Llama-VITS: Enhancing TTS Synthesis with Semantic Awareness

Xincan Feng[†‡], Akifumi Yoshimoto[‡]
[†]NARA Institute of Science and Technology, [‡]CyberAgent Inc
feng.xincan.fy2@is.naist.jp, yoshimoto_akifumi_xa@cyberagent.co.jp

Abstract

Recent advancements in Natural Language Processing (NLP) have seen Large-scale Language Models (LLMs) excel at producing high-quality text for various purposes. Notably, in Text-To-Speech (TTS) systems, the integration of BERT for semantic token generation has underscored the importance of semantic content in producing coherent speech outputs. Despite this, the specific utility of LLMs in enhancing TTS synthesis remains considerably limited. This research introduces an innovative approach, Llama-VITS, which enhances TTS synthesis by enriching the semantic content of text using LLM. Llama-VITS integrates semantic embeddings from Llama2 with the VITS model, a leading end-to-end TTS framework. By leveraging Llama2 for the primary speech synthesis process, our experiments demonstrate that Llama-VITS matches the naturalness of the original VITS (ORI-VITS) and those incorporate BERT (BERT-VITS), on the LJSpeech dataset, a substantial collection of neutral, clear speech. Moreover, our method significantly enhances emotive expressiveness on the EmoV_DB_bea_sem dataset, a curated selection of emotionally consistent speech from the EmoV_DB dataset, highlighting its potential to generate emotive speech.

Keywords: Text-To-Speech, Emotive Speech, Large-scale Language Model, Semantic Embedding

1. Introduction

Text-to-Speech (TTS) synthesis is a technology that transforms written text into its spoken equivalent, thereby enhancing content accessibility. This technology finds application in the production of audiobooks (Chen et al., 2022) and virtual assistants (Wu et al., 2023). However, traditional TTS models, which primarily focus on the acoustic features, often fall short in comprehending the semantic and emotional information embedded within the text.

With the significant advancements in Natural Language Processing (NLP) technologies, particularly through Language Models (LMs) such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2018; Brown et al., 2020), which have demonstrated formidable capabilities in understanding and generating natural language, researchers have proposed various BERT-based TTS models (Mukherjee et al., 2022; Abbas et al., 2022; Li et al., 2023; Guo et al., 2022) to improve the expressiveness of synthesized speech. Nonetheless, the effectiveness and flexibility of BERT-based TTS models in diverse applications are limited due to the smaller parameter size of BERT models and the necessity for designing specific fine-tuning tasks to enhance their capabilities.

On the other hand, Large-scale Language Models (LLMs), such as Llama2 (Touvron et al., 2023), not only require decreasing computational resources and achieve higher levels of text generation but also possess excellent zero-shot learning capabilities. Moreover, they can achieve improvements comparable to fine-tuning by adjusting only a minimal number of parameters through prompt tuning (Liu et al., 2022; Tu et al., 2022). However, the potential of these LLMs for TTS tasks has not been fully explored.

In light of this context, we introduce Llama-VITS, a model that leverages semantic representations extracted from Llama2 on top of a state-of-the-art TTS model, VITS (Kim et al., 2021), enabling the generated speech to retain acoustic information while understanding and expressing semantics and emotions. Through comprehensive objective and subjective evaluations, Llama-VITS has been verified to surpass TTS baselines without semantic input or those integrated with BERT.

The main contributions encapsulate:

- We propose Llama-VITS model that utilizes the semantic understanding and expression capabilities of Llama2, offering equal or superior acoustic performance compared to baseline models, along with a significantly enhanced ability to understand and express semantics and emotions.
- Through empirical analysis, we demonstrate that global tokens in Llama-VITS provide more significant improvements than sequential tokens, contrasting with observations in BERT-based TTS models.
- We introduce a new evaluation framework that includes traditional objective metrics for acoustic performance of synthesized speech and incorporates ESMOS, a subjective metric to assess emotional expression.

Our code, models, audio demos, and the filtered single female speaker emotional dataset

LREC-COLING 2024, pages 10642–10656

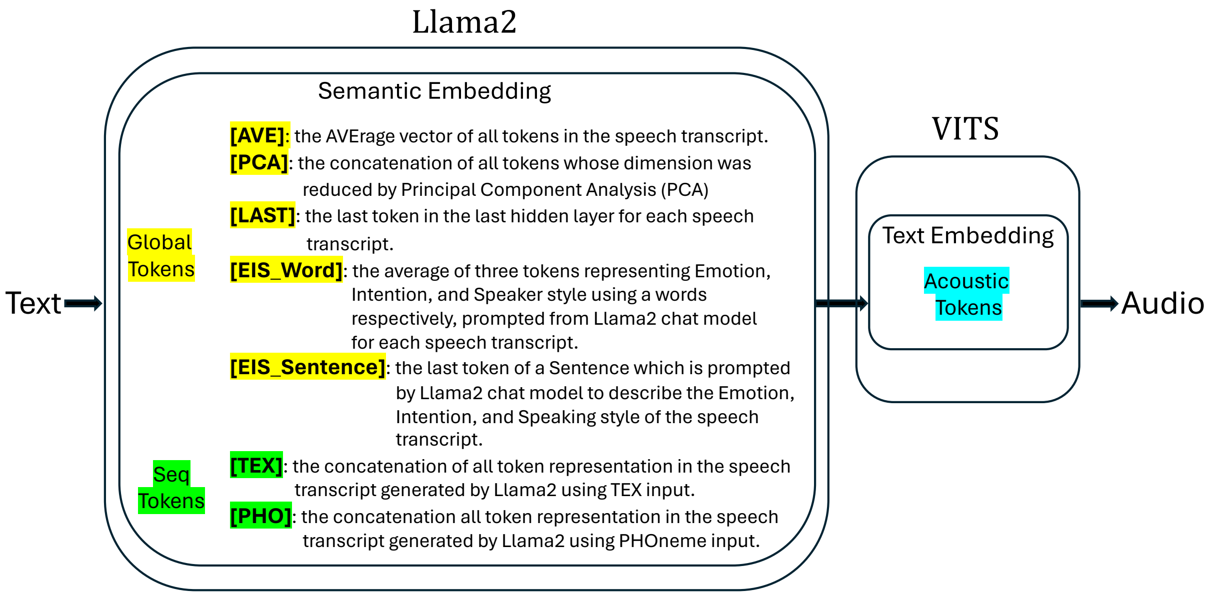

Figure 1: Pipeline of Llama-VITS

```
# Prompt for extracting [EIS_Word] token
# Build dialogue for each sentence
for sentence in sentences:
    dialogs=[
      [
        {"role": "system", "content": "Always answer within a word."},
        {"role": "user", "content": f"what is the emotion of the sentence: {sentence}"},
      ],
      [
        {"role": "system", "content": "Always answer within a word."},
        {"role": "user", "content": f"what is the intention of the sentence: {sentence}"},
      ],
      [
        {"role": "system", "content": "Always answer within a word."},
        {"role": "user", "content": f"what is the speaking style of the sentence: {sentence}"},
      ],
    ]
```

