

Article

BERTIVITS: The Posterior Encoder Fusion of Pre-Trained Models and Residual Skip Connections for End-to-End Speech Synthesis

Zirui Wang [†] , Minqi Song [†] and Dongbo Zhou ^{*} 

Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China; wangzirui0426@mails.ccnu.edu.cn (Z.W.); pisesminqi@mails.ccnu.edu.cn (M.S.)

* Correspondence: zhoudongbo@ccnu.edu.cn

[†] These authors contributed equally to this work.

Abstract: Enhancing the naturalness and rhythmicity of generated audio in end-to-end speech synthesis is crucial. The current state-of-the-art (SOTA) model, VITS, utilizes a conditional variational autoencoder architecture. However, it faces challenges, such as limited robustness, due to training solely on text and spectrum data from the training set. Particularly, the posterior encoder struggles with mid- and high-frequency feature extraction, impacting waveform reconstruction. Existing efforts mainly focus on prior encoder enhancements or alignment algorithms, neglecting improvements to spectrum feature extraction. In response, we propose BERTIVITS, a novel model integrating BERT into VITS. Our model features a redesigned posterior encoder with residual connections and utilizes pre-trained models to enhance spectrum feature extraction. Compared to VITS, BERTIVITS shows significant subjective MOS score improvements (0.16 in English, 0.36 in Chinese) and objective Mel-Cepstral coefficient reductions (0.52 in English, 0.49 in Chinese). BERTIVITS is tailored for single-speaker scenarios, improving speech synthesis technology for applications like post-class tutoring or telephone customer service.

Keywords: pre-trained model; text to speech; neural TTS; speech synthesis; end-to-end model



Citation: Wang, Z.; Song, M.; Zhou, D. BERTIVITS: The Posterior Encoder Fusion of Pre-Trained Models and Residual Skip Connections for End-to-End Speech Synthesis. *Appl. Sci.* **2024**, *14*, 5060. <https://doi.org/10.3390/app14125060>

Academic Editor: Christos Bouras

Received: 16 May 2024

Revised: 2 June 2024

Accepted: 5 June 2024

Published: 10 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, deep learning technology has made remarkable progress in the field of speech synthesis. Traditional speech synthesis systems typically rely on separate modules, such as acoustic models and vocoders, which are designed based on manually crafted features and rules. However, these systems often have difficulty capturing subtle variations and voice characteristics in speech, limiting the naturalness and fluency of the generated speech.

Against this backdrop, the introduction of end-to-end speech synthesis methods has brought new hopes to the field of speech synthesis. Compared to traditional systems, end-to-end speech synthesis methods learn the entire synthesis process as a unified model, greatly simplifying system design and enhancing the performance of speech synthesis.

However, end-to-end speech synthesis still encounters several challenges that urgently demand resolution, including prolonged training times and unnatural prosodic expressions. As we shift our focus towards similar sequence-to-sequence tasks encountered in the realm of Natural Language Processing (NLP), we observe that with the emergence of transformer [1], attention mechanism [1], and subsequent models such as BERT [2] and its derivatives like RoBERTa [3], substantial advancements have been made in the field of natural language processing. These models, trained on large-scale corpora, exhibit the capability to generate high-quality language representations and have demonstrated remarkable performance across various downstream tasks. Their advantages encompass contextual awareness, transfer learning capability, unsupervised learning proficiency, and profound understanding of language. In the realm of speech synthesis, we can draw

upon the principles and methodologies of models like BERT. By devising pre-training models, engaging in transfer learning, and incorporating contextual information, we aim to further augment the performance and efficacy of speech synthesis models. These endeavors hold promise for making significant contributions to the advancement of speech synthesis technology and the progression of intelligent speech interaction systems.

Therefore, our focus has been directed towards models such as HuBERT [4] and wav2vec2.0 [5], which have demonstrated significant improvements in speech recognition performance they were when pre-trained on the LibriSpeech [6] dataset. The objective of extracting hidden representations from speech preprocessing tasks aligns with that of speech recognition in end-to-end speech synthesis. Consequently, we propose that integrating pre-trained models into existing speech synthesis frameworks would substantially enhance the naturalness and prosody of the synthesized speech. For instance, in the VITS [7] model, XPhoneBERT [8] can be employed as a replacement for the text encoder segment. XPhoneBERT shares a similar structure to BERT, albeit differing in input; while BERT's input is a sequence of textual tokens, XPhoneBERT's input comprises a sequence of phonemes. By substituting VITS's basic text encoder with a pre-trained phoneme encoder, the model's understanding of textual input can be markedly enhanced, thereby further amplifying the expressive capabilities of speech synthesis.

Therefore, this paper aims to further explore how to leverage pre-trained models to enhance the quality and effectiveness of speech synthesis within the existing end-to-end speech synthesis framework VITS. We selected the pre-trained model HuBERT as our research subject and introduced it into the posterior encoder component of the VITS model. By combining the robust speech recognition capabilities of HuBERT with the end-to-end structure of VITS, we anticipate augmenting the model's proficiency in rhythm and prosody performance, thereby generating more natural and fluent speech. Through these enhancements, we aspire to advance both the quality and efficacy of speech synthesis within the current end-to-end speech synthesis framework while contributing to the future developments in intelligent speech interaction systems.

2. Related Works

End-to-end speech synthesis has seen significant advancements with the development of deep learning technologies. This literature review aims to explore recent progress and identify key areas for further research, particularly focusing on improving the naturalness and rhythmicity of synthesized speech.

2.1. Comprehensive Literature Search

To compile relevant studies, a comprehensive literature search was conducted using databases such as IEEE Xplore, Google Scholar, and ACM Digital Library. Keywords included "end-to-end speech synthesis", "VITS model", "speech naturalness", and "rhythmicity". Studies were selected based on their relevance, publication date (within the recent five years), and contributions to the field.

