
A Variational Framework for Improving Naturalness in Generative Spoken Language Models

Li-Wei Chen¹ Takuya Higuchi² Zakaria Aldeneh² Ahmed Hussen Abdelaziz² Alexander Rudnicky¹

Abstract

The success of large language models in text processing has inspired their adaptation to speech modeling. However, since speech is continuous and complex, it is often discretized for autoregressive modeling. Speech tokens derived from self-supervised models (known as semantic tokens) typically focus on the linguistic aspects of speech but neglect prosodic information. As a result, models trained on these tokens can generate speech with reduced naturalness. Existing approaches try to fix this by adding pitch features to the semantic tokens. However, pitch alone cannot fully represent the range of paralinguistic attributes, and selecting the right features requires careful hand-engineering. To overcome this, we propose an end-to-end variational approach that automatically learns to encode these continuous speech attributes to enhance the semantic tokens. Our approach eliminates the need for manual extraction and selection of paralinguistic features. Moreover, it produces preferred speech continuations according to human raters. Code, samples and models are available at <https://github.com/b04901014/vae-gslm>.

1. Introduction

Large language models (LLMs) have achieved tremendous success in text processing (OpenAI, 2024), offering new ways to interact with machines. This progress has motivated efforts to extend their capabilities to speech to enable more natural spoken interactions with machines. However, modeling speech presents unique challenges due to its continuous

and complex nature. As a result, previous works (Lakhotia et al., 2021; Borsos et al., 2023; Maiti et al., 2024) tokenized speech into simpler discrete units to enable the application of language modeling techniques originally developed for text. However, these *semantic tokens* are typically derived by performing k -means clustering on features extracted from self-supervised pre-trained speech models, such as HuBERT (Hsu et al., 2021). We use the term *semantic tokens* to distinguish them from acoustic tokens (Borsos et al., 2023), which capture general acoustic information. These models primarily capture the linguistic aspects of speech, such as phonetic information, while often overlooking paralinguistic features, such as prosody (Weston et al., 2021). As a result, training an autoregressive model solely with such semantic tokens restricts the model’s ability to fully capture and represent the diverse information encoded in speech.

To address the aforementioned limitation, Kharitonov et al. (2022) augmented the tokens with extracted fundamental frequency (F_0 , or pitch) to enable prosody-aware modeling. However, augmenting semantic tokens with manually defined paralinguistic attributes can be inherently suboptimal. First, pitch alone cannot capture the full range of paralinguistic information encoded in speech. For instance, energy-related (e.g., loudness, zero-crossing-rate) and spectral-related (e.g., mel-frequency cepstral coefficients) features are also important paralinguistic features (Schuller et al., 2009; 2013; Eyben et al., 2015). Furthermore, training a correct pitch tracker introduces additional complexity (Kim et al., 2018).

Instead of relying on hand-engineered paralinguistic features, we propose an approach to learning these features directly from the input signal, within an autoregressive framework. These learned features are optimized to both: 1) reconstruct the input speech, and 2) enhance the autoregressive modeling process. Our approach allows the learned features to complement semantic tokens, removing the need for pre-extracted paralinguistic features as required in previous methods. As a result, our method generates more natural-sounding speech compared to baseline models while maintaining comparable meaningfulness of the syntheses.

¹Language Technology Institute, Carnegie Mellon University ²Apple. Correspondence to: Li-Wei Chen <liweiche@cs.cmu.edu>, Takuya Higuchi <takuya_higuchi@apple.com>, Zakaria Aldeneh <zaldeneh@apple.com>, Ahmed Hussen Abdelaziz <ahussenabdelaziz@apple.com>, Alexander Rudnicky <air@cs.cmu.edu>.

2. Preliminaries

In this work, we work on mel-spectrogram, and consider vocoding, the act of turning mel-spectrogram back to raw waveform, as a problem that has already been addressed. We denote the mel-spectrogram as $\mathbf{X} = (x_t \in \mathbb{R}^{d_x})_{t=1}^T$, where d_x represents the number of filter-banks, T is the total number of time frames in the spectrogram, and x_t is the frame at time t . We use $\mathbf{X}_{i:j}$ to denote the sub-sequence $(x_t)_{t=i}^j$, and define $\mathbf{X}_{1:0} = \emptyset$. Our goal is to model $p(\mathbf{X})$ using a generative approach.

Token-based Speech Language Model We describe the general framework of speech language models that rely on the use of semantic tokens, as seen in works like Lakhota et al. (2021); Borsos et al. (2023); Maiti et al. (2024). This approach consists of three components: a speech tokenizer, an autoregressive model, and a decoder. The speech tokenizer maps \mathbf{X}^1 to a sequence of discrete semantic tokens $\mathbf{Z}^d = (z_t^d \in \mathbb{N}_k)_{t=1}^T$, where $\mathbb{N}_k = \{1, 2, \dots, k\}$, and k is the vocabulary size of the semantic tokens. We use $p(\mathbf{Z}^d | \mathbf{X})$ to denote the implicit distribution of the pre-trained speech tokenizer. The autoregressive model, parameterized by ψ , models the probability of token sequences \mathbf{Z}^d as $p_\psi(\mathbf{Z}^d) = \prod_{t=1}^T p_\psi(z_t^d | \mathbf{Z}_{1:t-1}^d)$. Finally, the decoder, parameterized by θ , is trained to convert \mathbf{Z}^d back to \mathbf{X} by modeling $p_\theta(\mathbf{X} | \mathbf{Z}^d)$. However, this framework is limited to semantic tokens \mathbf{Z}^d , which primarily capture linguistic information and ignore paralinguistic information. As a result, the decoder θ may struggle with accurate reconstruction, and the autoregressive model ψ can have difficulty incorporating paralinguistic information. To address this limitation, we propose to incorporate the variational autoencoder framework to learn continuous features to complement \mathbf{Z}^d .

Variational Autoencoder (VAE) Latent variable models introduce unobserved latent variables $\mathbf{Z}^c = (z_t^c \in \mathbb{R}^{d_z^c})_{t=1}^T$ that influence the observed variable \mathbf{X} . d_z^c is the dimension of each z_t^c , and is a hyper-parameter chosen prior to training. In a VAE, the likelihood of the observed data given the latent variable, $p_\theta(\mathbf{X} | \mathbf{Z}^c)$, is modeled by a neural decoder, parameterized by θ . The variational posterior, $q_\phi(\mathbf{Z}^c | \mathbf{X})$, is modeled by a neural encoder, parameterized by ϕ . Using this modeling setup, the log-likelihood of the data, $\log p_\theta(\mathbf{X})$, can be written as:

$$\underbrace{\mathbb{E}_{q_\phi(\mathbf{Z}^c | \mathbf{X})}[\log p_\theta(\mathbf{X} | \mathbf{Z}^c)] - D_{KL}(q_\phi(\mathbf{Z}^c | \mathbf{X}) || p(\mathbf{Z}^c))}_{\mathcal{O}_{ELBO}} + D_{KL}(q_\phi(\mathbf{Z}^c | \mathbf{X}) || p_\theta(\mathbf{Z}^c | \mathbf{X})), \quad (1)$$

¹Speech tokenizers can operate on mel-spectrograms or directly on raw waveforms.

where D_{KL} is the Kullback–Leibler (KL) divergence between two distributions, and $p(\mathbf{Z}^c)$ is a fixed prior distribution (usually a Gaussian). In Equation 1, \mathcal{O}_{ELBO} is known as the evidence lower bound (ELBO), which provides a lower bound for $\log p_\theta(\mathbf{X})$ since $D_{KL}(q_\phi(\mathbf{Z}^c | \mathbf{X}) || p_\theta(\mathbf{Z}^c | \mathbf{X}))$ is always nonnegative. Therefore, instead of directly optimizing $\mathbb{E}_{\mathbf{X}}[\log p_\theta(\mathbf{X})]$, the VAE maximizes the tractable lower bound $\mathbb{E}_{\mathbf{X}}[\mathcal{O}_{ELBO}]$. Here, we refer to the learned continuous latent \mathbf{Z}^c from the VAE as the *variational features*.

3. Proposed Framework

Figure 1 provides an overview of our proposed framework. This section is organized as follows: Section 3.1 introduces our setup that combines a VAE with an autoregressive model for the latent variables. Section 3.2 describes how we integrate semantic tokens into the framework. Section 3.3 discusses how to balance the different loss terms that arise in our setup. Section 3.4 describes the use of normalizing flows to improve the expressive power of the autoregressive prior. Finally, Section 3.5 introduces the diffusion decoder and the utterance encoder used in the framework.

3.1. VAE with an Autoregressive Prior

Our method starts by modeling the prior of the VAE, which is typically a fixed Gaussian distribution, with a trainable autoregressive model $p_\psi(\mathbf{Z}^c) = \prod_{t=1}^T p_\psi(z_t^c | \mathbf{Z}_{1:t-1}^c)$. We refer to this framework as *VAE with an autoregressive prior*. We note that VAE with an autoregressive prior has been explored in previous works (Vahdat & Kautz, 2020; Zhu et al., 2020) within the computer vision domain. Additionally, Sun et al. (2020) also applied a similar framework for TTS, but with prior and posterior distributions optimized separately instead of jointly. Here, we adopt the VAE framework with an autoregressive prior for speech continuation and further integrate it with discrete token-based models to enhance the naturalness of the synthesis. We use a diagonal Gaussian distribution to model the variational posterior, where the statistics are predicted by a neural network:

$$q_\phi(z_t^c | \mathbf{X}) = \mathcal{N}(z_t^c, \mu_\phi(\mathbf{X}, t), \sigma_\phi(\mathbf{X}, t)). \quad (2)$$

Since each z_t^c is conditionally independent given \mathbf{X} , we can express the posterior as: $q_\phi(\mathbf{Z}^c | \mathbf{X}) = \prod_{t=1}^T q_\phi(z_t^c | \mathbf{X})$. With this decomposition, and the parameterized autoregressive prior, the \mathcal{O}_{ELBO} in Equation 1 can be further derived²

²See Appendix A.1 for detailed derivation.