(a) Prompts used for extracting [EIS_Word] tokens

```
# Prompt for extracting [EIS_Sentence] token
# Build dialogue for each sentence
for sentence in sentences:
  dialogs = [
    [
      {
        "role": "system",
        "content": "Always answer within a sentence even if there are multiple requirements. \
            Never chat about yourself."
      },
      {
        "role": "user",
        "content": f"Describe the emotion, intention, and speaking style of the sentence \
            in an easy-to-understand sentence: {sentence}"
      },
    ]
  ]
```

(b) Prompts used for extracting [EIS_Sentence] tokens.

Figure 2: Prompts used for extracting EIS tokens

EmoV_DB_bea_sem are available at `https://github.com/xincanfeng/vitsgpt.git`.

2. Related Work

TTS technology has significantly advanced in learning acoustic features through the above structural evolution. However, comprehending and conveying semantics and emotions remain challenging. Since BERT-like LMs have demonstrated profound capabilities in understanding semantics through extensive pre-training on vast text corpora, some studies have integrated BERT-like LMs with TTS technology to enhance the expression in synthesized speech. Nonetheless, research on incorporating GPT-like LMs within TTS technology is notably scarce.

2.1. Text-To-Speech Models

TTS task aims to generate natural, fluent, and easily comprehensible speech. Traditional TTS systems, e.g., a Statistical Parametric Speech Synthesis (SPSS) system (Taylor, 2009), usually comprise multiple distinct components. These include a frontend module that converts text into linguistic features (such as duration and pitch), an acoustic model that maps these linguistic features to acoustic features, and a vocoder responsible for generating speech waveforms from the acoustic features. Over the past decades, the complexity of traditional models has been notable, attributed to their reliance on manually engineered features and the intricate communication required between modules.

Transitioning from Hidden Markov Models (HMM) based models (Black et al., 2007), through Deep Neural Networks (DNN) models (Zen et al., 2013), to Generative Adversarial Networks (GAN) based models (Saito et al., 2017), there has been a notable enhancement in voice quality, yet the architectural complexity remains significant.

The advent of end-to-end TTS models marks a significant milestone in TTS technology, increasingly reducing the distinction between synthesized speech and human voice. End-to-end models are capable of transforming raw text directly into final speech output, which not only streamlines the structural complexity of TTS systems and facilitates easier deployment but also significantly reduces the dependency on manual feature engineering, simplifying the training process. Moreover, they notably enhance the naturalness and intelligibility of the speech, thereby becoming the predominant architecture in TTS models. For instance, Char2Wav (Sotelo et al., 2017) introduces an attentive encoder-decoder framework for direct speech synthesis from text input. Tacotron (Wang et al., 2017) undertakes training from the ground up and directly predicts linear spectrograms. Furthermore, the speech produced by Tacotron2 (Shen et al., 2018) closely mirrors the natural human voice.

In the realm of end-to-end TTS models, many have adopted a non-autoregressive model architecture. This architecture enables parallel data processing, where the model's output generation does not depend on the output of the previous time step, thereby enhancing processing speed. It also circumvents the error accumulation issue inherent in traditional autoregressive models during audio generation, which significantly boosts TTS performance. FastSpeech (Ren et al., 2019) and its variants exemplify this trend. FastSpeech employs a transformer-based architecture to generate mel-spectrograms in parallel. Building on FastSpeech, FastPitch (Łańcucki, 2021) predicts pitch contours during inference, enabling the production of more expressive and high-quality speech. FastSpeech2 (Ren et al., 2022) further refines the model by incorporating explicit duration prediction and introducing pitch and energy as conditional inputs for optimization.

Previous non-autoregressive approaches typically involve distinct training phases for acoustic models and vocoders. However, VITS (Kim et al., 2021) introduces a more natural-sounding output compared to these two-stage systems through its one-stage parallel end-to-end architecture. Innovatively, VITS incorporates variational inference combined with normalizing flows and employs an adversarial training methodology. Due to VITS's exemplary performance across multiple TTS benchmarks, it has been selected as the foundational TTS model for our system.

2.2. Fine-tuning BERT-like models for TTS

While TTS models have increasingly advanced in replicating acoustic features, insufficient training data can hinder the model's ability to learn the emotional nuances of the same input across different contexts, thus limiting its expressive capabilities. Consequently, researchers have turned to leveraging the robust transfer learning capabilities of BERT-like models. Ultimately, TTS systems that incorporate pre-trained and fine-tuned BERT-like models have achieved better understandings of semantics and enhanced generated speech, marking a significant advancement in TTS technology.

Hayashi et al. (2019) utilized a pre-trained BERT model as an auxiliary input to enhance a Tacotron2-based TTS system, resulting in improved speech naturalness. Similarly, Yang et al. (2019) applied a pre-trained BERT to achieve enhanced front-end accuracy. Kenter et al. (2020) demonstrated that integrating a BERT model, pre-trained on extensive unlabeled data and fine-tuned for speech, into an RNN-based speech synthesis system enhances prosody. Kenter et al. (2020) specifically suggest updating the BERT network's parameters during the training of their RNN-based speech synthesis model, emphasizing the critical role of fine-tuning the BERT component for optimal outcomes. As prompt tuning draws wide attention in guiding text or image generation, PromptTTS (Guo et al., 2022) takes a prompt with both style and content descriptions as input to synthesize the corresponding speech. Specifically, PromptTTS consists of a style encoder and a content encoder to extract the corresponding representations from the prompt, and a speech decoder to synthesize speech according to the extracted style and content representations. Experiments show that PromptTTS can generate speech with precise style control and high speech quality.