2.2. Classification of Works

The reviewed literature was classified based on the following criteria:

Publication Year: Focus on recent advancements from the recent five years.

Research Methodology: Empirical studies, theoretical analyses, and experimental results.

Technologies Covered: VITS, WaveNet, Tacotron, HuBERT, and other prominent models.

Geographic Areas: Research conducted in various regions to understand global advancements.

2.3. Critical Analysis of the Literature

2.3.1. Enhancements in Emotional and Prosodic Features

Zhang et al. proposed a speech emotion recognition method [9] that integrates energy frame time-frequency fusion based on differences in speech prosodic features, showing

improvements in weighted accuracy (WA) and unweighted accuracy (UA) metrics on the IEMOCAP [10] dataset. Further enhancements in the Tacotron [11] model by Zhang et al. involved modifying prosodic parameters [12] to boost emotional expressiveness.

Cui Lin et al. developed a parallel hybrid model incorporating a fusion multi-head attention mechanism [13], achieving high accuracy on the RAVDESS [14] and EMO-DB [15] datasets. The MAHCNN [16] model demonstrated superior capabilities in speech feature extraction and classification, which are beneficial for monitoring emotional dynamics in educational settings.

2.3.2. Advances in Model Architectures

Wav2vec was the first to apply unsupervised pre-training to speech recognition with a fully convolutional model and achieved a WER (Word Error Rate; WER is the ratio of errors in a transcript to the total words spoken) of 2.43% by introducing this method on the WSJ (Wall Street Journal; the Wall Street Journal corpus defines a test group of 283 speakers) test set. EVA-GAN [17] introduced improvements in spectral and high-frequency reconstruction, while VITS2 [18] and Glow-WaveGAN2 [19] focused on enhancing synthesis quality and efficiency through adversarial learning and large-scale pre-trained models.

Tacotron2 [20] leveraged a sequence-to-sequence recurrent network with a modified WaveNet [21] vocoder to achieve near-human audio quality. EmoQ-TTS [22] introduced fine-grained emotion intensity control for better emotional expressiveness.

Kim et al. simplified the TTS pipeline by dividing it into semantic and acoustic modeling stages, reducing training complexity [23]. Trini TTS [24] and NSV-TTS [25] focused on pitch-controllable models and self-supervised learning to extract unsupervised linguistic units, respectively.

2.3.3. Innovations in High-Frequency Reconstruction and Robustness

CLONE [26] addressed TTS system limitations in high-frequency reconstruction with a dual parallel autoencoder. Natural speech [27] employed a variational autoencoder (VAE) [28] to enhance the prior capacity from text and reduce posterior complexity from speech. Period VITS [29] introduced an explicit periodicity generator for high-quality speech synthesis.

2.3.4. Using Pre-Training Models

XPhoneBERT and MelHuBERT [30] improved multilingual TTS models and training efficiency, respectively. BDDM [31] utilized a bilateral denoising diffusion model to generate high-fidelity audio samples.

2.3.5. Synthesis of Results

The reviewed literature highlights significant progress in emotional and prosodic feature enhancements, model architectures, and high-frequency reconstruction. However, gaps remain in improving the robustness of posterior encoders and scalability to multi-speaker scenarios. Future research should focus on leveraging pre-trained models and innovative architectural designs to address these challenges and further advance the field of end-to-end speech synthesis.

3. Methods

The speech synthesized by the VITS model lacks naturalness and prosody. Through spectrogram comparison analysis, whose results are shown in the last figure, it was observed that while the synthesized audio closely matches the original in the low-frequency range, significant discrepancies exist in the mid- and high-frequency regions, where many details remain unrecovered. The VITS model struggles to effectively learn the mid- and high-frequency components of the speech data in the dataset, resulting in synthesized speech that fails to accurately reproduce prosody. After a careful analysis, we discovered that VITS can be seen as a variational autoencoder (VAE) architecture network. Its prior

and posterior encoders extract the prior and posterior distributions of latent variables z from text and spectrograms, respectively. These distributions are then aligned and fed into a vocoder to reconstruct the audio. The final optimization objective is not the log-likelihood of the latent variable z , but rather the maximization of the Evidence Lower Bound (ELBO).

To enhance the expressiveness and effectiveness of VITS speech synthesis, considerable effort has been directed towards improving the capability of the prior encoder to extract textual features or optimizing alignment algorithms to enhance alignment accuracy. However, there has been no related work focused on modifying the posterior encoder to improve the extraction efficiency of spectrogram features. As the prior encoder in VITS is essentially a text encoder, with the remarkable performance of pre-trained text encoders like BERT in the NLP domain, related work targeting the prior encoder of the VITS model has proposed corresponding pre-trained models. For instance, pre-trained models such as XPhoneBERT have been introduced specifically for phoneme-based applications. At the same time, since the prior encoder handles phonemes rather than raw text, there are also related studies suggesting that using pseudo-phonemes instead of raw text as input can improve the model's expressiveness. However, these approaches either require significant time to collect and process corresponding corpora (XPhoneBERT necessitates a large volume of text converted into phoneme sequences as input and adopts a structure similar to BERT, thus requiring extensive pre-training time) or entail extensive data preprocessing (the pseudo-phoneme approach involves multiple rounds of clustering and merging of all pseudo-phonemes and then the hidden representation in the entire dataset), making training quite complex.

Our proposed model architecture is shown in Figures 1 and 2. Figure 1 illustrates the structure of the model during the training phase. The model receives text sequences and the corresponding waveforms as input. The prior encoder, also known as the text encoder, is responsible for computing the prior distribution of the latent variable z . The posterior encoder is responsible for computing the posterior distribution of the latent variable z from the spectrogram obtained through the Fast Fourier Transform (FFT) of the waveform. The alignment algorithm calculates the alignment matrix between the spectrogram and the text. The decoder samples from the posterior distribution to predict the waveform. The duration predictor learns the duration of the corresponding text in the waveform from the alignment matrix.