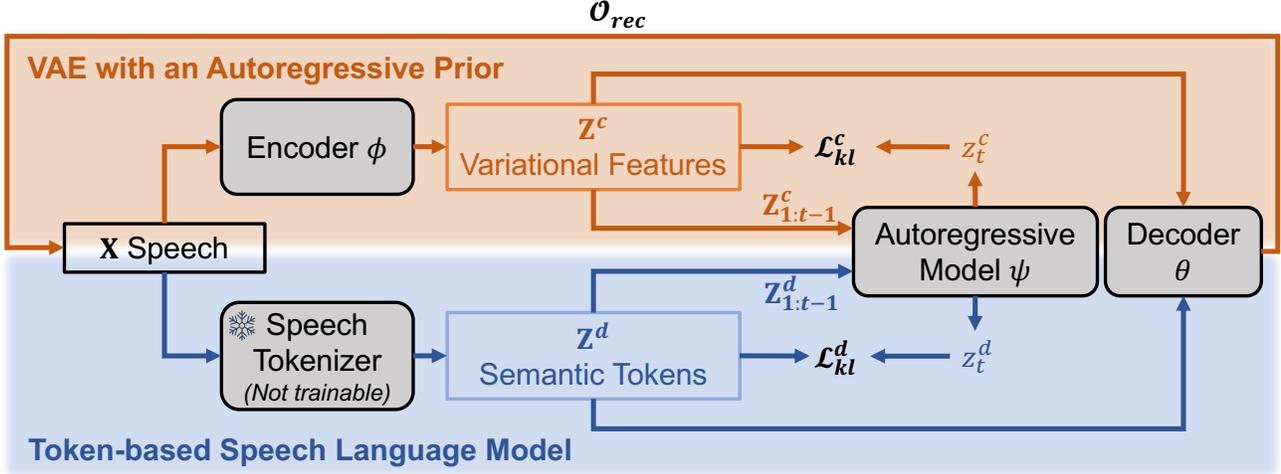


Figure 1. Overview of our proposed approach. Our method integrates the token-based speech language model (outlined in Section 2, represented by the lower shaded region) with a variational autoencoder (VAE with autoregressive prior, shown in the upper shaded region). This setup allows the model to learn variational features \mathbf{Z}^c that complement the pre-extracted semantic speech tokens \mathbf{Z}^d . In our proposed joint setup, the variational features \mathbf{Z}^c are trained to 1) reconstruct speech \mathbf{X} alongside \mathbf{Z}^d (by maximizing \mathcal{O}_{rec}); 2) facilitate the prediction of the next speech token z_t^d (by minimizing \mathcal{L}_{kl}^d); 3) support the sequential prediction of the variational features themselves (by minimizing \mathcal{L}_{kl}^c).

into:

$$\mathcal{O}_{ELBO} = \underbrace{\mathbb{E}_{\mathbf{Z}^c \sim q_\phi(\mathbf{Z}^c | \mathbf{X})} [\log p_\theta(\mathbf{X} | \mathbf{Z}^c)]}_{\mathcal{O}_{rec}} - \quad (3)$$

$$\underbrace{\sum_{t=1}^T \mathbb{E}_{\mathbf{Z}_{1:t-1}^c} [D_{KL}(q_\phi(z_t^c | \mathbf{X}) || p_\psi(z_t^c | \mathbf{Z}_{1:t-1}^c))]}_{\mathcal{L}_{kl}^c}.$$

By maximizing \mathcal{O}_{ELBO} , we maximize the first term, the reconstruction objective \mathcal{O}_{rec} , and minimize the second term, the variational feature prediction loss \mathcal{L}_{kl}^c . We note that training a model to maximize Equation 3 is feasible without incorporating discrete semantic tokens \mathbf{Z}^d . This token-free approach is also depicted as the upper shaded region in Figure 1 (VAE with an Autoregressive Prior), and its properties are further explored in Section 5.

3.2. Incorporating the Semantic Tokens with VAE

We now integrate the semantic tokens \mathbf{Z}^d with the VAE with an autoregressive prior. Using these tokens, the model no longer needs to encode as much phonetic information as in \mathbf{Z}^c , allowing \mathbf{Z}^c to focus on other attributes of continuous speech. To this end, we introduce a joint latent variable $\mathbf{Z} = (z_t \in \mathbb{R}^{d_z} \times \mathbb{N}_k)^T_{t=1}$, where z_t is the concatenation of z_t^c and z_t^d . Since \mathbf{Z}^d and \mathbf{Z}^c are conditional independent given \mathbf{X} , we can express the new variational posterior as: $q_\phi(\mathbf{Z} | \mathbf{X}) = q_\phi(\mathbf{Z}^c | \mathbf{X})p(\mathbf{Z}^d | \mathbf{X})$. Then, we model $p_\psi(z_t | \mathbf{Z}_{1:t-1}) = p_\psi(z_t^d | \mathbf{Z}_{1:t-1})p_\psi(z_t^c | \mathbf{Z}_{1:t-1})$, assum-

ing the conditional independence of z_t^d and z_t^c given the past generations. We further discuss this modeling assumption in Appendix I. This allows us to re-write³ \mathcal{O}_{ELBO} from Equation 1 as:

$$\mathcal{O}_{ELBO} = \underbrace{\mathbb{E}_{\mathbf{Z}^d \sim p(\mathbf{Z}^d | \mathbf{X}), \mathbf{Z}^c \sim q_\phi(\mathbf{Z}^c | \mathbf{X})} [\log p_\theta(\mathbf{X} | \mathbf{Z}^d, \mathbf{Z}^c)]}_{\mathcal{O}_{rec}} - \quad (4)$$

$$\underbrace{\sum_{t=1}^T \mathbb{E}_{\mathbf{Z}_{1:t-1}} [D_{KL}(q_\phi(z_t^c | \mathbf{X}) || p_\psi(z_t^c | \mathbf{Z}_{1:t-1}))]}_{\mathcal{L}_{kl}^c} -$$

$$\underbrace{\sum_{t=1}^T \mathbb{E}_{\mathbf{Z}_{1:t}} [-\log p_\psi(z_t^d | \mathbf{Z}_{1:t-1})]}_{\mathcal{L}_{kl}^d}.$$

From Equation 4, our training objective \mathcal{O}_{ELBO} consists of three terms: \mathcal{O}_{rec} , \mathcal{L}_{kl}^c , and \mathcal{L}_{kl}^d . \mathcal{O}_{rec} is the *reconstruction objective*. Maximizing \mathcal{O}_{rec} trains the decoder θ to reconstruct \mathbf{X} from both \mathbf{Z}^c and \mathbf{Z}^d , while encouraging the encoder ϕ to generate \mathbf{Z}^c with helpful information to reconstruct \mathbf{X} . \mathcal{L}_{kl}^c is the *variational feature prediction loss*. Minimizing \mathcal{L}_{kl}^c trains the autoregressive model ψ to predict the next variational feature z_t^c and encourages the encoder ϕ to generate \mathbf{Z}^c that is easier for ψ to model. \mathcal{L}_{kl}^d is the *semantic token prediction loss*, which trains the autoregres-

³See Appendix A.2 for detailed derivation.

sive model ψ to predict the next semantic token given the previous \mathbf{Z}^d and \mathbf{Z}^c .

3.3. Balancing the loss terms

In Equation 4, the terms \mathcal{O}_{rec} , \mathcal{L}_{kl}^c , and \mathcal{L}_{kl}^d can work against each other. For instance, the encoder ϕ optimizes both \mathcal{O}_{rec} and \mathcal{L}_{kl}^c . Maximizing \mathcal{O}_{rec} encourages the variational features \mathbf{Z}^c to encode more information about \mathbf{X} , while minimizing \mathcal{L}_{kl}^c regularize \mathbf{Z}^c to be simpler for the autoregressive model ψ to predict. Similarly, optimizing \mathcal{L}_{kl}^c and \mathcal{L}_{kl}^d with the autoregressive model ψ is a multi-task learning scenario, where ψ learns to predict two different objectives given the same input. Moreover, these terms may operate on different scales due to how the losses are computed, necessitating a balancing mechanism. As a result, inspired by β -VAE (Higgins et al., 2017), we introduce two scalars: β and γ , to balance the loss terms as follows:

$$\mathcal{O}_{ELBO} = \mathcal{O}_{rec} - \beta \cdot \mathcal{L}_{kl}^c - \gamma \cdot \mathcal{L}_{kl}^d. \quad (5)$$

Here, a larger β favors a simple $p(\mathbf{Z}^c)$, while a smaller β encourages the variational features \mathbf{Z}^c to encode more information about \mathbf{X} . Larger γ encourages the autoregressive model ψ to prioritize accurate predictions of \mathbf{Z}^d over \mathbf{Z}^c . In practice, we employ a linear warm-up strategy for β , increasing it from zero to its final value during the early stages of training. This approach, inspired by prior works on text generation (Bowman et al., 2016; Fu et al., 2019), helps mitigate posterior collapse. Empirically, we find that this strategy allows for higher values of β without causing \mathcal{L}_{kl}^c to collapse to zero.