In particular, Mukherjee et al. (2022) utilized a pre-trained BERT model to develop a text emotion classification model, employing the final hidden states of the initial [CLS] token as a comprehensive representation of the text. Researchers such as Kenter et al. (2020); Li et al. (2021); Abbas et al. (2022) have applied word-level BERT to capture the semantic and syntactic structure of sentences, thereby aiding TTS synthesis. Li et al. (2023) introduced a phoneme-level BERT, designed with a preliminary task of predicting corresponding graphemes in addition to regular masked phoneme predictions, to enhance the naturalness of speech synthesized from out-of-distribution (OOD) texts.

However, despite BERT's acknowledged capacity to provide detailed word importance, syntactic and semantic insights, and general knowledge (Hayashi et al., 2019; Kenter et al., 2020), its effectiveness is constrained by the particularities of fine-tuning approaches. Furthermore, BERT's inherent non-generative nature might limit its ability to account for information outside the immediate sentence context.

2.3. Integrating GPT-like models for TTS

Considering semantic understanding and expression capabilities, BERT is primarily utilized for comprehension tasks. In comparison, GPT excels not only in understanding text but also in generating natural and coherent text. Moreover, with the larger model parameters, GPT is particularly adept at zero-shot or few-shot learning, enabling its direct application to various tasks with little to no need for fine-tuning or structural modifications.

However, research on leveraging GPT-like models to aid semantic understanding in TTS systems is very limited. For instance, Stephenson et al. (2021) explores the potential of improving speech synthesis prosody with the GPT model by modifying text input. Such an approach potentially restricts TTS applications, as altering the input is often undesirable. Furthermore, the findings were not particularly significant.

In our study, we selected Llama2, a GPT-like LM, for integration into our TTS system, motivated by its technological advancements and potential for diverse applications. Llama2 stands out as one of the largest publicly accessible LMs, rivaling proprietary models such as GPT3.5 (OpenAI et al., 2024) and PaLM (540B) (Chowdhery et al., 2022), and surpasses other open-source alternatives like MPT[1] and Falcon (Almazrouei et al., 2023) in benchmark evaluations. Additionally, the novel architecture of Llama2 not only ensures enhanced security but also facilitates the extension of various downstream tasks (Touvron et al., 2023).

Related research that employs Llama2 in speech and other multimodal tasks (Radhakrishnan et al., 2023; Zhang et al., 2023), coupled with the ongoing efforts to reduce computing costs associated with Llama2[2], underscores the model's significant research interest and its promising prospects in multimodal applications.

3. Methodology

We propose leveraging semantic embeddings derived from a GPT-like LM to improve TTS synthesis. In our work, Llama2 is employed as the GPT-like model, as elaborated in Section §2.3, and VITS is utilized as the TTS model for generating audio from phoneme embeddings, as detailed in Section §2.1. In essence, we extract semantic embeddings E_s from the final hidden layer of Llama2 and integrate them with the original acoustic text embeddings E_a of VITS, forming enhanced text embeddings E_{as} for speech synthesis. Specifically, either a global token or a sequence of tokens is used to encapsulate the semantic attributes of an input sentence for varying objectives. The distinctions between these two token types are further explicated in Section §3.1.

[1] https://www.databricks.com/blog/mpt-30b

[2] https://huggingface.co/4bit/Llama-2-70b-chat-hf

3.1. Semantic Embeddings Derived from Llama2

For each input sentence s, we extract information from the final hidden layer before the output of Llama2. Different strategies are employed to create various tokens that serve as the semantic embedding for the sentence.

Let E_s denote the semantic embedding of sentence s, and $H^F_{\text{Llama}}(s)$ represent the output of the Llama2 model for sentence s at the final hidden layer F. Therefore, E_s can be expressed as:

$$E_s = H^F_{\text{Llama}}(s) \tag{1}$$

Here, $H^F_{\text{Llama}}(s)$ is a vector that encapsulates the semantic representation of sentence s after processing through all layers of the Llama2, culminating in the final layer.

Formulation for Global Tokens We explored five types of global tokens to represent the overarching semantic features of an input sentence, namely [AVE], [PCA], [LAST], [EIS_Word], and [EIS_Sentence], with each strategy employing a single token.

In the [AVE] strategy, the semantic token is derived by calculating the average of all tokens' output vectors for sentence s, formulated as:

$$E_s^{\text{AVE}} = \frac{1}{n}\sum_{i=1}^{n} H^F_{\text{Llama}}(s, i) \tag{2}$$

Here, E_s^{AVE} denotes the semantic token obtained using the [AVE] strategy, and $H^F_{\text{Llama}}(s, i)$ represents the output of the ith token of sentence s at the final hidden layer F of Llama2, with s comprising n tokens.

For the [PCA] strategy, we apply Principal Component Analysis to the output vectors of sentence s to extract principal components and rescale the mean of the PCA results according to the original data's value range. This rescaling ensures that the PCA-processed data maintains a scale consistent with the original data, preserving the relative importance of semantic information numerically. Formulated as:

$$E_s^{\text{PCA}} = \text{PCA\_rescale}\left(H^F_{\text{Llama}}(s)\right) \tag{3}$$

In the [LAST] strategy, the semantic token is obtained by selecting the last token from the output vector of sentence s, as shown in the formula:

$$E_s^{\text{LAST}} = H^F_{\text{Llama}}(s, n) \tag{4}$$

where $H^F_{\text{Llama}}(s,n)$ refers to the representation of the last token of sentence s after processing through all layers of Llama2 at the final layer.

In the [EIS_Word] and [EIS_Sentence] strategies, unlike the above approaches that utilize the sentence itself for representation, we derive the semantic representation of sentence s based on Llama2's comprehension u. Specifically, we employ prompts, as illustrated in 2a and 2b, respectively, to obtain Llama2's understanding of sentence s in terms of Emotion, Intention, and Speaking Style, denoted as u, and calculate the average of this understanding's representation to serve as the semantic token.