Figure 2 shows the structure of the model during the inference phase. During inference, the model receives text sequences as input. After processing is concluded by the prior encoder, the model samples are based on the duration predictor's instructions. This is followed by normalization flow. Then, the samples are passed to the posterior encoder, and, finally, the decoder generates the waveform.

The original posterior encoder in VITS takes linear spectrogram input x and possible conditional input g and then performs forward propagation operations. Firstly, if there is a conditional input g , it is processed through conditional layers. Then, for each layer, the input x is convolved through input layers (*in_layers*) and then processed through an activation function. After the activation function processing, residual connections and skip connections are added to the input and Dropout is applied. Finally, the outputs of all layers are summed and multiplied by the input mask (x_mask). Such posterior encoder structures tend to be relatively simple, lacking parallel information and failing to effectively extract features from various frequency bands of speech.

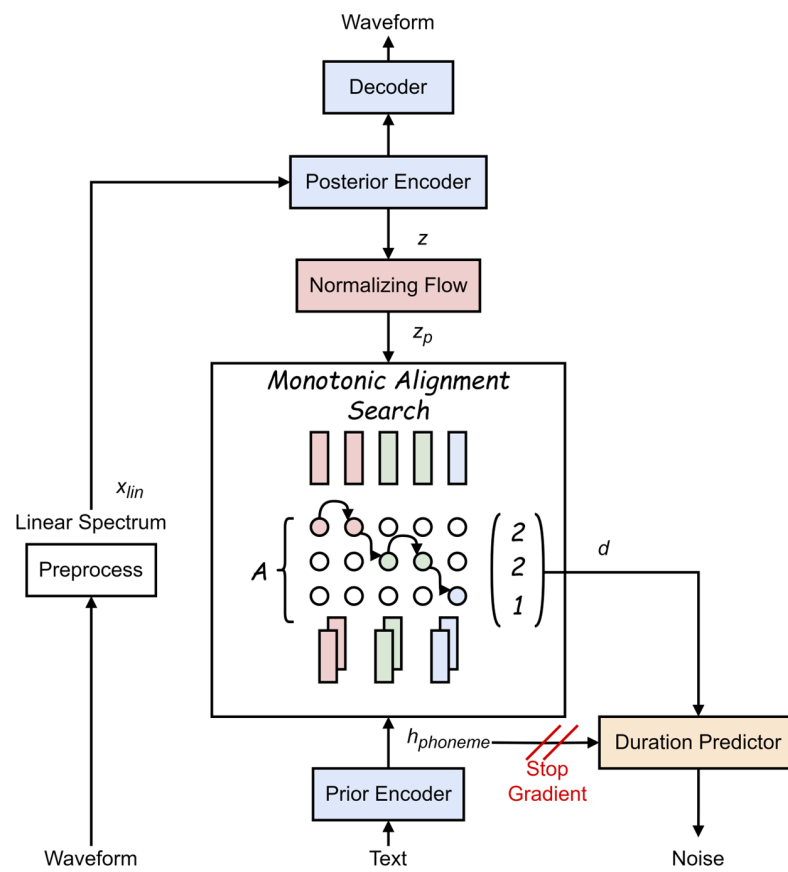


Figure 1. The training procedure of our new model architecture.

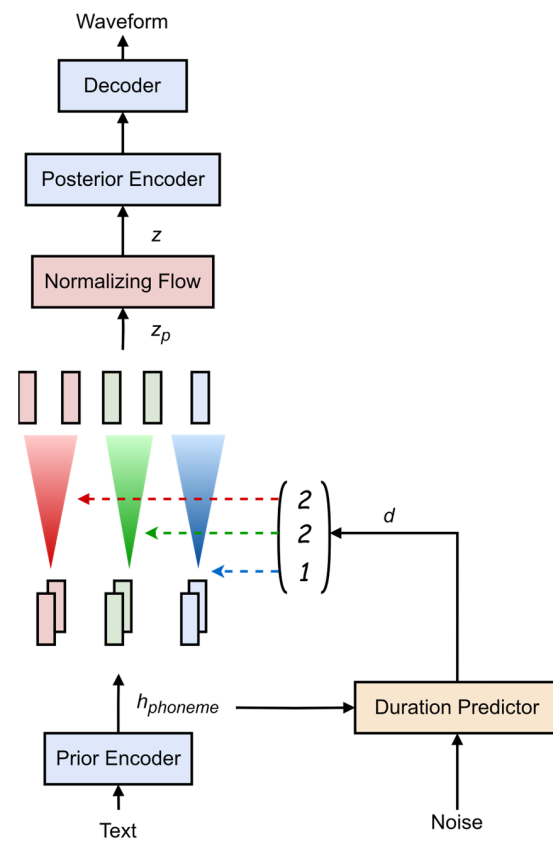


Figure 2. The inferencing procedure of our new model architecture.

The original posterior encoder is capable of handling linear spectrograms and synthesizing relatively natural speech. However, when compared to the original data, it consistently falls short in capturing rhythm and intonation. Therefore, building upon the original posterior encoder, we introduced the pre-trained model HuBERT, which is a self-supervised speech representation learning method. It leverages an offline clustering step to provide aligned target labels for BERT-like prediction losses. We found that introducing pre-trained models can enhance the posterior encoder's learning effectiveness of spectrogram features.

As shown in Figure 3, by using the pre-trained model to extract corresponding hidden states, which are then projected onto the same dimensionality as the output of the original WaveNet (stat), followed by self-attention pooling and fusion of hidden states and stat, we can significantly improve the model's ability to reconstruct mid- and high-frequency components.