3.4. Time-wise Normalizing Flow

We employ a lightweight normalizing flow (Rezende & Mohamed, 2015) that is shared across time to improve the expressive power of the autoregressive prior $p_\psi(z_t^c | \mathbf{Z}_{1:t-1})$. Specifically, an invertible flow network f_ψ maps each z_t to a point in the Gaussian distribution, and sampling can be realized by running the network in reverse. By using the change of variables, we can write:

$$p_\psi(z_t^c | \mathbf{Z}_{1:t-1}) = \mathcal{N}(f_\psi(z_t^c), \mu_\psi(\mathbf{Z}_{1:t-1}), \sigma_\psi(\mathbf{Z}_{1:t-1})) \left| \det \frac{\partial f_\psi(z_t^c)}{\partial z_t^c} \right|, \quad (6)$$

where μ_ψ, σ_ψ are modeled by autoregressive neural networks (i.e., transformer). We choose affine coupling layers (Dinh et al., 2017) as the backbone of our normalizing flow due to their simple implementation and efficient computation. We note that similar approaches using normalizing flows to enhance prior distributions have also been observed in Kim et al. (2021; 2020) for text-to-speech.

3.5. Other Components

We describe the modeling of the our decoder $p_\theta(\mathbf{X} | \mathbf{Z})$ and the utterance encoder designed to capture static information. While these components are not the main focus of our study, they help ensure a fair comparison between different methods. We use these components for all methods in our experiments and focus on how changing the inputs to the autoregressive model affects performance.

Diffusion Decoder We model the decoder $p_\theta(\mathbf{X} | \mathbf{Z})$ with Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020). We choose DDPM due to its flexibility in modeling complex distributions. We condition the diffusion process on \mathbf{Z} . For back-propagation through the encoder ϕ , we use the reparameterization trick (Kingma & Welling, 2019) to sample from $q_\phi(\mathbf{Z}^c | \mathbf{X})$, and combine it with embedded semantic tokens \mathbf{Z}^d . The outcome is then concatenated with each intermediate layer of the diffusion decoder for conditional diffusion. We train all diffusion decoders with 1000 DDPM steps. Note that our proposed approach is not limited to a specific decoder. Although we opted for a diffusion-based decoder for ease of training, our method is compatible with various decoding strategies. There are no constraints on the type of decoder used to parameterize $p_\theta(\mathbf{X} | \mathbf{Z}^d, \mathbf{Z}^c)$.

Utterance Encoder Static features, such as speaker information and recording environments, often vary little across a given utterance. In our current modeling approach, this static information would be redundantly encoded at each time step. To address this issue, we introduce an additional utterance-level feature encoder that encourages \mathbf{Z} to focus on time-varying signals. Specifically, we randomly segment a portion of the mel-spectrogram \mathbf{X} and feed it to the utterance encoder to produce an utterance-level embedding. This embedding is then concatenated with \mathbf{Z} before being provided to the diffusion decoder. The utterance encoder is trained end-to-end with the entire system.

4. Experimental Setup

4.1. Datasets

We use two datasets in our experiments: LibriSpeech (Panayotov et al., 2015) and Libri-light (Kahn et al., 2020), consisting of audiobooks narrated in English. LibriSpeech contains 960 hours of speech, while Libri-light contains 60k hours of speech. For semantic token extraction, we follow Hassid et al. (2023); Maiti et al. (2024) and use tokens derived from HuBERT representations (Hsu et al., 2021). We use the official HuBERT checkpoints,

pre-trained on LibriSpeech⁴ and Libri-light⁵. We run k -means clustering with $k = 200$ on the output of the last transformer layer of HuBERT using 10% of data randomly sampled from the training set. We pick $k = 200$ after testing values from $\{50, 200, 1000\}$ and choosing the one that produced the best language modeling performance. The result is also consistent with Maiti et al. (2024). More details on the choice of k are provided in Appendix F.

4.2. Methods

We compare our proposed approach to methods that use only semantic tokens in the autoregressive model, as well as methods that use semantic tokens with added pitch features in the autoregressive model. To ensure a fair comparison, we fix the autoregressive model architecture to be the same for all methods, varying only the input and output layers. We also use the same configuration for the diffusion decoder and utterance encoder across all methods.⁶ For the neural vocoder (i.e., mapping the mel-spectrogram back to waveform), we train HiFi-GAN (Kong et al., 2020) on LibriSpeech and use it for all of the methods. We leave the detailed configuration of model architectures in Appendix B. Below, we provide further details on the three approaches.

Token-LM We adopt the token-based speech language model (described in Section 2) as our baseline, representing approaches such as Lakhotia et al. (2021); Borsos et al. (2023); Maiti et al. (2024), which apply only discrete semantic tokens to the autoregressive model.

Token-LM + Pitch In this baseline approach, we augment the semantic tokens of token-based speech language model (described in Section 2) with log pitch features before passing them into the autoregressive model. The pitch features are extracted using CREPE (Kim et al., 2018). Additionally, we introduce a pitch regression task alongside the standard next-token prediction task, optimizing it with L1 loss. This method incorporates hand-engineered paralinguistic features, similar to the approach used by Kharitonov et al. (2022).

Token-LM + Acoustic In this comparison method, we augment semantic tokens with acoustic tokens (Borsos et al., 2023; Défossez et al., 2023). Specifically, we train a residual vector quantization (RVQ) autoencoder to discretize speech into four levels of acoustic tokens. At each transformer time step, the model first predicts the semantic token, followed by the acoustic tokens, which are autoregressively generated over the code levels using an additional transformer layer,

similar to Chen et al. (2023); Défossez et al. (2024). We include this baseline to compare with recent methods (Défossez et al., 2024) that integrate acoustic tokens into the autoregressive generation process.

Variational speech modeling approach (Proposed) This is our proposed approach introduced in Section 3. In this approach, we learn to extract variational features that supplement the semantic tokens while jointly training the autoregressive model. The learned variational features are used by both the autoregressive model and the decoder. This approach eliminates the need for the selection and extraction of paralinguistic features based on hand-made engineering. Additionally, we set our latent dimension $d_z^c = 4$. While we observed performance improvements with larger d_z^c , we opted for a smaller value to ensure a fairer comparison, as it results in less variation in parameter size. Our additional experiments on the latent dimension d_z^c is in Appendix E.

For inference, we use temperature-based sampling similar to Lakhotia et al. (2021). Specifically, we set the temperature to 0.85 for both semantic tokens \mathbf{Z}^d and continuous variational features \mathbf{Z}^c . For variational features, the temperature is the scalar multiplied to the standard deviation of the normal distribution in Equation 6 before sampling, as done in Kim et al. (2020). For the diffusion decoder, we use denoising diffusion implicit models (DDIM) from Song et al. (2021) with $\eta = 0.5$ and 100 diffusion steps. Training details are provided in Appendix C.

4.3. Evaluation Metrics

We evaluate the comparison methods on both reconstruction and speech continuation. The reconstruction metrics, introduced in Section 4.3.1, involve only the encoder-decoder pair and indicate how much information is preserved in the extracted representations. The remaining metrics focus on speech continuation, which is our primary objective, where the performance of the autoregressive model is also assessed.

4.3.1. OBJECTIVE METRICS

Reconstruction Metrics We use F_0 -RMSE, mel-cepstral distortion (MCD), and character error rate (CER) to measure the quality of the reconstructed signal. F_0 -RMSE measures the root mean squared difference between the pitch contour of the ground-truth signal and the reconstructed one. We use CREPE (Kim et al., 2018) to extract pitch and only consider the voiced parts of the signal when computing the difference. MCD measures the Euclidean distance between the 23 mel-cepstral coefficients (MCEPs) extracted from the ground-truth and reconstructed signals. For calculating CER, we use a pre-trained Whisper (Radford et al.,

⁴<https://huggingface.co/facebook/hubert-base-ls960>

⁵<https://huggingface.co/facebook/hubert-large-ll60k>

⁶Except for the *Token-LM + Acoustic Tokens* method, which uses the RVQ decoder directly.

2023) automatic speech recognition model.⁷ We use the `dev-clean` and `dev-other` subsets of LibriSpeech for evaluating reconstruction. To ensure deterministic results, instead of sampling each z_t^c from $q_\phi(z_t^c | \mathbf{X})$, we directly use the Gaussian mean $\mu_\phi(\mathbf{X}, t)$ from Equation 2. In practice, we observed that the stochastic noise of $q_\phi(z_t^c | \mathbf{X})$ has little effect on the reconstructed syntheses.

