moreover, we adapt the classifier-free guidance technique [53, 54]. Specifically, in training, we do not use the prompt with a probability of $p_{\text{cfg}} = 0.15$. In inference, we extrapolate the model output 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 [55].

3.4 Connections to the NaturalSpeech Series

NaturalSpeech 3 is an advanced TTS system of the NaturalSpeech series. Compared with the previous versions NaturalSpeech [4] and NaturalSpeech 2 [5], NaturalSpeech 3 has the following connections and distinctions:

- *Goal*. The NaturalSpeech series aims to generate natural speech with high quality and diversity. We approach this goal in several stages: 1) Achieving high-quality speech synthesis in single-speaker scenarios. To this end, NaturalSpeech [4] generates speech with quality on par with human recordings and only tackles single-speaker recording-studio datasets (e.g., LJSpeech). 2) Achieving high-quality and diverse speech synthesis on multi-style, multi-speaker, and multi-lingual scenarios. NaturalSpeech 2 [5] firstly focuses on speech diversity by exploring the zero-shot synthesis ability based on large-scale, multi-speaker, and in-the-wild datasets. Furthermore, NaturalSpeech 3 further achieves human-level naturalness on the multi-speaker dataset (e.g., LibriSpeech).
- *Architecture.* The NaturalSpeech series shares the basic components such as encoder/decoder for waveform reconstruction and duration prediction for non-autoregressive speech generation. Different from NaturalSpeech which utilizes flow-based generative models and NaturalSpeech 2 which leverages latent diffusion models, NaturalSpeech 3 proposes the concept of factorized diffusion models to generate each factorized speech attribute in a divide-and-conquer way.
- *Speech Representations.* Due to the complexity of speech waveform, the NaturalSpeech series uses an encoder/decoder to obtain speech latent for high-quality speech synthesis. NaturalSpeech utilizes naive VAE-based continuous representations, NaturalSpeech 2 leverages the continuous

representations from the neural audio codec with residual vector quantizers, while NaturalSpeech 3 proposes a novel FACodec to convert complex speech signal into disentangled subspaces (i.e., prosody, content, acoustic details, and timbre) and reduces the speech modeling complexity.

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.

Implementation Details. We use Librilight [56], which contains 60K hours of 16KHz unlabeled speech data and around 7000 distinct speakers from LibriVox audiobooks, as the training set. In duration diffusion, we further improve the performance by conditioning phoneme-level prosody codes. 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. Please refer to Appendix B.1 for details of the FACodec and Appendix A.2 for more details of our factorization diffusion model.

Evaluation Dataset. We employ two benchmark datasets: 1) LibriSpeech [57] test-clean, a widely-used testset for zero-shot TTS task. It contains 40 distinct speakers and 5.4-hour speech. Following [5], 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 [58], 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 speaker-similarity (SIM-O and SIM-R), speech quality (UTMOS), and robustness (WER). In specific, 1) for SIM-O and SIM-R, we employ the WavLM-TDCNN[4] 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 [59] which is a surrogate objective metric of MOS; 3) for Word Error Rate (WER), we use an ASR model[5] 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 [60][6]. In the RAVDESS benchmark, we evaluate the prosody similarity (MCD and MCD-Acc). In specific, 1) following [61], we adopt Mel-Ceptral Distortion (MCD) for prosody evaluation by measuring the differences between generated samples and ground truth samples. 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.

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

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

[6] `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). 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.

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

Evaluation Baselines. We compare NaturalSpeech 3 with baselines: 1) VALL-E [6]. 2) NaturalSpeech 2 [5]. 3) Voicebox [11]. 4) Mega-TTS 2 [62]. 5) UniAudio [35]. 6) StyleTTS 2 [24]. 7) HierSpeech++ [25]. Please refer to Appendix A.3 for details.

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 our method 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.

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.

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**

4.2.3 Robustness

We assess the robustness of our zero-shot TTS by measuring the word error rate of generated speech on the LibriSpeech test-clean benchmark. The results in Table 1 indicate that 1) NaturalSpeech 3 achieves a better WER than the ground truth, 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 9 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 [4] has achieved human-level quality on the single-speaker LJSpeech dataset.

4.3 Ablation Study and Method Analyses

4.3.1 Ablation Study

In this subsection, we conduct ablation studies to verify the effectiveness of 1) factorization; 2) classier-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. 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.

Classier-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.

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

Table 4: 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

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 [7, 63, 64]) 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 4 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).

Table 5: 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. Abbreviation: H (Hop Size), N (Codebook Number).

Models	Sampling Rate	H	N	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**

better than Encodec on speech

4.3.2 Method Analyses

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

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, allowing for a fair comparison. We name it VALL-E + FACodec. As shown in Table 6, VALL-E + FACodec 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.