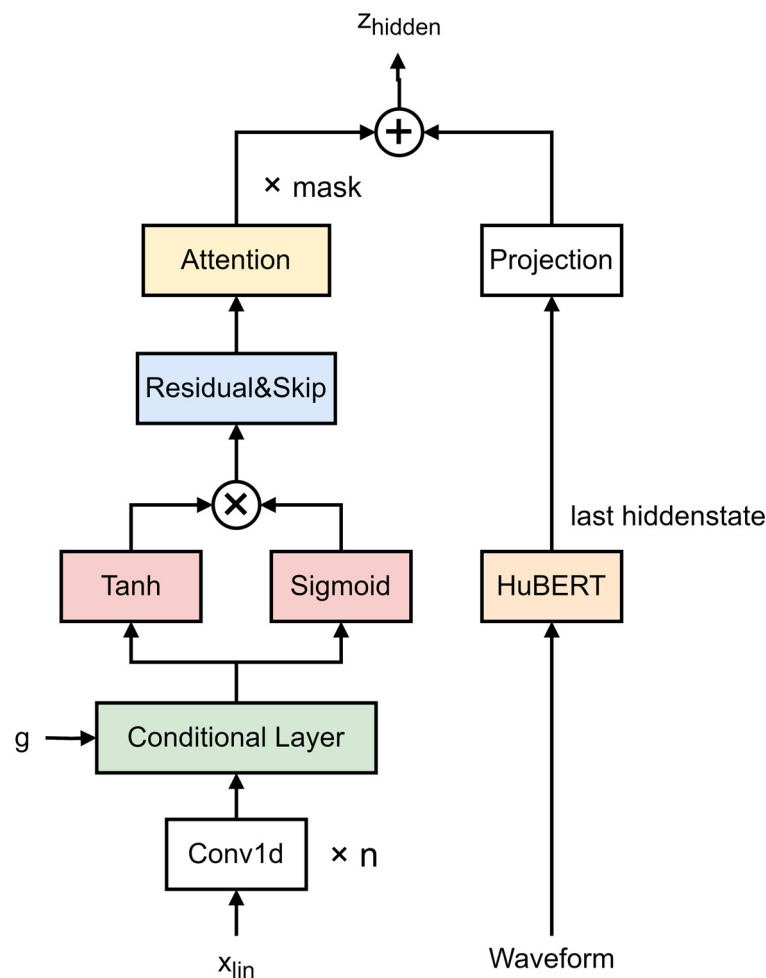


Figure 3. Our new posterior encoder.

The corresponding optimization objective has also changed, as depicted in Equation (1), as follows:

$$\begin{aligned}
 A &= \arg \max_{\tilde{A}} \log p(x|c_{phoneme}, \tilde{A}) \\
 &= \arg \max \log N\left(f(x); \mu(c_{phoneme}, \tilde{A}), \sigma(c_{phoneme}, \tilde{A})\right),
 \end{aligned} \tag{1}$$

and the posterior distribution of the latent variable has also changed, as outlined in Equation (2):

$$p_{\theta}(z_{\text{hidden}} | c_{\text{phoneme}}) = N\left(f_{\theta}(z_{\text{hidden}}); \mu_{\theta}(c_{\text{phoneme}}), \sigma_{\theta}(c_{\text{phoneme}})\right) \left| \det \frac{\partial f_{\theta}(z_{\text{hidden}})}{\partial z_{\text{hidden}}} \right|, \quad (2)$$

The c_{phoneme} is the combination of the phoneme of the text and the alignment matrix of the corresponding spectrum, as depicted in Equation (3):

$$c_{\text{phoneme}} = [c_{\text{phoneme}}, A] \quad (3)$$

The framework we proposed is based on the VITS architecture, which utilizes a conditional variational autoencoder (CVAE) [32] structure. VITS consists of a posterior encoder, a prior encoder, a waveform decoder, and a duration predictor. Building upon VITS, we enhanced the expressiveness of the model by modifying the original posterior encoder. The task of the original posterior encoder is to extract the posterior distribution $q_{\phi}(z|x_{\text{lin}})$ of latent variables z given the linear spectrogram x_{lin} , where the decoder generates output audio from sampled z . The new posterior encoder we proposed aims to obtain the distribution of the last hidden state, $q_{\phi}(z_{\text{hidden}}|x_{\text{lin}})$, given the linear spectrogram x_{lin} and the original waveform y , and then sample new latent z_{hidden} from the distribution to generate the output audio. The prior encoder consists of a normalizing flow and a text encoder, matching z_{hidden} to the conditional prior distribution $p_{\theta}(z_{\text{hidden}} | c_{\text{phoneme}}, A_{\text{phoneme}})$ given phonemes c_{phoneme} —obtained from preprocessing the given text and alignment matrix A_{phoneme} . By using the Monotonic Alignment Search (MAS) algorithm [7], the alignment matrix A_{phoneme} is computed. The duration predictor utilizes A_{phoneme} to learn to predict the duration of each phoneme. We found that the original text encoder structure is too simple to effectively learn some hidden states of the text. Therefore, we decided to replace the original text encoder with pre-trained models such as BERT, XPhoneBERT, etc. The learned hidden state is then projected back to the output dimension of the original VITS text encoder to replace a part of the text encoder. Building upon this, we also observed the work of replacing the text encoder of the VITS model with a pseudo-phoneme [33] encoder. The specific process involves using wav2vec 2.0 to process the waveform, indexing, clustering, and merging the resulting hidden states to obtain representations of pseudo-phonemes. These pseudo-phonemes are then input into the pseudo-phoneme encoder, and the encoded results are fed into the MAS algorithm. We have considered that HuBERT is a follow-up work to wav2vec 2.0. Thus, based on the work with the pseudo-phoneme encoder, we plan to replace wav2vec 2.0 with HuBERT. We will then perform the necessary feature projection to ensure dimensionality consistency and conduct comparative experiments.