ZeroSpeech Metrics We adopt the commonly-used metrics (Borsos et al., 2023; Hassid et al., 2023; Maiti et al., 2024) from the ZeroSpeech challenge (Nguyen et al., 2020): sWUGGY and sBLIMP to measure language capability objectively. For these two metrics, speech utterances are given in positive-negative pairs, with each model scoring both utterances. The model’s accuracy is the percentage of instances where the positive example receives a higher score than the negative one. sWuggy measures if the model scores a real word higher than a phonetically similar non-word (e.g., “brick” v.s. “blick”). sBLIMP measures if a model scores a grammatically correct sentence higher than a similar but incorrect one (e.g., “the dogs sleep” vs. “the dog sleep”). Both metrics use text-to-speech to generate the examples. In line with Borsos et al. (2023), we evaluate sWUGGY using only words existing in LibriSpeech (referred as the “in-vocab” version). We use the test split for evaluation. See Appendix G for detailed description on how we estimate the scores for the methods.

4.3.2. SUBJECTIVE METRICS

We use subjective human evaluations to assess the naturalness and meaningfulness of the generated speech. We randomly sampled 100 utterances from the LibriSpeech `dev-clean` and `dev-other` subsets, cropping the first three seconds to use as prompts. Each audio sample was rated by seven annotators. For naturalness, annotators rated how human-like the generated speech sounded on a five-point Likert scale, where one corresponds to “Very unnatural” and five to “Very natural.” For meaningfulness, they rated the grammar and content of the speech on a five-point Likert scale, where one corresponds to “Very Poor” and five to “Excellent.” Additional details on the subjective evaluations are provided in Appendix D.

5. Experimental Results

5.1. Main Results

Tables 1 and 2 present the results for the three methods described in Section 4.2. Table 1 reports objective metrics for speech reconstruction, while Table 2 provides both objective and subjective results for speech continuation. We discuss our observations below.

⁷<https://huggingface.co/openai/whisper-medium>

Table 1. Results of speech reconstruction evaluation (F_0 -RMSE, MCD, CER) for the models discussed in Section 4.2. The evaluation metrics are detailed in Section 4.3. All models were trained on the Libri-light dataset.

Method	F_0 -RMSE(↓)	MCD(↓)	CER(↓)
Ground-truth	<i>n/a</i>	<i>n/a</i>	2.35
<i>Token-LM</i>	43.90	7.55	10.19
+ <i>Pitch</i>	25.46	6.90	6.59
+ <i>Acoustic</i>	15.05	2.58	3.73
<i>Proposed</i>	16.56	5.43	4.35

Reconstruction Quality. First, the results in Table 1 show that compared to *Token-LM* and *Token-LM + Pitch*, our proposed approach improves the reconstruction of the original signal. These findings highlight three key points: 1) discrete semantic tokens alone are insufficient to capture all the components necessary for faithful reconstruction, 2) incorporating only pitch information is not enough, and 3) the learned variational features \mathbf{Z}^c in our approach effectively complement the discrete semantic tokens \mathbf{Z}^d , leading to better reconstruction of the speech signal. On the other hand, our proposed method achieves slightly lower reconstruction quality than *Token-LM + Acoustic*. Since the variational features are continuous, they should be able to encode more information than four levels of acoustic tokens. Therefore, our results suggest that the information encoded in the variational features is effectively regularized by the autoregressive losses: \mathcal{L}_{kl}^c and \mathcal{L}_{kl}^d .

Speech continuation of our approach is more natural compared to the speech generated from the baselines.

The subjective evaluation of speech continuation, measured by the mean opinion score of naturalness (N-MOS) in Table 2, shows that *the syntheses produced by our proposed approach have significantly higher naturalness compared to all baselines*. This finding further supports our hypothesis that the variational features \mathbf{Z}^c learned by our approach improve the quality of the synthesis. While *Token-LM + Acoustic* achieves the best reconstruction in Table 1, the autoregressive model struggles to effectively process the additional information encoded in the RVQ tokens, resulting in significantly lower speech continuation performance, as shown in Table 2. Additionally, Table 2 compares the number of parameters between different methods. The result indicates that the overhead of the proposed method is relatively small (< 1% of the total parameters), while still achieving noticeably better performance.

Speech generated using our proposed approach achieves subjective meaningfulness (as measured by M-MOS) comparable to the baselines. The results in Table 2 in-

Table 2. Results of speech continuation evaluation for the models discussed in Section 4.2. The evaluation metrics are detailed in Section 4.3. M-MOS refers to the meaningfulness mean opinion score. N-MOS refers to the naturalness mean opinion score. Both M-MOS and N-MOS are evaluated on speech continuation and are presented along with 95% confidence intervals. All models were trained on the Libri-light dataset. ‘# Param.’ refers to the number of parameters used during inference. ‘M’ stands for million.

Method	# Param.	sWUGGY(↑)	sBLIMP(↑)	M-MOS(↑)	N-MOS(↑)
Ground-truth	<i>n/a</i>	<i>n/a</i>	<i>n/a</i>	3.94 ± 0.08	3.89 ± 0.09
<i>Token-LM</i>	219M	61.75	58.31	3.24 ± 0.09	3.19 ± 0.11
<i>Token-LM + Pitch</i>	219M	60.75	56.92	3.29 ± 0.09	3.08 ± 0.10
<i>Token-LM + Acoustic</i>	226M	56.23	52.03	2.75 ± 0.09	3.03 ± 0.10
<i>Proposed</i>	221M	60.48	56.56	3.45 ± 0.09	3.60 ± 0.10

indicate that our proposed approach produces syntheses that are comparable to or better than baselines, as reflected by its higher meaningfulness mean opinion score (M-MOS). However, all compared methods show lower sWUGGY and sBLIMP scores than *Token-LM*. This outcome is expected, as the model must predict additional acoustic information beyond semantic tokens, which primarily encode linguistic content. Consequently, given a fixed model parameter budget, language modeling performance naturally declines as the model allocates capacity to model acoustic information. This effect is also evident in the low M-MOS of *Token-LM + Acoustic*, where the acoustic tokens may capture excessively detailed information, such as recording noise, which does not contribute meaningfully to synthesis.

However, one may question why the trend in the sWUGGY and sBLIMP scores does not align with the M-MOS evaluation. We analyze the ASR transcriptions from the compared methods and observe that the transcriptions of *Token-LM* do have higher meaningfulness than those of other approaches, consistent with the trend of the sWUGGY and sBLIMP scores. However, after listening to the audio samples, we found that the natural prosody of our proposed method significantly improves intelligibility. Although Whisper ASR can still transcribe speech of unnatural prosody generated by *Token-LM*, human raters often needed multiple passes to fully comprehend the linguistic content. In practical applications, interactive dialogue systems must generate speech that users can easily understand in a single pass. The M-MOS score serves as an indicator of the suitability of a system in this regard.

5.2. Impact of Loss-balancing Parameters

Here, we study the effect of varying the loss-balancing hyper-parameters: β and γ , which are described in Section 3.3.

Varying β Table 3 shows that, for reconstruction metrics (F_0 -RMSE, MCD, CER), lower values of β result in smaller errors, indicating better reconstruction. However, for the

sWUGGY and sBLIMP metrics, performance decreases as β increases. This finding aligns with our discussion in Section 3.3, where we discussed how lower β values encourage better reconstruction, but make it harder for the autoregressive model to effectively model \mathbf{Z}^c .

Varying γ Table 4 shows that increasing γ leads to worse pitch reconstruction, as measured by F_0 -RMSE, but improves CER. This result indicates that γ governs the type of information captured in the variational feature \mathbf{Z}^c . With a higher γ , the system prioritizes the prediction of semantic tokens. Therefore, the variational feature \mathbf{Z}^c is encouraged to encode more phonetic information, resulting in lower CER and MCD. In contrast, a lower γ encourages \mathbf{Z}^c to focus more on encoding pitch-related information, as indicated by the lower F_0 -RMSE. Then, we analyze subjective measures and observe that both M-MOS and N-MOS favor a lower γ . We attribute the performance decline to the increased difficulty in autoregressive generation of \mathbf{Z}^c . By increasing the weight of \mathcal{L}_{kl}^d , the model sacrifices its focus on minimizing \mathcal{L}_{kl}^c , which in turn compromises its ability to model \mathbf{Z}^c .

5.3. Removing the Semantic Tokens

Here, we evaluate the utility of the semantic tokens in our proposed approach by training a model that uses only variational features \mathbf{Z}^c . This removal corresponds to only training a VAE with an autoregressive prior with Equation 3 without the use of discrete semantic tokens.

Table 3 shows the impact of removing the discrete semantic tokens from our proposed approach, which is denoted as *Proposed (-tokens)*. We find that excluding semantic tokens leads to a slight improvement in the sWUGGY metric compared to including them. However, this exclusion significantly worsens the CER, indicating poorer phonetic reconstruction. These results suggest that without discrete semantic tokens, our approach struggles to effectively encode abstract phonetic information in the variational features (\mathbf{Z}^c) but still performs well on sWUGGY, possibly by leveraging other cues. One possible explanation is that the

Table 3. Results showing the impact of varying the β parameter (as described in Section 3.3) and the effect of removing phonetic tokens from our proposed approach on both language modeling and speech reconstruction performance. The γ parameter (as described in Section 3.3) for the proposed methods is fixed to 0.5. All models here were trained on the LibriSpeech dataset for lower computation cost.