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

Table 6: The comparison between autoregressive approach with (VALL-E + FACodec) 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 + FACodec	**0.57 / 0.65**	**5.60**	**+0.24**	**3.61**
VALL-E♦	0.47 / 0.51	6.11	0.00	3.46

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[7].

4.3.3 Experimental Results on FACodec

We compare the proposed FACodec in terms of the reconstruction quality with strong baselines, such as EnCodec [15], HiFi-Codec [65], Descript-Audio-Codec (DAC) [32], and our reproduced SoundStream [14]. Table 5 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.

4.4 Effectiveness of Data and Model Scaling

In this section, we study the effectiveness of data and model scaling on the proposed factorized diffusion model. We evaluate the zero-shot TTS performance in terms of speaker similarity (Sim-O) and robustness (WER) on an internal test set consisting of 30 audio clips.

Data Scaling. With a fixed model size of 500M parameters, we train NaturalSpeech 3, including both FACodec and the factorized diffusion model, on three datasets: 1) a 1K-hour subset randomly drawn from the Librilight dataset, 2) a 60K-hour Librilight dataset, and 3) an internal dataset with 200K hours of speech. In Table 7, we observe that: 1) even with a mere 1K hours of speech data, our model attains a Sim-O score of 0.64 and a WER of 3.94. It shows that with the speech factorization, NaturalSpeech 3 can generate the speech effectively. 2) As we scale up training data from 1K hours to 60K hours, and then to 200K hours, NaturalSpeech 3 displays continuously enhanced performance, with an improvement of 0.08 and 0.09 in terms of Sim-O, and 0.91 and 1.83 in terms of WER, respectively, thus confirming the benefits of data scaling.

Model Scaling. We scale up the model size of the factorized diffusion model from 500M to 1B parameters with the internal 200K hours dataset. Specifically, we double the number of transformer layers from 12 to 24. The results in Table 8 show a boost in both speaker similarity (0.05 in Sim-O) and robustness (0.40 in WER), validating the effectiveness of model scaling. In the future, we will scale up the model size even larger to achieve better results.

5 Conclusion

In this paper, we develop a TTS system that consists of 1) a novel neural speech codec with factorized vector quantization (i.e., FACodec) 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-the-art TTS system on speech quality, similarity, prosody, and intelligibility. We also show

[7] `https://speechresearch.github.io/naturalspeech3`

Table 7: The performance of NaturalSpeech 3 on an internal test set, with 500M model size and different hours of training data.

	Sim-O↑	WER↓
1K	0.64	3.94
60K	0.72	3.03
200K	**0.73**	**2.11**

Table 8: The performance of NaturalSpeech 3 on an internal test set, with 200K hours of training data and different model sizes.

	Sim-O↑	WER↓
500M	0.73	2.11
1B	**0.78**	**1.71**

that NaturalSpeech 3 can enable speech attribute manipulation, by customizing speech attribute prompts. Furthermore, we demonstrate that NaturalSpeech 3 achieves human-level performance on the multi-speaker LibriSpeech dataset for the first time and better performance by scaling to 1B parameters and 200K hours of training data. We list the limitations and future works in Appendix C.

6 Boarder Impact

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

[1] 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. *Proc. Interspeech 2017*, pages 4006–4010, 2017.

[2] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[3] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, robust and controllable text to speech. In *NeurIPS*, 2019.

[4] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

[6] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[7] Ziyue Jiang, Yi Ren, Zhenhui Ye, Jinglin Liu, Chen Zhang, Qian Yang, Shengpeng Ji, Rongjie Huang, Chunfeng Wang, Xiang Yin, et al. Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias. *arXiv preprint arXiv:2306.03509*, 2023.

[8] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *arXiv preprint arXiv:2106.06103*, 2021.

[9] Dan Lim, Sunghee Jung, and Eesung Kim. Jets: Jointly training fastspeech2 and hifi-gan for end to end text to speech. *arXiv preprint arXiv:2203.16852*, 2022.

[10] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-TTS: A diffusion probabilistic model for text-to-speech. *arXiv preprint arXiv:2105.06337*, 2021.

[11] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *arXiv preprint arXiv:2306.15687*, 2023.

[12] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*, 2022.

[13] Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023.

[14] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[15] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[16] 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*, pages 7836–7846. PMLR, 2020.

[17] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. AutoVC: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, pages 5210–5219. PMLR, 2019.

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

[19] Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*, 2023.

[20] Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Luping Liu, Zhenhui Ye, Ziyue Jiang, Chao Weng, Zhou Zhao, and Dong Yu. Make-a-voice: Unified voice synthesis with discrete representation. *arXiv preprint arXiv:2305.19269*, 2023.

[21] Dongchao Yang, Songxiang Liu, Rongjie Huang, Guangzhi Lei, Chao Weng, Helen Meng, and Dong Yu. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *arXiv preprint arXiv:2301.13662*, 2023.