4. Experiments

The training data consist of the LJSpeech [34] dataset, which is a publicly available dataset for speech synthesis research. It contains English speech samples from a single speaker. This dataset is commonly used for training and evaluating text-to-speech models. The speech samples in the dataset are of high quality, often characterized by good clarity and fluency. They cover a wide range of sentence types, including simple everyday phrases and complex literary works. The speech samples are stored in WAV format and are accompanied by text documents containing the corresponding textual content for each speech sample. We also utilized the Databaker [35] dataset, provided by Beijing Databaker Technology Co., Ltd., Beijing, China. This Chinese speech dataset is primarily used for research and development in fields such as speech recognition and synthesis. The dataset contains Chinese speech samples from the same speaker, covering various everyday phrases and scenarios, including news reports, movie clips, etc. These speech samples are typically stored in WAV format and come with corresponding text annotations. The Databaker dataset also boasts high-quality speech with excellent clarity and fluency, making it well suited for training and evaluating speech-related models.

We expanded the VITS model by modifying its WaveNet-based posterior encoder and concurrently introduced other pre-trained models for comparison. During the experiments, we trained the original VITS model using the best hyperparameters as specified in the original VITS paper. This includes using the AdamW [36] optimizer with parameters $\beta_1 = 0.8$, $\beta_2 = 0.99$, weight decay $\lambda = 0.01$, and initial learning rate of 2×10^{-4} . We run the model for 10K training steps with a batch size of 64 (equivalent to approximately 500 training epochs for both English and Chinese data). During training the new model, after a comparison, we found that freezing the parameters of HuBERT and introducing a separate optimizer to update only the posterior encoder part can prevent adverse consequences like gradient vanishing. However, this approach results in slower training compared to the original posterior encoder. The original VITS model trained on 2 RTX 6000 Ada GPUs (NVIDIA, Santa Clara, CA, USA) with a batch size of 64 achieves a training speed of approximately 1 min per epoch, but the new BERTIVITS trained on the same 2 RTX 6000 Ada GPUs with a reduced batch size of 16 achieves a training speed of approximately 3 min per epoch. Due to memory constraints, BERTIVITS needs to reduce the batch size for training without changing the devices.

During the experimental process, we found that the performance of the original text encoder in VITS was not satisfactory. However, improvements made by related works such as XPhoneBERT and pseudo-phonemes significantly enhance the expressiveness of the synthesized speech in VITS. Therefore, based on these related works, we introduced a new posterior encoder and observed positive experimental results.

Due to the introduction of the pre-trained model, the loss becomes very large during the initial training stages. Moreover, following the original optimization strategy, using a single optimizer for updating parameters of the entire generator quickly leads to gradient vanishing. Therefore, we set up two optimizers. The first optimizer retains the original hyperparameters, namely the AdamW optimizer, which was mentioned earlier. The second optimizer is a new one we introduced, which is the SGD [37] optimizer. We set its learning rate to 1×10^{-4} , momentum to 0.9, and weight decay to 2×10^{-5} . The first optimizer is responsible for optimizing parameters in the generator (*net_g*) except for the posterior encoder. The second optimizer is specifically tasked with optimizing the parameters of the modified posterior encoder. Additionally, we need to freeze the parameters of HuBERT; otherwise, gradient explosion occurs around the 40th epoch, unaffected by hyperparameters. We conducted multiple experiments, including adjusting learning rates and changing optimizers, but we encountered uncontrollable errors between the 30th and 50th epochs, leading us to prematurely terminate training for these models.

5. Results and Discussion

We evaluated the performance of the TTS model using subjective and objective metrics. For a subjective evaluation of naturalness, we randomly selected 50 real test audio samples along with their text transcripts for each language and measured the Mean Opinion Score. The Mean Opinion Score (MOS) is a subjective rating system in which human evaluators assess the perceived quality of synthesized speech on a scale from 1 to 5. It is one of the most widely used evaluation methods for TTS systems. Here, for each text transcript, we synthesize speech using six different models, including the baseline VITS and the extended BERTIVITS model with the new posterior encoder. For each language, we enlist 10 professionals to rate each of the five types of speech (four synthesized speech samples and one real speech sample). Naturalness is rated on a scale from 1 to 5, with increments of 1 point. Here, each evaluator is unaware of which model produced which speech sample. The results are as shown in Tables 1 and 2.

Table 1. Comparison of the MOS scores for all models on the LJSpeech dataset.

Model	MOS ($\uparrow$)
Ground Truth	4.72 ± 0.06
VITS-Baseline	4.00 ± 0.08
VITS-XPB	4.14 ± 0.07
VITS-Pseudo Phoneme (wav2vec2.0)	3.99 ± 0.10
VITS-Pseudo Phoneme (HuBERT)	4.11 ± 0.07
BERTIVITS (our)	4.16 ± 0.06

Table 2. Comparison of the MOS scores for all models on the Databaker dataset.

Model	MOS ($\uparrow$)
Ground Truth	4.66 ± 0.04
VITS-Baseline	3.53 ± 0.07
VITS-XPB	3.87 ± 0.08
VITS-Pseudo Phoneme (wav2vec2.0)	3.66 ± 0.06
VITS-Pseudo Phoneme (HuBERT)	3.78 ± 0.07
BERTIVITS (our)	3.89 ± 0.07

To objectively evaluate the distortion and prosodic differences between real and synthesized speech, we calculate the Mel-Cepstral Distortion (MCD; dB) for comparison. Mel-cepstral distortion is an objective metric used to assess synthetic voices. It quantifies the difference between two sequences of Mel (Melody) cepstra. The results are presented in Tables 3 and 4.

Table 3. Comparison of the MCD for all models on the LJSpeech dataset.

Model	MCD ($\downarrow$)
Ground Truth	0.00
VITS-Baseline	7.04
VITS-XPB	6.63
VITS-Pseudo Phoneme (wav2vec2.0)	6.77
VITS-Pseudo Phoneme (HuBERT)	6.65
BERTIVITS (our)	6.52

Table 4. Comparison of the MCD for all models on the Databaker dataset.