Method	β	sWUGGY(\uparrow)	sBLIMP(\uparrow)	F_0 -RMSE(\downarrow)	MCD(\downarrow)	CER(\downarrow)
<i>Proposed</i>	0.03	65.56	51.12	16.76	5.19	5.06
	0.04	65.96	51.40	16.88	5.53	5.43
	0.05	66.46	51.77	17.20	5.75	5.45
<i>Proposed (-tokens)</i>	0.04	69.33	51.85	17.47	5.48	13.02

Table 4. Results showing the impact of varying the γ parameter (as described in Section 3.3) in our proposed approach on both language modeling and speech reconstruction performance. The β parameter (as described in Section 3.3) is fixed to 0.04. M-MOS denotes the meaningfulness mean opinion score, and N-MOS denotes the naturalness mean opinion score, both presented with 95% confidence intervals. All models were trained on the Libri-light dataset.

γ	sWUGGY(\uparrow)	sBLIMP(\uparrow)	F_0 -RMSE(\downarrow)	MCD(\downarrow)	CER(\downarrow)	M-MOS(\uparrow)	N-MOS(\uparrow)
0.5	60.48	59.88	16.56	5.43	4.35	3.45 \pm 0.09	3.60 \pm 0.10
1.0	59.41	59.12	17.06	5.36	4.05	3.31 \pm 0.09	3.46 \pm 0.10
2.0	58.19	58.19	17.41	5.21	3.75	3.07 \pm 0.09	3.26 \pm 0.11

Table 5. Results of speech continuation evaluation for comparison on different semantic token extraction methods detailed in Section 5.4. M-MOS and N-MOS refer to the meaningfulness and naturalness mean opinion score, presented along with 95% confidence intervals. All models were trained on the Libri-light dataset.

Method	M-MOS(\uparrow)	N-MOS(\uparrow)
Ground Truth	3.87 \pm 0.08	3.97 \pm 0.08
<i>SpeechTokenizer-LM</i>	3.26 \pm 0.09	3.33 \pm 0.10
<i>Proposed</i>	3.68 \pm 0.09	3.61 \pm 0.10

synthesized non-existent words in sWUGGY, being out-of-domain for the text-to-speech system, may exhibit subtle prosodic irregularities that our model is able to detect. On the other hand, the best reconstruction results are obtained when semantic tokens are included, as removing them leads to worse reconstruction metrics.

5.4. Generalization to Different Semantic Tokens

In Section 5.1, we demonstrated the effectiveness of our proposed approach using semantic tokens derived from HuBERT representations. Here, we investigate its performance with an alternative approach to extracting semantic tokens, SpeechTokenizer (Zhang et al., 2024). SpeechTokenizer quantizes speech using Residual Vector Quantization (RVQ), which optimizes for reconstruction. However, its first-level RVQ tokens additionally minimizing distillation loss with HuBERT representations to encode content. We replace the semantic tokens in *Token-LM* with the first-level RVQ tokens from SpeechTokenizer, naming this new baseline *SpeechTokenizer-LM*. Our proposed method was similarly

adapted to this new set of semantic tokens. For our experiments, we used the official SpeechTokenizer checkpoint ⁸.

As shown in Table 5, our approach achieved superior naturalness and meaningfulness scores compared to *SpeechTokenizer-LM*. This verifies that our framework effectively enhances various approaches to extracting semantic tokens.

Flexibility with Different Decoders Additionally, for both *SpeechTokenizer-LM* and our proposed method, we did not adopt the diffusion decoder mentioned in Section 3.5. Instead, we predicted the remaining RVQ tokens from the semantic tokens (or semantic tokens and variational features for our approach) and leveraged the pre-trained SpeechTokenizer decoder for speech reconstruction. As noted in Section 3.5, our training framework is adaptable and is not tied to a specific decoder type. We adopted a diffusion-based decoder for simplified training and fair comparisons in our previous work. The empirical results in Table 5 further validate this flexibility, as our model still achieves high human evaluation MOS scores with a different decoder.

6. Related Work

Emerging speech language models typically use discrete semantic tokens for autoregressive modeling. These tokens are often obtained by k -means clustering of features extracted from self-supervised pre-trained models (Hsu et al., 2021; Chen et al., 2022). For instance, Lakhota et al. (2021) used semantic tokens for generative spoken language mod-

⁸<https://github.com/ZhangXInFD/SpeechTokenizer/>

eling (GSLM). Subsequently, [Kharitonov et al. \(2022\)](#) enhanced this approach by incorporating pitch information alongside semantic tokens as joint inputs to the autoregressive model. Our proposed approach improves upon this line of research by using a variational autoencoder to automatically learn paralinguistic speech attributes in conjunction with the autoregressive model. [Borsos et al. \(2023\)](#) proposed a two-stage approach for the decoder that used acoustic tokens ([Zeghidour et al., 2022](#); [Défossez et al., 2023](#)). This type of framework is also widely used in text-to-speech systems ([Chen et al., 2025](#); [2024](#)). In contrast, our approach focuses on the joint modeling of linguistic and paralinguistic features by enhancing the inputs to the autoregressive model rather than improving the decoder.

Recently, a line of research has emerged focusing on improving speech language models through the integration of text-based models. [Hassid et al. \(2023\)](#) initialized their speech language model using a pre-trained text-based large language model (LLM). Similarly, [Rubenstein et al. \(2023\)](#); [Maiti et al. \(2024\)](#) expanded the vocabulary of pre-trained text-based LLMs by integrating the semantic tokens. Building on this, [Yang et al. \(2024\)](#); [Du et al. \(2024\)](#) further explored multi-task training involving text-conditioned generative speech tasks, combining text and audio within a single LLM. We note that our proposed approach takes a different direction but can still be integrated with these approaches. For example, one could initialize the transformer in our autoregressive model using parameters from a text-based LLM.

Recent works ([Défossez et al., 2024](#)) incorporate discrete acoustic tokens directly into autoregressive modeling. However, these approaches often require complex designs, such as delay patterns and text-based pretraining. In Section 5, we demonstrate that directly incorporating acoustic tokens to autoregressive modeling significantly affects the generation of linguistic content, while our method does not.

7. Conclusion

In this work, we proposed an approach that combines a variational autoencoder with existing token-based speech language models. We conducted experiments to evaluate its effectiveness in terms of language capability and synthesis naturalness. Empirical evaluations suggest that our proposed approach, in contrast with other recent techniques, is capable of producing synthesis with better subjective meaningfulness and naturalness. Additionally, we examined the effects of the weights of different loss terms, β and γ , on performance. Our findings indicate that β governs the amount of information encoded from the mel-spectrogram into the variational feature, whereas γ controls the type of information encoded within the variational feature.

8. Limitations and Future Work

Our results indicate that the performance of our proposed approach is sensitive to the choice of hyper-parameters β and γ . Future work will explore automated methods for tuning these hyper-parameters. Additionally, our evaluation is limited to English datasets, and it remains unclear if the approach generalizes to languages with different prosodic patterns. Future work will extend training and evaluation to additional languages to assess cross-lingual applicability. Finally, our model has a relatively small number of parameters and is trained on a smaller dataset compared to existing frameworks ([Hassid et al., 2023](#); [Rubenstein et al., 2023](#)). We plan to scale up both the model and the training data to examine whether our findings hold with increased computational resources and larger datasets.

Acknowledgments

This work was supported by a grant from Apple. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies or position, either expressed or implied, of Apple. We thank Professor Shinji Watanabe for his valuable feedback. We also thank the reviewers and the area chair for their insightful comments and suggestions.

Impact Statement

We proposed an approach that improves the naturalness of speech language models without compromising their language proficiency, which can be leveraged by existing paradigms in this literature. While a model that generates more natural speech can enhance the user experience in conversational agents, it can also be exploited for harmful purposes, such as creating fake videos or making spam phone calls.

References

- Adigwe, A., Tits, N., Haddad, K. E., Ostadabbas, S., and Dutoit, T. The emotional voices database: Towards controlling the emotion dimension in voice generation systems, 2018. URL <https://arxiv.org/abs/1806.09514>.
- Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audioldm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409.

- Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. In Riezler, S. and Goldberg, Y. (eds.), *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL <https://aclanthology.org/K16-1002>.
- Chen, L.-W., Watanabe, S., and Rudnicky, A. A vector quantized approach for text to speech synthesis on real-world spontaneous speech. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11): 12644–12652, Jun. 2023. doi: 10.1609/aaai.v37i11.26488. URL <https://ojs.aaai.org/index.php/AAAI/article/view/26488>.
- Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., and Wei, F. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.
- Chen, S., Liu, S., Zhou, L., Liu, Y., Tan, X., Li, J., Zhao, S., Qian, Y., and Wei, F. VALL-E 2: Neural codec language models are human parity zero-shot text to speech synthesizers, 2024. URL <https://arxiv.org/abs/2406.05370>.
- Chen, S., Wang, C., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S., and Wei, F. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025. doi: 10.1109/TASLPRO.2025.3530270.
- Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL <https://openreview.net/forum?id=ivCd8z8zR2>. Featured Certification, Reproducibility Certification.
- Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=HkpbhH91x>.
- Du, Z., Wang, J., Chen, Q., Chu, Y., Gao, Z., Li, Z., Hu, K., Zhou, X., Xu, J., Ma, Z., Wang, W., Zheng, S., Zhou, C., Yan, Z., and Zhang, S. LauraGPT: Listen, attend, understand, and regenerate audio with GPT, 2024. URL <https://arxiv.org/abs/2310.04673>.
- Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N. Moshi: a speech-text foundation model for real-time dialogue, 2024. URL <https://arxiv.org/abs/2410.00037>.
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.
- Fu, H., Li, C., Liu, X., Gao, J., Celikyilmaz, A., and Carin, L. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 240–250, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1021. URL <https://aclanthology.org/N19-1021>.
- Hassid, M., Remez, T., Nguyen, T. A., Gat, I., Conneau, A., Kreuk, F., Copet, J., Defossez, A., Synnaeve, G., Dupoux, E., Schwartz, R., and Adi, Y. Textually pretrained speech language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 63483–63501. Curran Associates, Inc., 2023.
- Hendrycks, D. and Gimpel, K. Gaussian error linear units (GELUs). In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Bk0MRI5lg>.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=Sy2fzU9gl>.
- Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.
- Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhota, K., Salakhutdinov, R., and Mohamed, A. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Au-*