[22] Chenpeng Du, Yiwei Guo, Feiyu Shen, Zhijun Liu, Zheng Liang, Xie Chen, Shuai Wang, Hui Zhang, and Kai Yu. Unicats: A unified context-aware text-to-speech framework with contextual vq-diffusion and vocoding. *arXiv preprint arXiv:2306.07547*, 2023.

[23] Eliya Nachmani, Alon Levkovitch, Julian Salazar, Chulayutsh Asawaroengchai, Soroosh Mariooryad, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Lms with a voice: Spoken language modeling beyond speech tokens. *arXiv preprint arXiv:2305.15255*, 2023.

[24] Yinghao Aaron Li, Cong Han, Vinay S Raghavan, Gavin Mischler, and Nima Mesgarani. 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.

[25] Sang-Hoon Lee, Ha-Yeong Choi, Seung-Bin Kim, and Seong-Whan Lee. 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.

[26] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[27] Aäron van den Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, Florian Stimberg, et al. Parallel WaveNet: Fast high-fidelity speech synthesis. In *International conference on machine learning*, pages 3918–3926. PMLR, 2018.

[28] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aäron Courville, and Yoshua Bengio. Char2wav: End-to-end speech synthesis. 2017.

[29] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep Voice 3: 2000-speaker neural text-to-speech. *Proc. ICLR*, pages 214–217, 2018.

[30] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with Transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6706–6713, 2019.

[31] 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, 2020.

[32] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *arXiv preprint arXiv:2306.06546*, 2023.

[33] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron Weiss, and Yonghui Wu. Parallel Tacotron: Non-autoregressive and controllable TTS. *arXiv preprint arXiv:2010.11439*, 2020.

[34] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. DiffSinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11020–11028, 2022.

[35] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.

[36] 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.

[37] Hyeong-Seok Choi, Jinhyeok Yang, Juheon Lee, and Hyeongju Kim. Nansy++: Unified voice synthesis with neural analysis and synthesis. *arXiv preprint arXiv:2211.09407*, 2022.

[38] Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*, 2021.

[39] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 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)*, pages 244–250. IEEE, 2021.

[40] 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.

[41] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *Proc. Interspeech 2019*, pages 3465–3469, 2019.

[42] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023.

[43] 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.

[44] Xue Jiang, Xiulian Peng, Yuan Zhang, and Yan Lu. Disentangled feature learning for real-time neural speech coding. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[45] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, sheng zhao, and Tie-Yan Liu. AdaSpeech: Adaptive text to speech for custom voice. In *International Conference on Learning Representations*, 2021.

[46] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021.

[47] SiCheng Yang, Methawee Tantrawenith, Haolin Zhuang, Zhiyong Wu, Aolan Sun, Jianzong Wang, Ning Cheng, Huaizhen Tang, Xintao Zhao, Jie Wang, et al. Speech representation disentanglement with adversarial mutual information learning for one-shot voice conversion. *arXiv preprint arXiv:2208.08757*, 2022.

[48] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[49] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

[50] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

[51] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022.

[52] José Lezama, Huiwen Chang, Lu Jiang, and Irfan Essa. Improved masked image generation with token-critic. In *European Conference on Computer Vision*, pages 70–86. Springer, 2022.

[53] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[54] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[55] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5404–5411, 2024.

[56] Jacob Kahn, Morgane Riviere, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, 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)*, pages 7669–7673. IEEE, 2020.

[57] 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)*, pages 5206–5210. IEEE, 2015.

[58] Steven R Livingstone and Frank A Russo. 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.

[59] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022.

[60] Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, 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)*, pages 6381–6385. IEEE, 2019.

[61] Mohammed Salah Al-Radhi, Tamás Gábor Csapó, and Géza Németh. 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)*, pages 176–181. IEEE, 2023.

[62] Ziyue Jiang, Jinglin Liu, Yi Ren, Jinzheng He, Chen Zhang, Zhenhui Ye, Pengfei Wei, Chunfeng Wang, Xiang Yin, Zejun Ma, et al. Mega-tts 2: Zero-shot text-to-speech with arbitrary length speech prompts. *arXiv preprint arXiv:2307.07218*, 2023.

[63] Hyung-Seok Oh, Sang-Hoon Lee, and Seong-Whan Lee. Diffprosody: Diffusion-based latent prosody generation for expressive speech synthesis with prosody conditional adversarial training. *arXiv preprint arXiv:2307.16549*, 2023.

[64] 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)*, pages 7577–7581. IEEE, 2022.

[65] Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023.

[66] 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. In *INTERSPEECH*, 2019.

[67] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022.

[68] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*, 2022.

[69] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[70] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6309–6318, 2017.