Model	MOS ($\downarrow$)
Ground Truth	0.00
VITS-Baseline	8.12
VITS-XPB	7.73
VITS-Pseudo Phoneme (wav2vec2.0)	7.86
VITS-Pseudo Phoneme (HuBERT)	7.75
BERTIVITS (our)	7.71

During the training process, we also recorded the Root Mean Square Error (RMSE) and training speed (batch size = 64) of different models on the same dataset. The RMSE indicates the absolute difference between the generated speech and the original speech, while the training speed roughly reflects the difficulty of training different models. When training different models, we kept the hyperparameters consistent and set the gradients of the pre-trained model to be non-updatable. The comparison of the RMSE and training speed is shown in Tables 5 and 6.

Table 5. Comparison of the RMCE and train speed for all models on the LJSpeech dataset.

Model	RMSE	Train Speed (min/Epoch)
Ground Truth	0.00	0
VITS-Baseline	18.26	1.04
VITS-XPB	15.37	2.85
VITS-Pseudo Phoneme (wav2vec2.0)	16.44	4.91
VITS-Pseudo Phoneme (HuBERT)	15.66	5.14
BERTIVITS (our)	14.71	3.22

Table 6. Comparison of the RMCE and train speed for all models on the Databaker dataset.

Model	RMSE	Train Speed (min/Epoch)
Ground Truth	0.00	0
VITS-Baseline	21.77	1.27
VITS-XPB	19.89	2.94
VITS-Pseudo Phoneme (wav2vec2.0)	20.16	5.11
VITS-Pseudo Phoneme (HuBERT)	19.92	5.37
BERTIVITS (our)	18.82	3.73

Tables 1 and 2 both provide MOS score results on two datasets. It can be observed that our model contributes to improving the performance of the VITS model on this evaluation metric. Our model achieved significant improvements of 0.16 and 0.36 on the two datasets respectively. This clearly demonstrates the effectiveness of our model. In terms of Mel-Cepstral Distortion (MCD), as shown in Tables 3 and 4, our model contributes to improving the objective quality of speech synthesized by the VITS model. Mel-Cepstral coefficients decreased by 0.52 and 0.49 on the two datasets respectively. Here, we observe that the magnitude of the MCD reduction and MOS improvement are not always directly proportional, which is consistent with the findings in [38]. Additionally, as indicated in Tables 5 and 6, our model did not significantly increase training time, although the training time per epoch for our model is higher than that of VITS-XPB. However, XPhoneBERT has not been pre-trained on a large corpus of Chinese data; it has primarily focused on English and Vietnamese. Given significant phonetic differences among languages, we have not fine-tuned the original XPhoneBERT for Chinese. Thus, our current work has not fully explored the potential of XPhoneBERT in the Chinese domain. In the future, we will consider adopting the training methodology and data preprocessing techniques of XPhoneBERT to create a large-scale Chinese phoneme corpus and conduct pre-training on it.

We selected the same text and used VITS, VITS-XPB, VITS-Pseudo Phoneme (HuBERT), and BERTIVITS to plot spectrograms, as shown in Figure 4. The comparison is shown in the figure, indicating that our model can improve the spectral details of the output of the VITS model. Observing the spectrograms, our model performs significantly better than VITS-Baseline in restoring frequencies above 4000 Hz and does well in restoring real audio in the 2000–4000 Hz frequency range. That is, it can effectively restore human voice while capturing more high-frequency details.