- dio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.
- 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., Lee, A., Polyak, A., Adi, Y., Copet, J., Lakhotia, K., Nguyen, T. A., Riviere, M., Mohamed, A., Dupoux, E., and Hsu, W.-N. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8666–8681, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.593. URL <https://aclanthology.org/2022.acl-long.593>.
- Kim, J., Kim, S., Kong, J., and Yoon, S. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8067–8077. Curran Associates, Inc., 2020.
- Kim, J., Kong, J., and Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5530–5540. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/kim21f.html>.
- Kim, J. W., Salamon, J., Li, P., and Bello, J. P. Crepe: A convolutional representation for pitch estimation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, 2018. doi: 10.1109/ICASSP.2018.8461329.
- Kingma, D. P. and Welling, M. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8245. doi: 10.1561/22000000056. URL <http://dx.doi.org/10.1561/22000000056>.
- Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17022–17033. Curran Associates, Inc., 2020.
- Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., and Dupoux, E. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl_a_00430. URL <https://aclanthology.org/2021.tacl-1.79>.
- Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Maiti, S., Peng, Y., Choi, S., Jung, J.-W., Chang, X., and Watanabe, S. VoxTlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13326–13330, 2024. doi: 10.1109/ICASSP48485.2024.10447112.
- Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., and Dupoux, E. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling, 2020. URL <https://arxiv.org/abs/2011.11588>.
- OpenAI. GPT-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.
- 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, 2015. doi: 10.1109/ICASSP.2015.7178964.
- Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11671. URL <https://ojs.aaai.org/index.php/AAAI/article/view/11671>.
- Press, O., Smith, N., and Lewis, M. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=R8sQPpGCv0>.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul

2023. URL <https://proceedings.mlr.press/v202/radford23a.html>.
- Reyes-González, H. and Torre, R. Testing the boundaries: Normalizing flows for higher dimensional data sets. *Journal of Physics: Conference Series*, 2438 (1):012155, 02 2023. Copyright - Published under licence by IOP Publishing Ltd. This work is published under <http://creativecommons.org/licenses/by/3.0/> (the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2023-11-28.
- Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. URL <https://proceedings.mlr.press/v37/rezende15.html>.
- Ronneberger, O., Fischer, P., and Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F. (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.
- Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., de Chaumont Quitry, F., Chen, P., Badawy, D. E., Han, W., Kharitonov, E., Muckenhirn, H., Padfield, D., Qin, J., Rozenberg, D., Sainath, T., Schalkwyk, J., Sharifi, M., Ramanovich, M. T., Tagliasacchi, M., Tudor, A., Velimirović, M., Vincent, D., Yu, J., Wang, Y., Zayats, V., Zeghidour, N., Zhang, Y., Zhang, Z., Zilka, L., and Frank, C. AudioPaLM: A large language model that can speak and listen, 2023. URL <https://arxiv.org/abs/2306.12925>.
- Schuller, B., Steidl, S., and Batliner, A. The interspeech 2009 emotion challenge. In *Interspeech 2009*, pp. 312–315, 2009. doi: 10.21437/Interspeech.2009-103.
- Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Wening, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Interspeech 2013*, pp. 148–152, 2013. doi: 10.21437/Interspeech.2013-56.
- Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=StlgjarCHLP>.
- Sun, G., Zhang, Y., Weiss, R. J., Cao, Y., Zen, H., Rosenberg, A., Ramabhadran, B., and Wu, Y. Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and autoregressive prosody prior. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6699–6703, 2020. doi: 10.1109/ICASSP40776.2020.9053436.
- Ulyanov, D., Vedaldi, A., and Lempitsky, V. Instance normalization: The missing ingredient for fast stylization, 2017. URL <https://arxiv.org/abs/1607.08022>.
- Vahdat, A. and Kautz, J. Nvae: A deep hierarchical variational autoencoder. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19667–19679. Curran Associates, Inc., 2020.
- Weston, J., Lenain, R., Meepegama, U., and Fristed, E. Learning de-identified representations of prosody from raw audio. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11134–11145. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/weston21a.html>.
- Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., and Liu, T. On layer normalization in the transformer architecture. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/xiong20b.html>.
- Yamagishi, J., Veaux, C., and MacDonald, K. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), 2019. URL <https://doi.org/10.7488/ds/2645>.
- Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Guo, H., Chang, X., Shi, J., Zhao, S., Bian, J., Zhao, Z., Wu, X., and Meng, H. M. UniAudio: Towards universal audio generation with large language models. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 56422–56447. PMLR, 21–27 Jul 2024. URL <https://proceedings.mlr.press/v235/yang24x.html>.
- 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*, 30:495–507, 2022. doi: 10.1109/TASLP.2021.3129994.

Zhang, B. and Sennrich, R. Root mean square layer normalization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Zhang, X., Zhang, D., Li, S., Zhou, Y., and Qiu, X. SpeechTokenizer: Unified speech tokenizer for speech language models. In *International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=AF9Q8Vip84>.

Zhu, Y., Min, M. R., Kadav, A., and Graf, H. P. S3vae: Self-supervised sequential vae for representation disentanglement and data generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

A. Mathematical Derivations

A.1. Equation 3

For notation simplicity, we drop the superscript c of \mathbf{Z}^c into \mathbf{Z} in this proof.

With the parameterized prior, the modeling distribution of \mathbf{X} now also depends on ψ :

$$\begin{aligned} p_{\theta,\psi}(\mathbf{X}) &= \int p_{\theta}(\mathbf{X} | \mathbf{Z}) p_{\psi}(\mathbf{Z}) d\mathbf{Z}, \\ p_{\theta,\psi}(\mathbf{Z} | \mathbf{X}) &= \frac{p_{\theta,\psi}(\mathbf{X}, \mathbf{Z})}{p_{\theta,\psi}(\mathbf{X})} = \frac{p_{\theta}(\mathbf{X} | \mathbf{Z}) p_{\psi}(\mathbf{Z})}{p_{\theta,\psi}(\mathbf{X})}. \end{aligned}$$

Following a similar proof in Kingma & Welling (2019):

Proof.

$$\begin{aligned} \log p_{\theta,\psi}(\mathbf{X}) &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} [\log p_{\theta,\psi}(\mathbf{X})] \\ &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{p_{\theta,\psi}(\mathbf{X}, \mathbf{Z})}{p_{\theta,\psi}(\mathbf{Z} | \mathbf{X})} \right] \right] \\ &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{p_{\theta,\psi}(\mathbf{X}, \mathbf{Z}) q_{\phi}(\mathbf{Z} | \mathbf{X})}{p_{\theta,\psi}(\mathbf{Z} | \mathbf{X}) q_{\phi}(\mathbf{Z} | \mathbf{X})} \right] \right] \\ &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{p_{\theta,\psi}(\mathbf{X}, \mathbf{Z})}{q_{\phi}(\mathbf{Z} | \mathbf{X})} \right] \right] + \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{q_{\phi}(\mathbf{Z} | \mathbf{X})}{p_{\theta,\psi}(\mathbf{Z} | \mathbf{X})} \right] \right] \\ &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{p_{\theta,\psi}(\mathbf{X}, \mathbf{Z})}{q_{\phi}(\mathbf{Z} | \mathbf{X})} \right] \right] + D_{KL}(q_{\phi}(\mathbf{Z} | \mathbf{X}) || p_{\theta,\psi}(\mathbf{Z} | \mathbf{X})). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathcal{O}_{ELBO} &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{p_{\theta,\psi}(\mathbf{X}, \mathbf{Z})}{q_{\phi}(\mathbf{Z} | \mathbf{X})} \right] \right] \\ &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{p_{\theta}(\mathbf{X} | \mathbf{Z}) p_{\psi}(\mathbf{Z})}{q_{\phi}(\mathbf{Z} | \mathbf{X})} \right] \right] \\ &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} [\log p_{\theta}(\mathbf{X} | \mathbf{Z})] + \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{p_{\psi}(\mathbf{Z})}{q_{\phi}(\mathbf{Z} | \mathbf{X})} \right] \right] \\ &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} [\log p_{\theta}(\mathbf{X} | \mathbf{Z})] - D_{KL}(q_{\phi}(\mathbf{Z} | \mathbf{X}) || p_{\psi}(\mathbf{Z})). \end{aligned}$$