In the [EIS_Word] strategy, Llama2 is prompted to describe Emotion, Intention, and Speaking Style with three separate words, resulting in the following formula for the final semantic token:

$$\begin{aligned} E_s^{\text{EIS\_Word}} =& \frac{1}{m}\Big[\sum_i H^F_{\text{Llama}}(u_{\text{E}}, i) \\ &+ \sum_j H^F_{\text{Llama}}(u_{\text{I}}, j) + \sum_k H^F_{\text{Llama}}(u_{\text{S}}, k)\Big] \end{aligned} \tag{5}$$

where $u_{\text{E}}, u_{\text{I}}, u_{\text{S}}$ are the representations of Llama2's output expressing the sentence's Emotion, Intention, and Speaking Style at the final hidden layer, respectively, with i, j, k indicating the tokens of each output word, and m being the total number of these tokens.

In the [EIS_Sentence] strategy, Llama2 is guided to describe its understanding of the input sentence's Emotion, Intention, and Speaking Style with an easy-to-understand sentence, leading to the following formula for the final semantic token:

$$E_s^{\text{EIS\_Sentence}} = \frac{1}{m}\sum_{i=1}^{m} H^F_{\text{Llama}}(u_{\text{EIS}}, i) \tag{6}$$

where u_{EIS} is the representation of Llama2's output expressing the understanding of the original sentence at the final hidden layer, and m is the total number of tokens in this sentence representation.

Formulation for Sequential Tokens In the implementation of sequential tokens strategies, we focus on utilizing the complete set of tokens from the input sentence to represent its semantic information. Unlike the global token approaches, sequential tokens strategies encompass representations based on either text or phonemes, aiming to better align with the TTS model's potential emphasis on acoustic features. The mathematical representations for these two strategies are as follows:

Under the [TEX] strategy, we directly employ all tokens from the textual form of sentence s to represent its semantic information. If the output of sentence s at the final hidden layer F of Llama2 consists of n tokens, then the semantic token T_s^{TEX} is represented as a sequence:

$$E_s^{\text{TEX}} = \{H^F_{\text{Llama}}(s,1), H^F_{\text{Llama}}(s,2), \ldots, H^F_{\text{Llama}}(s,n)\} \tag{7}$$

In the [PHO] strategy, we consider the complete set of tokens from the phonemic form. Here, s_{pho} denotes the phonemic representation of sentence s. If the output of s_{pho} at the final hidden layer F of Llama2 comprises m tokens, then the semantic token T_s^{PHO} is represented as a sequence:

$$\begin{aligned} E_s^{\text{PHO}} =& \{H^F_{\text{Llama}}(s_{\text{pho}}, 1), \\ & H^F_{\text{Llama}}(s_{\text{pho}}, 2), \ldots, H^F_{\text{Llama}}(s_{\text{pho}}, m)\} \end{aligned} \tag{8}$$

In both strategies, $H^F_{\text{Llama}}(s,i)$ and $H^F_{\text{Llama}}(s_{\text{pho}}, i)$ respectively represent the outputs of the ith token of sentence s in its textual and phonemic forms at the final hidden layer F of Llama2. This representation allows the TTS model to leverage the complete semantic information of a sentence, whether based on text or phonemes.

3.2. Fusing Semantic Embedding with Acoustic Embedding

To align the dimensions of semantic embedding extracted from Llama2, denoted as E_s, with the acoustic embeddings from VITS, denoted as E_a, we employ a linear projection. The original dimension of E_s, d_{Llama}, is projected to match the dimension of VITS acoustic embedding, d_{VITS}, using a linear transformation matrix W of dimensions $d_{\text{VITS}} \times d_{\text{Llama}}$. The projected semantic embedding, E'_s, is calculated as follows:

$$E'_s = W \cdot E_s \tag{9}$$

Fusing Global Embedding with Acoustic Embedding To obtain an embedding E_{as} that integrates both semantic and acoustic information, for global tokens, we simply add the dimensionally unified global embedding to VITS's acoustic embedding, as shown in the equation:

$$E_{as} = E_a + E'_s \tag{10}$$

Fusing Sequential Embeddings to Enhance Text Embeddings We utilize the Scaled Dot-Product Attention mechanism to merge sequential embeddings with VITS's original acoustic embedding to gain enhanced embedding E_{as}, which can be described by the following mathematical formulas:

First, calculate the attention scores A:

$$A = \frac{q \cdot k^T}{\gamma} \tag{11}$$

where q is the acoustic embedding E_a in VITS with dimensions $[b, t, d]$; k and v denotes the semantic embedding E'_s from Llama2, also with dimensions $[b, t, d]$; b is the batch size, t is the sequence length, and d is the embedding dimension; γ is temperature for scaling. k^T denotes the transpose of k, transforming k from $[b, t, d]$ to $[b, d, t]$ for matrix multiplication. The resulting A has dimensions $[b, t, t]$.

If a source mask or target mask is present, a masking operation is applied, setting the attention scores at masked positions to a very low value (e.g., $-6e4$) to nearly eliminate their weight contribution in the subsequent softmax step.

Next, apply the softmax function and dropout to the attention scores, obtaining the final attention weights W_{attn}:

$$W_{\text{attn}} = \text{Dropout}(\text{Softmax}(A)) \tag{12}$$

Finally, the output E_{as} is calculated by weighting v with the attention weights:

$$E_{as} = W_{\text{attn}} \cdot v$$

The output E_{as}, viewed as text embedding fused with semantic information, has dimensions $[b, t, d]$ that match those of q.

4. Experiments

4.1. Experimental Settings

We propose Llama-VITS which uses semantic tokens derived from Llama2 to enhance acoustic embedding in VITS for better TTS performance. To show the effectiveness of our method, we experimented with two baseline models. In the ORI-VITS baseline, we use the original VITS without external semantic information. In the BERT-VITS baseline, we extract various semantic tokens according to former research introduced in Section § 2.2. Specifically, we use the [CLS] token of BERT as the global token. To form the baseline of the sequential token in BERT, we use all the tokens in the sentence trained by text or phoneme, named [BERT_TEX] and [BERT_PHO], respectively. In our proposed Llama-VITS, we derive global token [AVE], [LAST], [PCA], [EIS_Word], and [EIS_Sentence], and sequential tokens [TEX] and [PHO] from Llama2, corresponding to those in BERT-VITS.