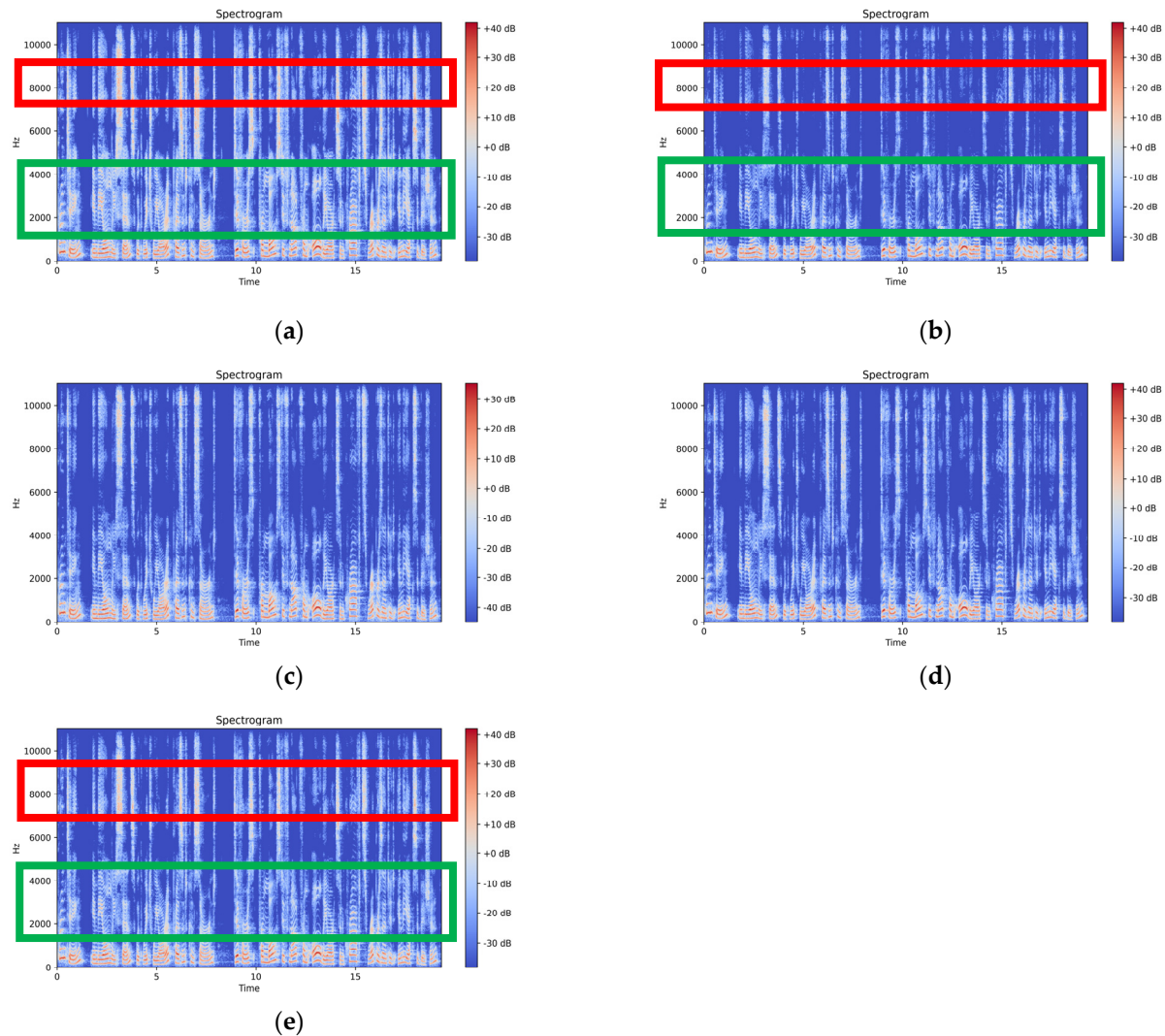


Figure 4. Visualization of the spectrograms by different models corresponding to the text “Printing, in the only sense with which we are at present concerned, differs from most if not from all the arts and crafts represented in the Exhibition”. (a) Ground Truth; (b) VITS-Baseline; (c) VITS-XPB; (d) VITS-Pseudo Phoneme (HuBERT); (e) BERRTIVITS.

In summary, our model significantly improves the performance of the VITS model in various aspects through multiple enhancements, including increased subjective evaluation scores, improved objective evaluation metrics, and optimization of spectrogram details.

Our method achieves significant results on single-speaker datasets, but its effectiveness is not pronounced in multi-speaker scenarios. The differences in voices of different individuals vary significantly, including variations in fundamental frequency, energy, and consonants, among many other aspects. The pre-trained model HuBERT was originally designed for tasks such as speech recognition, which may not fully align with the requirements of tasks like speech synthesis and voice cloning. Its effectiveness in our task stems from the fact that in single-speaker datasets, where only one person’s voice needs to be reconstructed, we only need to consider the characteristics of that individual’s voice. If considering multi-speaker scenarios, adjustments to the architecture of the pre-trained model are necessary. This involves placing greater emphasis on extracting features specific to different speakers and embedding speaker identity information. Alternatively, specialized optimization of corresponding modules in the VITS model, such as pre-training the vocoder and fine-tuning other modules, may be required to better accomplish the tasks at hand. Therefore, our method can be applied in single-speaker scenarios, such as educational and

customer service settings, involving one-on-one or one-to-many interactions. It significantly enhances the naturalness of synthesized speech in these scenarios, better capturing the voice of the target speaker.

6. Conclusions

In this paper, we have successfully modified the posterior encoder of the original VITS model to incorporate the pre-trained HuBERT model, achieving enhanced extraction of prosodic features during speech synthesis. This innovation retains the WaveNet block architecture while significantly improving the naturalness, intonation, and prosodic features of synthesized speech. Comparative experiments and subjective evaluations confirm the superior performance of our approach, demonstrating the effectiveness of integrating pre-trained models like HuBERT into speech synthesis frameworks.

Our work contributes to the theoretical understanding of prosodic feature extraction and their integration into end-to-end speech synthesis models. Practically, the proposed method paves the way for more realistic and natural-sounding speech synthesis, which is crucial for applications in intelligent speech interaction systems, such as virtual assistants and automated customer service. This study is distinct in its application of self-attention pooling and HuBERT for enhancing prosodic feature extraction in VITS models. It represents a novel approach to leveraging pre-trained models in speech synthesis, showcasing the potential for improved performance through cross-domain knowledge transfer.

While our findings are promising, it is important to acknowledge several limitations. These include the reliance on a specific dataset for training and evaluation, which may not be generalizable to all speech synthesis scenarios. Additionally, the computational cost associated with integrating HuBERT might pose challenges in real-time applications. Future research could explore the scalability of our method to different languages and dialects as well as investigate the impact of various pre-trained models on speech synthesis quality. Moreover, optimizing the computational efficiency of the integrated model would be beneficial for deployment in resource-constrained environments.

In summary, this research offers a novel strategy for enhancing speech synthesis technologies by integrating pre-trained models. It underscores the importance of leveraging linguistic knowledge from large-scale textual data to improve speech synthesis performance and realism. Our findings lay a solid foundation for the advancement of intelligent speech interaction systems, opening new avenues for creating more intelligent and natural speech synthesis systems.

Author Contributions: Conceptualization, Z.W. and M.S.; Methodology, Z.W. and M.S.; Software, Z.W. and M.S.; Validation, Z.W. and M.S.; Writing—original draft, Z.W. and M.S.; Writing—review & editing, Z.W., M.S. and D.Z.; Visualization, Z.W. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets we used are from reference [34,35].