With $q_{\phi}(\mathbf{Z} | \mathbf{X}) = \prod_{t=1}^T q_{\phi}(z_t | \mathbf{X})$, and $p_{\psi}(\mathbf{Z}) = \prod_{t=1}^T p_{\psi}(z_t | \mathbf{Z}_{1:t-1})$:

$$\begin{aligned} D_{KL}(q_{\phi}(\mathbf{Z} | \mathbf{X}) || p_{\psi}(\mathbf{Z})) &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{q_{\phi}(\mathbf{Z} | \mathbf{X})}{p_{\psi}(\mathbf{Z})} \right] \right] \\ &= \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{\prod_{t=1}^T q_{\phi}(z_t | \mathbf{X})}{\prod_{t=1}^T p_{\psi}(z_t | \mathbf{Z}_{1:t-1})} \right] \right] \\ &= \sum_{t=1}^T \mathbb{E}_{\mathbf{Z} \sim q_{\phi}(\mathbf{Z} | \mathbf{X})} \left[\log \left[\frac{q_{\phi}(z_t | \mathbf{X})}{p_{\psi}(z_t | \mathbf{Z}_{1:t-1})} \right] \right] \\ &= \sum_{t=1}^T \mathbb{E}_{\mathbf{Z}_{1:t-1}} [D_{KL}(q_{\phi}(z_t | \mathbf{X}) || p_{\psi}(z_t | \mathbf{Z}_{1:t-1}))], \end{aligned}$$

where $\mathbf{Z}_{1:t-1} \sim \prod_{t=1}^T q_{\phi}(z_t | \mathbf{X})$. □

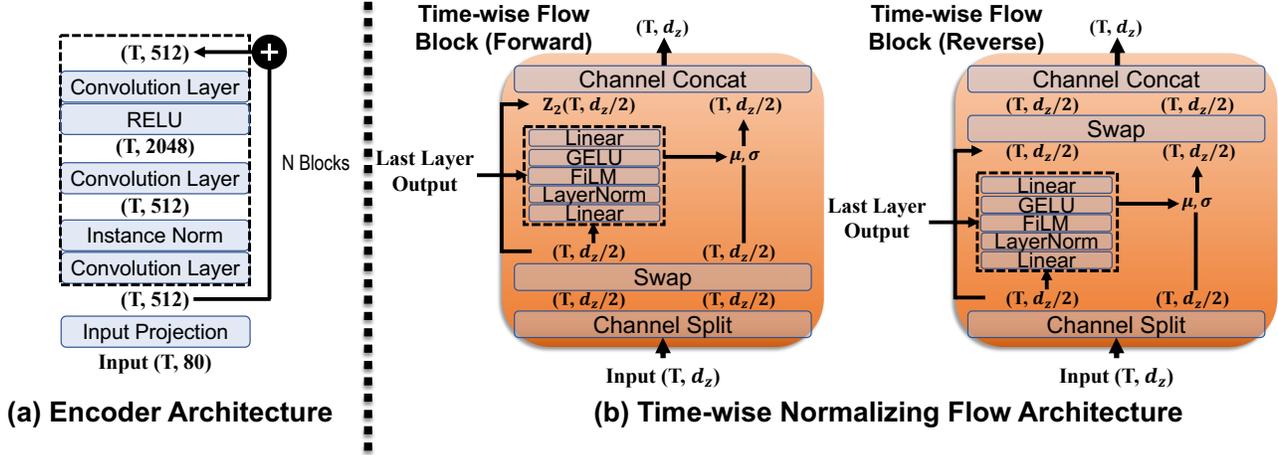


Figure 2. (a) Residual block architecture or the encoder ϕ . (b) Model architecture for the time-wise normalization flow introduced in Section 3.4.

A.2. Equation 4

Proof. Since \mathcal{O}_{rec} is straightforward to derive from Equation 1 (decompose \mathbf{Z} into \mathbf{Z}^c and \mathbf{Z}^d), here we show how \mathcal{L}_{kl}^c and \mathcal{L}_{kl}^d are derived from the $D_{KL}(q_\phi(\mathbf{Z} | \mathbf{X}) || p_\psi(\mathbf{Z}))$ in Equation 1.

With $q_\phi(\mathbf{Z} | \mathbf{X}) = q_\phi(\mathbf{Z}^c | \mathbf{X})p(\mathbf{Z}^d | \mathbf{X})$ and $p_\psi(z_t | \mathbf{Z}_{1:t-1}) = p_\psi(z_t^d | \mathbf{Z}_{1:t-1})p_\psi(z_t^c | \mathbf{Z}_{1:t-1})$:

$$\begin{aligned}
& D_{KL}(q_\phi(\mathbf{Z} | \mathbf{X}) || p_\psi(\mathbf{Z})) \\
&= \mathbb{E}_{\mathbf{Z}} \left[\log \left[\frac{q_\phi(\mathbf{Z} | \mathbf{X})}{p_\psi(\mathbf{Z})} \right] \right] \\
&= \mathbb{E}_{\mathbf{Z}} \left[\log \left[\frac{q_\phi(\mathbf{Z}^c | \mathbf{X})p(\mathbf{Z}^d | \mathbf{X})}{\prod_{t=1}^T p_\psi(z_t | \mathbf{Z}_{1:t-1})} \right] \right] \\
&= \mathbb{E}_{\mathbf{Z}} \left[\log \left[\frac{q_\phi(\mathbf{Z}^c | \mathbf{X})p(\mathbf{Z}^d | \mathbf{X})}{\prod_{t=1}^T p_\psi(z_t^c | \mathbf{Z}_{1:t-1})p_\psi(z_t^d | \mathbf{Z}_{1:t-1})} \right] \right] \\
&= \mathbb{E}_{\mathbf{Z}} \left[\log \left[\frac{q_\phi(\mathbf{Z}^c | \mathbf{X})}{\prod_{t=1}^T p_\psi(z_t^c | \mathbf{Z}_{1:t-1})} \right] \right] + \mathbb{E}_{\mathbf{Z}} \left[\log \left[\frac{p(\mathbf{Z}^d | \mathbf{X})}{\prod_{t=1}^T p_\psi(z_t^d | \mathbf{Z}_{1:t-1})} \right] \right] \\
&= \sum_{t=1}^T \mathbb{E}_{\mathbf{Z}_{1:t-1}} [D_{KL}(q_\phi(z_t^c | \mathbf{X}) || p_\psi(z_t^c | \mathbf{Z}_{1:t-1}))] - \sum_{t=1}^T \mathbb{E}_{\mathbf{Z}_{1:t}} [\log p_\psi(z_t^d | \mathbf{Z}_{1:t-1})] \\
&+ \mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{Z}^d | \mathbf{X})]
\end{aligned}$$

Since $\mathbb{E}_{\mathbf{Z}} [\log p(\mathbf{Z}^d | \mathbf{X})]$ does not depend on any parameters, it can be dropped during optimization. \square

B. Model Architectures

Encoder $q_\phi(\mathbf{Z} | \mathbf{X})$ We use a different number of residual blocks for the encoder. We use a kernel size of 7; the hidden dimensions used for all models are in Figure 2 (a). The architecture of the residual block is illustrated in Figure 2 (a). Finally, after 3 residual blocks, we apply another instance normalization, followed by separate linear heads to output the mean and log-variance of Equation 2. We used the same size encoder for experiments with LibriSpeech and Libri-light. Instance Norm refers to instance normalization (Ulyanov et al., 2017).

Autoregressive Transformer We follow the typical implementation of transformers with Post-LN (Xiong et al., 2020). We use RMSNorm (Zhang & Sennrich, 2019) and GELU activation (Hendrycks & Gimpel, 2017). We use ALiBi (Press

Table 6. Model configuration of the autoregressive transformer for training on LibriSpeech and Libri-light respectively. This configuration is shared for all comparing methods. ‘feed-forward size’ refers to the width of the feed-forward linear layer.

Dataset	# of layers	# of heads	hidden size	feed-forward size
LibriSpeech	4	8	512	2048
Libri-light	16	16	1024	4096

et al., 2022) for relative positional encoding. We used different model sizes for the LibriSpeech and Libri-light experiments, with the configuration summarized in Table 6. The same configuration is shared for all comparison methods.

Time-wise Normalizing Flow The architecture of our time-wise normalizing flow is illustrated in Figure 2 (b). Here, μ and σ are the mean and standard deviation that will be multiplied and added to the input. This part mainly follows the implementation of Dinh et al. (2017). The “Last Layer Output” in Figure 2 (b) refers to the output of the last transformer layer. “FiLM” refers to FiLM conditioning (Perez et al., 2018). “Swap” refers to the swapping of the two inputs in their channel order. We used 4 flow blocks for all experiments.

Diffusion Decoder For our diffusion decoder θ , we apply the same residual block as in Figure 2 (a). However, here we have additional skip connections between the output of residual blocks following the commonly-used U-Net architecture (Ronneberger et al., 2015). We encode the current diffusion step with sinusoidal positional encoding, linear project it and add it to each time frame of the output of the first convolution layer in each of the residual blocks. For both data sets, we used six residual blocks, with the same hidden dimensions and kernel size as the encoder ϕ .

Utterance Encoder The utterance encoder consists of 3 blocks, where each block sequentially includes a convolution with stride 2 and kernel size 4, followed by instance normalization (Ulyanov et al., 2017) and RELU activation. The hidden size of the convolution layer is: 128, 256, 512. Afterward, a simple time-averaging is applied to the output to generate an utterance-level embedding.