We use Llama2 (13b) to generate semantic embeddings of dimension 5120. [CLS] and [BERT_TEX] tokens are extracted from BERT-base-uncased model which has a parameter size of 110M that generates token embedding of 768 dimensions. [BERT_PHO] token is extracted from BERT-x-phone-base model whose parameter size is 88M to generate token embedding of 768 dimensions.

Global Token Extraction In our proposed Llama-VITS, global strategy [LAST] only uses the last token in the final hidden layer of Llama2 for each sentence. [AVE] uses the average of all tokens for each sentence. [PCA] uses the concatenation of all tokens whose dimension was reduced by Principal Component Analysis (PCA). [EIS_Word] and [EIS_Sentence] use the average of tokens for an answer, which is formed in three words or a sentence by prompts shown in Figure 2, to describe the Emotion, Intention, and Speaking style of the transcript.

In BERT-VITS baseline, global strategy [CLS] only uses the first token from the BERT-base-uncased model for each input sentence.

Sequential Token Extraction In our proposed Llama-VITS, sequential strategy [TEX] concatenates the sequence of tokens in a sentence generated by Llama2 using text input. [PHO] concatenates the sequence of tokens of a sentence generated by Llama2 using phonemic input.

In the baseline BERT-VITS, sequential strategy [BERT_TEX] concatenates all the tokens in a sentence extracted from BERT-base-uncased model. [BERT_PHO] concatenates all the tokens in a sentence extracted from BERT-x-phone-base model.

Datasets We utilized full LJSpeech, 1-hour LJSpeech, and EmoV_DB_bea_sem dataset for our experimental verification. LJSpeech [3] comprises 24 hours recorded of English speech by single female speaker, where we evaluate how the tokens extracted from Llama2 can help improve the speech naturalness. Besides the full LJSpeech dataset, we also randomly filtered 1-hour LJSpeech which contains only 1-hour records as an ablation study to show how dataset size influences the results. EmoV_DB (Adigwe et al., 2018) is a database of emotional speech that contains data for male and female actors in English and a male actor in French. EmoV_DB covers 5 emotion classes, amused, angry, disgusted, neutral, and sleepy. To factor out the effect of different speakers, we filtered the original EmoV_DB dataset into the speech of a specific female English speaker, bea. Then we use Llama2 to predict the emotion label of the transcript chosen from the above 5 emotion classes, and select the audio samples which has the same predicted emotion. The filtered dataset contains 22.8-min records for training. We named the filtered dataset EmoV_DB_bea_sem and investigated how the semantic tokens from Llama2 behave in naturalness and expressiveness on this filtered dataset. Refer to Appendix 11 for more dataset statistics.

[3] https://keithito.com/LJ-Speech-Dataset/

full LJSpeech							
Model	**Semantic Token**			**Evaluation 100k-step (batchsize=64)**			
						ASR	
	Type	**Fuse**	**Token**	**UTMOS**	**MCD**	**CER**	**WER**
ORI-VITS	-	-	-	4.19 ± 0.05	7.32 ± 0.61	6.2	16.5
BERT-VITS	gl	add	CLS	4.10 ± 0.06	7.38 ± 0.60	6	16.5
	seq	att	BERT_TEX	**4.22 ± 0.05**	**7.27 ± 0.61**	**5.9**	15.9
			BERT_PHO	4.08 ± 0.08	7.28 ± 0.62	**5.9**	**15.7**
Llama-VITS	gl	add	AVE	**4.21 ± 0.05**	7.30 ± 0.66	5.9	16
			PCA	4.19 ± 0.05	**7.23 ± 0.61**	**5.8**	**15.8**
			LAST	**4.21 ± 0.05**	7.39 ± 0.63	**5.8**	16.2
			EIS_Word	4.16 ± 0.06	7.32 ± 0.62	**5.8**	16.1
			EIS_Sentence	**4.21 ± 0.04**	7.26 ± 0.64	5.9	16.2
	seq	att	TEX	4.13 ± 0.06	7.48 ± 0.68	6	16
			PHO	4.19 ± 0.05	7.33 ± 0.65	6	16.2

1-hour LJSpeech							
Model	**Semantic Token**			**Evaluation 100k-step (batchsize=64)**			
						ASR	
	Type	**Fuse**	**Token**	**UTMOS**	**MCD**	**CER**	**WER**
ORI-VITS	-	-	-	4.02 ± 0.08	7.47 ± 0.63	6	16.1
BERT-VITS	gl	add	CLS	4.02 ± 0.07	**7.39 ± 0.62**	**6**	**16.3**
	seq	att	BERT_TEX	3.91 ± 0.08	7.54 ± 0.60	6.2	16.6
			BERT_PHO	**4.05 ± 0.07**	7.40 ± 0.65	6.2	**16.3**
Llama-VITS	gl	add	AVE	**4.10 ± 0.07**	7.37 ± 0.63	**6**	**16.4**
			PCA	4.04 ± 0.06	7.39 ± 0.60	6.4	16.7
			LAST	3.98 ± 0.08	7.38 ± 0.62	6.1	16.6
			EIS_Word	3.99 ± 0.08	7.37 ± 0.60	6.2	16.6
			EIS_Sentence	4.01 ± 0.08	**7.36 ± 0.59**	**6**	**16.4**
	seq	att	TEX	4.02 ± 0.08	7.55 ± 0.62	6.2	16.7
			PHO	3.95 ± 0.08	7.42 ± 0.57	6.7	17.3

EmoV_DB_bea_sem								
Model	**Semantic Token**			**Evaluation 150k-step (batchsize=16)**				
							ASR	
	Type	**Fuse**	**Token**	**ESMOS**	**UTMOS**	**MCD**	**CER**	**WER**
ORI-VITS	-	-	-	3.06 ± 0.08	3.61 ± 0.08	7.06 ± 1.19	4.5	18.5
BERT-VITS	gl	add	CLS	**3.02 ± 0.07**	3.50 ± 0.11	**7.11 ± 1.13**	5.4	20.5
	seq	att	BERT_TEX	2.92 ± 0.08	**3.61 ± 0.18**	7.21 ± 1.15	**4.4**	**18.7**
			BERT_PHO	2.96 ± 0.08	3.50 ± 0.15	7.21 ± 1.26	5	19.6
Llama-VITS	gl	add	AVE	3.01 ± 0.08	3.60 ± 0.13	7.15 ± 1.18	4.6	18.3
			PCA	-	3.57 ± 0.13	**7.13 ± 1.24**	4.8	19.2
			LAST	-	3.55 ± 0.11	**7.13 ± 1.17**	**4.3**	**17.4**
			EIS_Word	-	**3.61 ± 0.12**	7.20 ± 1.23	4.7	19.6
			EIS_Sentence	-	3.55 ± 0.12	**7.13 ± 1.23**	5	18.5
	seq	att	TEX	**3.22 ± 0.07**	3.40 ± 0.21	7.21 ± 1.35	4.5	18.3
			PHO	2.98 ± 0.07	3.52 ± 0.15	7.27 ± 1.30	4.7	18.5