[71] 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*, pages 2709–2720. PMLR, 2022.

[72] Zhichao Wang, Yuanzhe Chen, Lei Xie, Qiao Tian, and Yuping Wang. Lm-vc: Zero-shot voice conversion via speech generation based on language models. *arXiv preprint arXiv:2306.10521*, 2023.

A Details of Factorization Diffusion Model

A.1 Model Configuration

The phoneme encoder uses a similar architecture as [5] 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 shared 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. 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

We use Librilight [56], 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 [66], 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. 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 [67], 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 [6]. 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 [5]. It use a non-autoregressive model for continuous vectors generation. We obtain samples through communication with the authors.
- Voicebox [11]. 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.
- Mega-TTS 2 [62]. It use a non-autoregressive model for continuous vectors generation. We obtain samples through communication with the authors.
- UniAudio [35]. It use an autoregressive model for discrete token generation. We obtain samples through communication with the authors.
- StyleTTS 2 [24]. It use a non-autoregressive model for continuous vectors generation. We use official code and checkpoint[8].
- HierSpeech++ [25]. It use a non-autoregressive model for continuous vectors generation. We use official code and checkpoint[9]. We do not use its super resolution model for fair comparison.

[8] `https://github.com/yl4579/StyleTTS2`

[9] `https://github.com/sh-lee-prml/HierSpeechpp`

Table 9: 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.

	Sim-O ↑	Sim-R ↑	WER↓	WER* ↓	UTMOS ↑	CMOS↑	SMOS↑
Ground Truth	0.68	-	1.94	0.68	4.14	+0.08	3.85
VALL-E ♥	-	0.58	5.90	-	-	-	-
VALL-E ♦	0.47	0.51	6.11	4.87	3.68	-0.60	3.46
NaturalSpeech 2♠	0.55	0.62	1.94	1.24	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	1.24	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	1.58	3.94	-0.21	3.07
HierSpeech++♣	0.51	-	6.33	4.97	3.80	-0.41	3.50
NaturalSpeech 3	**0.67**	**0.76**	**1.81**	**1.13**	**4.30**	**0.00**	**4.01**

Table 10: 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

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[10]; 2) UTMOS [59], which is a surrogate objective metric of MOS. The results are shown in Table 9.

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 [59] [11], 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 10. 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.

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

[10] `https://huggingface.co/nvidia/stt_en_conformer_transducer_xlarge`

[11] `https://github.com/tarepan/SpeechMOS`

Table 11: 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

Table 12: 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

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 11, 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 12, 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 FACodec encoder and decoder follows [32] and employs the SnakeBeta activation function [68]. The timbre extractor consists of several conformer [69] 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.

$\mathcal{L}_{rec}$ - autoencoder recon loss (mel spectrogram)

$\mathcal{L}_{adv}$ - autoencoder adv loss (multiscale STFT discriminator)

$\mathcal{L}_{feat}$ - autoencoder waveform loss

$\mathcal{L}_{codebook}/\mathcal{L}_{commit}$ - losses for vector quantization

$\mathcal{L}_{ph}$ - phoneme supervision

$\mathcal{L}_{f0}$ - prosody supervision

$\mathcal{L}_{gr\text{-}ph}$ - phoneme grad reversal loss

$\mathcal{L}_{gr\text{-}f0}$ - prosody grad reversal loss

$\mathcal{L}_{gr\text{-}spk}$ - speaker grad reversal loss

Loss Functions. We utilize the multi-scale mel-reconstruction loss $\mathcal{L}_{\text{rec}}$ as detailed in [32]. 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 [32]. 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 [70]. 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 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 [15][12], HiFi-Codec [65][13], and Descript-Audio-Codec (DAC) [32][14]. We additionally reproduce SoundStream [14] following the original paper's implementation and experimental setup. Table 13 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.

Table 13: 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. Abbreviation: H (Hop Size), N (Codebook Number).

Models	Sampling Rate	H	N	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	<u>2.65</u>
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	<u>3.47</u>	**0.95**	<u>0.93</u>	**2.59**

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

[12] https://github.com/facebookresearch/encodec

[13] https://github.com/yangdongchao/AcademiCodec

[14] https://github.com/descriptinc/descript-audio-codec

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 [71], Make-A-Voice (VC) [20], LM-VC [72], and UniAudio [35]. We use VCTK dataset for comparison. We use Sim-O[15] to compare speaker similarity to baselines and WER to evaluate speech quality. Table 14 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 14: 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 our FACodec; 2) the effect of gradient reversal on the disentanglement of our 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 15 demonstrates that without the information bottleneck, the speaker similarity of zero-shot voice conversion decreases by 0.13.

Table 15: 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

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.

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

Although content, prosody, and timbre information already encompass the majority of speech information, Table 16 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 16: 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.