C. Training Details

For model training, we use the AdamW optimizer (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9, \beta_2 = 0.98$. We used a weight decay of 0.01 for LibriSpeech models and 0.1 for the Libri-light models. We trained the models with mixed precision. For Libri-light models, we use 2 L40S GPUs with gradient accumulation of step size 2. This makes the effective batch size 192. We trained for 600k update steps. We warm β from 0 to the final value in the first 30k update steps. It takes about 14 days to train the Libri-light models.

For LibriSpeech models, we discovered that methods involving discrete tokens suffer from early overfitting (but not in Libri-light). Therefore, we train these models (including our proposed approach) to only 100k steps. For the diffusion decoder of *Token-LM* and *Token-LM + pitch*, we separately train them to 500k steps, where we observe marginal improvement of loss functions between epochs. For pure variational approaches, we train to 400k steps as we did not observe overfitting. We used the same effective batch size on the 2 L40S GPUs but without gradient accumulation. For LibriSpeech models, we warm β from 0 to the final value in the first 20k update steps. It takes about 2 days to train for the 400k step models and less than 1 day for the 100k step models. For both models, we use an initial learning rate of $5e - 4$ and apply cosine learning rate decay to $5e - 5$.

For the input to the utterance encoder, we randomly cropped the segment to be between 2 and 4 seconds. For diffusion model, we use L1 loss to predict the diffusion noise and apply the cosine schedule for the diffusion noise variance.

D. Subjective Evaluation

We use crowd-sourcing for subjective human evaluation on speech meaningfulness and naturalness. The recruited raters speak English and were paid at least the minimum wage. We sample 100 prompts from LibriSpeech development subsets, crop the first 3 seconds, and feed to each model to produce a 10-second continuation (total 13 seconds). The same 100 prompts are used across all methods for a fair comparison. Since we do not train our model to predict the end of speech, we

Please listen to the computer-generated speech sample below and rate how well its grammar and content convey meaningful information. Focus on evaluating the grammar and the content, not the naturalness or quality of the speech.



Figure 3. A screenshot of the Meaningfulness (M-MOS) assessment task, as the crowd-sourced rater sees it.

Please listen to the computer generated speech sample below and rate how natural (i.e., human-like) it sounds on a scale from 1 (very unnatural) to 5 (very natural).

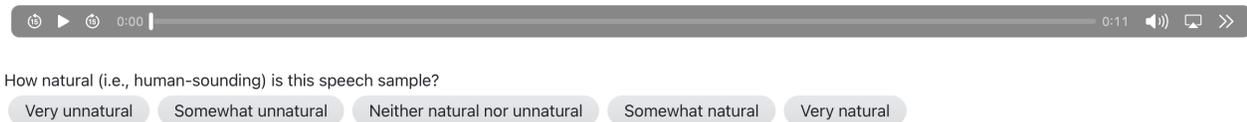


Figure 4. A screenshot of the Naturalness (N-MOS) assessment task, as the crowd-sourced rater sees it.

Table 7. Performance varying latent dimension d_z^c on our proposed approach (without speech tokens). Models are trained on LibriSpeech.

d_z^c	sWUGGY(↑)	sBLIMP(↑)	F0-RMSE(↓)	MCD(↓)	CER(↓)
4	69.33	51.85	17.47	5.48	13.02
16	73.49	51.69	16.68	5.37	7.80
64	73.25	50.91	17.37	5.35	7.79

observed that some synthesis ends earlier than 13 seconds. We use pre-trained voice activity detection from pyannote⁹ to post-process the samples, removing trailing silences and non-speech that might affect evaluation.

In Figures 3 and 4, we provide screenshots of what the raters see during the evaluation. Raters are presented with a spoken utterance and are instructed to rate its naturalness or meaningfulness on a five-point Likert scale, where 1 corresponds to very unnatural or meaningless and 5 corresponds to very natural or meaningful.

E. Dimension of the latent variable d_z^c

Table 7 presents our results of increasing the latent dimension d_z^c . We perform the sweep in the variational approach without semantic tokens for simplicity. From Table 7, we observe that increasing the latent dimension from 4 to 16 results in uniform improvements across the measures. However, further increasing the dimension from 16 to 64 leads to marginal degradation. We speculate that this performance plateau may arise from the difficulty normalizing flows face when modeling higher-dimensional distributions (Reyes-González & Torre, 2023).

F. Discrete token vocabulary size

Table 8 shows our evaluation results on semantic token models (*Token-LM*) trained with varying k . Here, k refers to number of clusters for the k -means clustering on obtaining the discrete token, which is equal to the vocabulary size of the discrete tokens. Our result is consistent with (Maiti et al., 2024), which shows that $k = 200$ obtains the best sWUGGY score.

⁹<https://huggingface.co/pyannote/voice-activity-detection>

Table 8. Comparison of model trained on different number of discrete tokens k . Models are trained on LibriSpeech.

k	sWUGGY(\uparrow)	sBLIMP(\uparrow)	F_0 -RMSE(\downarrow)	MCD(\downarrow)	CER(\downarrow)
50	59.63	52.49	41.11	6.49	11.87
200	67.32	52.46	35.41	6.23	5.40
1000	65.11	50.99	32.60	5.99	4.48

Table 9. Performance of speech emotion recognition models trained on different features. The features are extracted from models pre-trained on Libri-light using our proposed method.

Method	Emotion Recognition (ACC, %)
<i>Tokens</i>	57.46 \pm 1.59
<i>Variational Features</i>	91.57 \pm 0.35
<i>Tokens + Variational Features</i>	92.74 \pm 0.37

Reconstruction metrics indicate that $k = 200$ provides a significant improvement over $k = 50$, whereas increasing $k = 200$ to $k = 1000$ produces only a marginal gain. Interestingly, having larger k seems to negatively impact sBLIMP. We speculate that the small vocabulary size ($k = 50$) is adequate to distinguish word-level changes in sentences, but insufficient to detect subtle phonetic variations within words.

G. Scoring sWUGGY and sBLIMP

Token-LM To obtain the scores for sWUGGY and sBLIMP for semantic token only models, we follow (Borsos et al., 2023) and use the log-likelihood returned by the model normalized by the sequence length.

Token-LM + Pitch, Token-LM + Acoustic Tokens, Proposed Methods For methods that have additional inputs other than the discrete tokens, we only use the model’s log-likelihood of the discrete tokens. We do not use the log-likelihood of the Z^c , as we assume that the discrete tokens Z^d should contain all the information needed for sWUGGY and sBLIMP. In practice, we do observe that including the log-likelihood of the Z^c slightly lowers the score for our proposed method.

Proposed - token Since there are no discrete tokens involved in *Proposed - token*, we directly use the log-likelihood of Z^c . The likelihood can be estimated using Equation 6.

For *Proposed* and *Proposed w.o. token*, to ensure a deterministic outcome, we again use $\mu_\phi(\mathbf{X}, t)$ from Equation 2 directly as Z^c , instead of sampling Z^c from $q_\phi(z_t | \mathbf{X})$.

H. Side Experiments on inspecting learned features

Speech Emotion Recognition We evaluate speech emotion recognition on the EmoV-DB (Adigwe et al., 2018) dataset. We follow a 9:1 split on training and testing for the dataset. The dataset contains five emotion categories: amused, angry, neutral, disgust, and sleepiness. We train a classifier with the same structure to predict emotion categories based on different features. The experiments are repeated 20 times to report the mean and 95% confidence interval. From Table 9, we can observe that the variational features alone obtain significantly better performance compared to tokens, showcasing its capability of capturing paralinguistic information. Combining both tokens and variational features gives a slight improvement over using variational features alone.

Speaker Identification For speaker identification, we evaluate the performance on the VCTK (Yamagishi et al., 2019) dataset, which consists of read English sentences, with 400 sentences each from 110 speakers. We again follow a 9:1 train-test split and repeat each run 20 times to report the mean and 95% confidence interval. We also evaluate our embedding of utterance, which is designed to capture static utterance-level information (see Section 3.5). From Table 10, we can see that using tokens only results in poor speaker identification accuracy. With variational features, the classifier obtains improved accuracy. We attribute this improvement to the fact that speaking styles can be captured in the variational features to classify speakers. On the other hand, the utterance embedding outperforms the other features in this task. These results support our

Table 10. Performance of speaker identification models trained on different features. The features are extracted from models pre-trained on Libri-light using our proposed method.

Method	Speaker Identification (ACC, %)
<i>Tokens</i>	7.08 ± 0.40
<i>Variational Features</i>	63.41 ± 0.43
<i>Tokens + Variational Features</i>	63.13 ± 0.45
<i>Utterance Embedding</i>	94.06 ± 0.32

claim that the utterance encoder encodes global speaker information while variational features capture local paralinguistic attributes.

I. Conditional independence assumption of z_t^d and z_t^c

In general, \mathbf{Z}^c and \mathbf{Z}^d are not independent, since the language content can imply the paralinguistic information, and vice versa. However, our modeling assumes only conditional independence. Specifically, the past generations $\mathbf{Z}_{1:t-1}$ are first passed through the autoregressive transformer ψ to produce the intermediate representation $o_t = \text{Transformer}_\phi(\mathbf{Z}_{1:t-1})$. Then, two separate heads predict z_t^c and z_t^d based on o_t . This framework assumes that the transformer can learn o_t such that z_t^c and z_t^d become conditionally independent given o_t . Given the transformer’s modeling capacity, we believe it can extract shared information (o_t) between z_t^c and z_t^d from $\mathbf{Z}_{1:t-1}$, while delegating the distinct information to their respective heads.