Table 1: Results on full LJSpeech, 1-hour LJSpeech, and EmoV_DB_bea_sem dataset, respectively. Bold text show the best result in each model. Note that, for global tokens extracted from Llama2, ESMOS metric is only evaluated on [AVE] token to save human cost.

Implementation, Hyper-parameters, Training Our Llama-VITS system was built on the VITS (Kim et al., 2021) framework using its original implementation[4], augmented with semantic embeddings derived from Llama2 (Touvron et al., 2023) using its original implementation[5]. For training LJSpeech, we use the public configs in the original implementation of VITS. For EmoV_DB_bea_sem, we use the same config as LJSpeech but changed batch size from 64 to 16 since this dataset is much smaller. Besides implementing our proposed Llama-VITS, we extracted corresponding semantic tokens [CLS], [BERT_TEX] from BERT uncased base model[6] and [BERT_PHO] from BERT pre-trained on phoneme[7] for comparison.

In comparing the experimental results, we choose 100k-step results on both full LJSpeech and 1-hour LJSpeech datasets since they are rather large. On EmoV_DB_bea_sem, we used the pre-trained checkpoint of LJSpeech on 100k-step and compare the fine-tuning results on EmoV_DB_bea_sem at 150k-step since it is rather small.

Evaluation Metrics Both subjective and objective metrics are implemented for a comprehensive evaluation. In subjective evaluation, we conduct Emotion Similarity Mean Opinion Score (ESMOS) (Zhu et al., 2023) experiments to evaluate emotion similarity for EmoV_DB_bea_sem. In the subjective evaluation, we compared [AVE], [TEX] and [PHO] strategies in our Llama-VITS with the corresponding token [CLS], [BERT_TEX] and [BERT_PHO] extracted from different BERT models and the baseline ORI-VITS who does not contain semantic tokens, with the ground truth samples GT.

In evaluating ESMOS, we randomly chose 5 samples from the total 51 test samples proportionally divided by us and received 100 test results from different speakers on Amazon Mechanical Turk. The result significance level is thus 500. Each participant is asked to give a score on emotion similarity compared with ground truth in a 5-scale: Excellent Match 5, Good Match 4, Fair Match 3, Poor Match 2, Bad Match 1[8].

In objective evaluation, we utilize UTokyo-SaruLab Mean Opinion Score (UTMOS) (Saeki et al., 2022), Mel-Cepstral Distortion (MCD), and speech recognition performance measured by Character Error Rate (CER) and Word Error Rate (WER). UTMOS is a MOS prediction network using speech samples from previous Blizzard Challenges and Voice Conversion Challenges, which has reached the best performance in VoiceMOS Challenge 2022. We evaluate objective intelligibility by using Whisper-large (Radford et al., 2022). For calculating UTMOS, we use the implementation in SpeechMOS[9]. For calculating MCD and ASR, we use the evaluation implementation[10] of ESPnet (Hayashi et al., 2020, 2021).

5. Experimental Results

We evaluated our proposed Llama-VITS along with baselines ORI-VITS and BERT-VITS models on three distinct datasets: the full LJSpeech, the 1-hour LJSpeech, and EmoV_DB_bea_sem. The experimental outcomes provide a comprehensive understanding of the model performance and the impact of semantic tokens selection. A summary of these results is articulated below and can be referenced in Table 1.

5.1. Results on full LJSpeech

The ORI-VITS baseline, achieving a UTMOS of 4.19 ± 0.05, an MCD of 7.32 ± 0.61, a CER of 6.2, and a WER of 16.5.

Enhancements were observed with the BERT-VITS baseline. Specifically, BERT-VITS with [BERT_TEX] semantic tokens demonstrated superior performance in UTMOS (4.22 ± 0.05) and MCD (7.27 ± 0.61), indicating improved speech quality and reduced mel-cepstral distortion. Additionally, a reduced CER of 5.9 and WER of 15.9 were noted, highlighting enhanced automatic speech recognition accuracy.

Our proposed Llama-VITS, integrating various global and sequential semantic tokens, displayed competitive performance. The [PCA] strategy stood out, achieving an MCD of 7.23 ± 0.61, indicating optimal mel-cepstral distortion. The [EIS_Sentence], [AVE], and [LAST] tokens yielded a top-tier UTMOS of $4.21 \pm 0.04/0.05$, underscoring their effectiveness in enhancing perceived speech quality.

5.2. Results on 1-hour LJSpeech

In the more challenging 1-hour LJSpeech dataset, all models experienced a slight performance decrease, an expected outcome given the reduced training data size.

[4] https://github.com/jaywalnut310/vits
[5] https://github.com/facebookresearch/llama
[6] https://huggingface.co/bert-base-uncased
[7] https://huggingface.co/vinai/xphonebert-base
[8] Note that in the ESMOS experiments, participants are asked to ignore the speakers' voice, style, and audio quality and only consider the emotiveness of the speech.
[9] https://github.com/tarepan/SpeechMOS
[10] https://github.com/espnet/espnet

BERT-VITS baseline with [CLS] tokens exhibited notable MCD performance (7.39 ± 0.62), while the [BERT_PHO] excelled in UTMOS (4.05 ± 0.07), reflecting enhanced speech naturalness and reduced mel-cepstral distortion.

Llama-VITS with [AVE] tokens achieved the highest UTMOS (4.10 ± 0.07), while [EIS_Sentence] tokens resulted in the most favorable MCD (7.36 ± 0.59), illustrating the model's versatility and efficacy in different token configurations.