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2023**. [[CrossRef](#)]
2. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**. [[CrossRef](#)]
3. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**. [[CrossRef](#)]

4. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H.H.; Lakhota, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv* **2021**. [\[CrossRef\]](#)
5. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv* **2020**. [\[CrossRef\]](#)
6. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 5206–5210. [\[CrossRef\]](#)
7. Kim, J.; Kong, J.; Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 5530–5540.
8. Nguyen, L.T.; Pham, T.; Nguyen, D.Q. XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech. *arXiv* **2023**. [\[CrossRef\]](#)
9. Zhang, J.-H.; Zhang, Z.-H.; Yan, Q.; Wang, P.-W. Speech emotion recognition based on energy frames and time frequency fusion. *Comput. Sci.* **2023**, 1–11.
10. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [\[CrossRef\]](#)
11. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S. Tacotron: Towards End-to-End Speech Synthesis. *arXiv* **2017**. [\[CrossRef\]](#)
12. Zhang, X.; Hu, H.; Cao, X.; Wang, W. Expressive Speech Synthesis Method Based on Tacotron Model and Prosodic Correction. *Data Acquis. Process.* **2022**, *37*, 4. [\[CrossRef\]](#)
13. Cui, L.; Cui, C.; Liu, Z.; Xue, K. Speech Emotion Recognition Based on Improved MFCC and Parallel Hybrid Model. *Comput. Sci.* **2023**, *50*, S1.
14. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [\[CrossRef\]](#)
15. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B. A Database of German Emotional Speech. In Proceedings of the Interspeech 2005, Lissabon, Portugal, 4–8 September 2005.
16. Ou, Z.-G.; Liu, Y.-P.; Li, R.-L.; Qin, K. Research on Teacher Speech Emotion Recognition in International Chinese Language Classrooms. *Mod. Educ. Technol.* **2023**, *33*, 8.
17. Liao, S.; Lan, S.; Zachariah, A.G. EVA-GAN: Enhanced Various Audio Generation via Scalable Generative Adversarial Networks. *arXiv* **2024**. [\[CrossRef\]](#)
18. Kong, J.; Park, J.; Kim, B.; Kim, J.; Kong, D.; Kim, S. VITS2: Improving Quality and Efficiency of Single-Stage Text-to-Speech with Adversarial Learning and Architecture Design. *arXiv* **2023**. [\[CrossRef\]](#)
19. Lei, Y.; Yang, S.; Cong, J.; Xie, L.; Su, D. Glow-WaveGAN 2: High-quality Zero-shot Text-to-speech Synthesis and Any-to-any Voice Conversion. *arXiv* **2022**. [\[CrossRef\]](#)
20. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerry-Ryan, R. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *arXiv* **2018**. [\[CrossRef\]](#)
21. Van Den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A Generative Model for Raw Audio. *arXiv* **2016**. [\[CrossRef\]](#)
22. Im, C.-B.; Lee, S.-H.; Kim, S.-B.; Lee, S.-W. EMOQ-TTS: Emotion Intensity Quantization for Fine-Grained Controllable Emotional Text-to-Speech. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 6317–6321. [\[CrossRef\]](#)
23. Kim, M.; Jeong, M.; Choi, B.J.; Kim, S.; Lee, J.Y.; Kim, N.S. Utilizing Neural Transducers for Two-Stage Text-to-Speech via Semantic Token Prediction. *arXiv* **2024**. [\[CrossRef\]](#)
24. Ju, Y.; Kim, I.; Yang, H.; Kim, J.-H.; Kim, B.; Maiti, S.; Watanabe, S. TriniTTS: Pitch-controllable End-to-end TTS without External Aligner. In Proceedings of the Interspeech 2022, ISCA, Incheon, Republic of Korea, 18–22 September 2022; pp. 16–20. [\[CrossRef\]](#)
25. Zhang, H.; Yu, X.; Lin, Y. NSV-TTS: Non-Speech Vocalization Modeling and Transfer in Emotional Text-to-Speech. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5. [\[CrossRef\]](#)
26. Liu, Z.; Tian, Q.; Hu, C.; Liu, X.; Wu, M.; Wang, Y.; Zhao, H.; Wang, Y. Controllable and Lossless Non-Autoregressive End-to-End Text-to-Speech. *arXiv* **2022**. [\[CrossRef\]](#)
27. Tan, X.; Chen, J.; Liu, H.; Cong, J.; Zhang, C.; Liu, Y.; Wang, X.; Leng, Y.; Yi, Y.; He, L.; et al. NaturalSpeech: End-to-End Text to Speech Synthesis with Human-Level Quality. *arXiv* **2022**. [\[CrossRef\]](#)
28. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2022**. [\[CrossRef\]](#)
29. Shirahata, Y.; Yamamoto, R.; Song, E.; Terashima, R.; Kim, J.-M.; Tachibana, K. Period VITS: Variational Inference with Explicit Pitch Modeling for End-To-End Emotional Speech Synthesis. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5. [\[CrossRef\]](#)
30. Lin, T.-Q.; Lee, H.; Tang, H. MelHuBERT: A simplified HuBERT on Mel spectrograms. *arXiv* **2023**. [\[CrossRef\]](#)
31. Lam, M.W.Y.; Wang, J.; Su, D.; Yu, D. BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis. *arXiv* **2022**, arXiv:2203.13508.

32. Sohn, K.; Lee, H.; Yan, X. Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Glasgow, UK, 2015. Available online: https://papers.nips.cc/paper_files/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html (accessed on 14 May 2024).
33. Kim, M.; Jeong, M.; Choi, B.J.; Ahn, S.; Lee, J.Y.; Kim, N.S. Transfer Learning Framework for Low-Resource Text-to-Speech using a Large-Scale Unlabeled Speech Corpus. In *Proceedings of the Interspeech 2022, Incheon, Republic of Korea, 18–22 September 2022*; pp. 788–792. [CrossRef]
34. Ljspeech: Ito, K, The LJ Speech Dataset. 2017. Available online: <https://keithito.com/LJ-Speech-Dataset/> (accessed on 14 May 2024).
35. Databaker: Databaker Technology, Chinese Standard Female Voice Database. Available online: <https://www.data-baker.com/data/index/TNtts/> (accessed on 14 May 2024).
36. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2019**. [CrossRef]
37. Robbins, H.; Monro, S. A Stochastic Approximation Method. *Ann. Math. Stat.* **1951**, *22*, 400–407. [CrossRef]
38. Saeki, T.; Tachibana, K.; Yamamoto, R. DRSpeech: Degradation-Robust Text-to-Speech Synthesis with Frame-Level and Utterance-Level Acoustic Representation Learning. *arXiv* **2022**. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.