5.3. Results on EmoV_DB_bea_sem

On this even more challenging dataset, a small improvement observed in BERT-VITS only exists in the [BERT_TEX] with a CER of 4.4.

While our proposed Llama-VITS displayed notable enhancements. The [TEX] strategy achieves an ESMOS of 3.22 ± 0.07, indicating much more emotiveness. The [LAST] yielded the best performance on CER of 4.3 and WER of 17.4, other strategies also perform better than or comparable to BERT-VITS, underscoring its effectiveness in enhancing perceived speech expressiveness.

5.4. Analysis

Speaking of the strengths of different tokens, BERT-based tokens generally contribute to improving MCD and ASR scores, indicating the enriched semantic understanding translated to speech quality. Tokens of Llama-VITS exhibited a balanced performance across all metrics, with specific token configurations excelling in particular aspects. For instance, [PCA] token emerged as a strong contender in reducing MCD, [AVE] enhanced the UTMOS scores, [TEX] had superior performance to improve ESMOS score.

In individual comparisons, Llama-VITS's five global tokens generally outperformed BERT-VITS on the UTMOS metric for naturalness. In the ESMOS metric for emotional expression, Llama-VITS's two sequential tokens also generally surpassed BERT-VITS, particularly the [TEX] token. Therefore, we can infer that GPT-like LMs may have greater potential for TTS tasks than BERT-like models.

Further, our results reflect different patterns of gains from GPT-like and BERT-like models in TTS tasks. For instance, in the UTMOS naturalness metric, Llama-VITS's global tokens often outperformed sequential tokens, which is the opposite for BERT-VITS; in the ESMOS emotion metric, Llama-VITS's sequential token [TEX] significantly outperformed other tokens, while for BERT-VITS, global tokens performed better.

Overall, Llama-VITS showed a different pattern in UTMOS compared to BERT-VITS, and superior performance in ESMOS. These results highlight the potential for further exploration of semantic token types and fusion methods to achieve more significant enhancements in speech synthesis, particularly in scenarios constrained by limited and complex training data.

6. Discussions

In this section, we discuss factors influencing current outcomes. Based on this discussion, we also point out the directions for future work in Appendix 11.

6.1. GPT-like vs BERT-like

Initial observations from our experiments indicate that, even without any fine-tuning of Llama2, Llama-VITS significantly outperforms both BERT-VITS and ORI-VITS in terms of emotional expressiveness. This finding opens up avenues for future research into emotive TTS tasks.

Furthermore, a comparison between BERT-VITS and Llama-VITS highlights their distinct performance traits. BERT-VITS, leveraging deep contextual embeddings, provides profound semantic insights yet encounters challenges in customization and adaptability across a range of TTS tasks. Conversely, Llama-VITS can provide a more versatile and adaptable approach, with its array of token types demonstrating particular advantages across various evaluation metrics.

6.2. Semantic Token Strategy

The varying effectiveness of distinct semantic tokens underscores the importance of careful selection and integration tailored to the particular goals of TTS systems. Optimizing the type of token and method of fusion can be instrumental in enhancing aspects such as speech naturalness, emotional expressiveness, Mel Cepstral Distortion (MCD), or Automatic Speech Recognition (ASR) performance.

7. Conclusion

In summary, this study exemplifies a significant stride towards optimized TTS synthesis by integrating semantic tokens, leveraging the strengths of Llama-VITS. Our findings, validated by comprehensive experiments on the LJSpeech and EmoV_DB_bea_sem datasets, underscore the pivotal role of semantic embeddings in enhancing speech quality, naturalness, and emotiveness. The adaptability and efficacy of Llama-VITS, especially, open new vistas for customized and context-sensitive TTS applications.

8. Limitations

Compared with our baseline which uses different BERT models, we only tested our method using Llama2. As Kenter et al. (2020) indicate for their BERT-based TTS model, small BERT models work better than big ones, but the parameter size of our proposed GPT-based TTS influence is yet studied by our research. Although BERT-based TTS models are normally finetuned on speech tasks to provide more explicit acoustic information for TTS, we didn't try designing prompts to generate acoustic features and only studied how general semantic information can help. Our experiments were conducted only on clean datasets with limited size, and the effect on more complex datasets is to be further explored. The integration of Llama2's embeddings introduces additional computational costs, potentially limiting real-time applications.

9. Acknowledgements

This research was conducted with the support of team members who contributed to varying extents. Particular gratitude is owed to Koichi Miyazaki for his support regarding foundational knowledge and his assistance in implementing the subjective evaluation during the latter phases of the project.

10. Bibliographical References

Ammar Abbas, Thomas Merritt, Alexis Moinet, Sri Karlapati, Ewa Muszynska, Simon Slangen, Elia Gatti, and Thomas Drugman. 2022. Expressive, variable, and controllable duration modelling in tts.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models.

Alan W Black, Heiga Zen, and Keiichi Tokuda. 2007. Statistical parametric speech synthesis. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–1229–IV–1232.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023a. Audiolm: a language modeling approach to audio generation.

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Xueyuan Chen, Shun Lei, Zhiyong Wu, Dong Xu, Weifeng Zhao, and Helen Meng. 2022. Unsupervised multi-scale expressive speaking style modeling with hierarchical context information for audiobook speech synthesis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7193–7202, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani

Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2022. Prompttts: Controllable text-to-speech with text descriptions.

Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Shubham Toshniwal, and Karen Livescu. 2019. Pre-trained text embeddings for enhanced text-to-speech synthesis. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-September:4430–4434.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Tom Kenter, Manish Kumar Sharma, and Rob Clark. 2020. Improving prosody of rnn-based english text-to-speech synthesis by incorporating a bert model.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech.

Yinghao Aaron Li, Cong Han, Xilin Jiang, and Nima Mesgarani. 2023. Phoneme-level bert for enhanced prosody of text-to-speech with grapheme predictions.

Zhu Li, Yuqing Zhang, Mengxi Nie, Ming Yan, Mengnan He, Ruixiong Zhang, and Caixia Gong. 2021. Improving prosody for unseen texts in speech synthesis by utilizing linguistic information and noisy data.

Songxiang Liu, Yuewen Cao, and Helen Meng. 2020. Emotional voice conversion with cycle-consistent adversarial network.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2022. P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 61–68, Dublin, Ireland. Association for Computational Linguistics.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of gpt-3 in grammatical error correction: A study on performance and controllability in prompt-based methods.

Arijit Mukherjee, Shubham Bansal, Sandeepkumar Satpal, and Rupesh Mehta. 2022. Text aware Emotional Text-to-speech with BERT. In *Proc. Interspeech 2022*, pages 4601–4605.

Eliya Nachmani, Alon Levkovitch, Yifan Ding, Chulayuth Asawaroengchai, Heiga Zen, and Michelle Tadmor Ramanovich. 2023. Translatotron 3: Speech to speech translation with monolingual data.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez-Cabrero, and Jesper N. Tegner. 2023. Whispering llama: A cross-modal generative error correction framework for speech recognition.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2022. Fastspeech 2: Fast and high-quality end-to-end text to speech.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: Fast, robust and controllable text to speech.

Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2017. Statistical parametric speech synthesis incorporating generative adversarial networks.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions.

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

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jose Sotelo, Soroush Mehri, Kundan Kumar, Joo Felipe Santos, Kyle Kastner, Aaron C. Courville, and Yoshua Bengio. 2017. Char2wav: End-to-end speech synthesis. In *International Conference on Learning Representations*.

Brooke Stephenson, Thomas Hueber, Laurent Girin, and Laurent Besacier. 2021. Alternate endings: Improving prosody for incremental neural tts with predicted future text input.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Paul Taylor. 2009. *Text-to-Speech Synthesis*. Cambridge University Press.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lifu Tu, Caiming Xiong, and Yingbo Zhou. 2022. Prompt-tuning can be much better than fine-tuning on cross-lingual understanding with multilingual language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5478–5485, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Thomas Vakili, Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Downstream task performance of BERT models pre-trained using automatically de-identified clinical data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4245–4252, Marseille, France. European Language Resources Association.

Xingchen Wan, Ruoxi Sun, Hanjun Dai, Sercan Arik, and Tomas Pfister. 2023. Better zero-shot reasoning with self-adaptive prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3493–3514, Toronto, Canada. Association for Computational Linguistics.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023. Neural codec language models are zero-shot text to speech synthesizers.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. pages 4006–4010.

Te-Lin Wu, Satwik Kottur, Andrea Madotto, Mahmoud Azab, Pedro Rodriguez, Babak Damavandi, Nanyun Peng, and Seungwhan Moon. 2023. SIMMC-VR: A task-oriented multimodal dialog dataset with situated and immersive VR streams. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6273–6291, Toronto, Canada. Association for Computational Linguistics.

Bing Yang, Jiaqi Zhong, and Shan Liu. 2019. Pre-Trained Text Representations for Improving

Front-End Text Processing in Mandarin Text-to-Speech Synthesis. In *Proc. Interspeech 2019*, pages 4480–4484.

Heiga Zen, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7962–7966.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding.

Xinfa Zhu, Yi Lei, Kun Song, Yongmao Zhang, Tao Li, and Lei Xie. 2023. Multi-speaker expressive speech synthesis via multiple factors decoupling.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction.

11. Language Resource References

Adaeze Adigwe and Noé Tits and Kevin El Haddad and Sarah Ostadabbas and Thierry Dutoit. 2018. *The Emotional Voices Database: Towards Controlling the Emotion Dimension in Voice Generation Systems*.

Tomoki Hayashi and Ryuichi Yamamoto and Katsuki Inoue and Takenori Yoshimura and Shinji Watanabe and Tomoki Toda and Kazuya Takeda and Yu Zhang and Xu Tan. 2020. *ESPnet-TTS: Unified, Reproducible, and Integratable Open Source End-to-End Text-to-Speech Toolkit*.

Tomoki Hayashi and Ryuichi Yamamoto and Takenori Yoshimura and Peter Wu and Jiatong Shi and Takaaki Saeki and Yooncheol Ju and Yusuke Yasuda and Shinnosuke Takamichi and Shinji Watanabe. 2021. *ESPnet2-TTS: Extending the Edge of TTS Research*.

Takaaki Saeki and Detai Xin and Wataru Nakata and Tomoki Koriyama and Shinnosuke Takamichi and Hiroshi Saruwatari. 2022. *UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022*.

Appendix A. Dataset Statistics

Appendix B. Future Work

Given the promising potential of incorporating advanced semantic tokens to transform the TTS synthesis landscape, enhancing speech naturalness, emotiveness, and overall quality, future research could focus on investigating various token types, refining fusion strategies, and assessing performance across an expanded range of datasets and contexts.

Furthermore, with the rapid evolution of LM architectures and advancements in quantization techniques, a wider array of LM options has become feasible. For instance, Mixtral(Jiang et al., 2024), a Sparse Mixture of Experts (SMoE) LM, is considered comparable in capability to Llama2.

Revisiting the initial rationale for selecting GPT-like models, given that the effectiveness of Llama-VITS in improving speech synthesis performance has been established, it is conceivable to leverage the more adaptable fine-tuning or prompting features of GPT-like models over BERT-like models. This approach could explore utilizing GPT-like models for enhanced control over TTS synthesis tasks and refining fine-tuning or prompting methodologies.

Looking ahead, examining the dynamics of various TTS models when combined with semantic tokens could also be pivotal in unlocking the full potential of semantic-enhanced TTS systems.

Dataset	Train	Dev	Test	Total	Emotions
full LJSpeech	22.8hour	10.6min	54.5min	24hour	neutral
1-hour LJSpeech	1hour	10.6min	54.5min	2hour	neutral
EmoV_DB_bea_sem	15min	4min	3.8min	22.8min	neutral, amused, angry, disgusted, sleepy

Table 2: Dataset statistics. Note that, the LJSpeech dataset, a public-domain speech dataset recorded by a single female speaker, necessitated the filtration of the original EmoV_DB dataset to isolate recordings by a single female speaker, Bea, resulting in our creation of the EmoV_DB_bea dataset. This process was undertaken to facilitate a more accurate comparison. Additionally, the EmoV_DB_bea dataset underwent further refinement by selecting audio recordings whose expressed emotions aligned with those identified in the transcripts by Llama2, culminating in the EmoV_DB_bea_sem dataset. This dataset enables the training of models to recognize and learn emotional expressions.