applied sciences

Article

An Emotion Speech Synthesis Method Based on VITS

Wei Zhao [1,*] and Zheng Yang [2]

[1] School of Data Science and Intelligent Media, Communication University of China, Beijing 100024, China
[2] College of Information and Communication Engineering, Communication University of China, Beijing 100024, China
* Correspondence: zhao_wei@cuc.edu.cn

Abstract: People and things can be connected through the Internet of Things (IoT), and speech synthesis is one of the key technologies. At this stage, end-to-end speech synthesis systems are capable of synthesizing relatively realistic human voices, but the current commonly used parallel text-to-speech suffers from loss of useful information during the two-stage delivery process, and the control features of the synthesized speech are monotonous, with insufficient expression of features, including emotion, leading to emotional speech synthesis becoming a challenging task. In this paper, we propose a new system named Emo-VITS, which is based on the highly expressive speech synthesis module VITS, to realize the emotion control of text-to-speech synthesis. We designed the emotion network to extract the global and local features of the reference audio, and then fused the global and local features through the emotion feature fusion module based on the attention mechanism, so as to achieve more accurate and comprehensive emotion speech synthesis. The experimental results show that the Emo-VITS system's error rate went up a little bit compared with the network without emotionality and does not affect the semantic understanding. However, this system is superior to other networks in naturalness, sound quality, and emotional similarity.

Keywords: IoT; Emo-VITS; emotional speech synthesis; emotion feature fusion

check for updates

Citation: Zhao, W.; Yang, Z. An Emotion Speech Synthesis Method Based on VITS. *Appl. Sci.* **2023**, *13*, 2225. https://doi.org/10.3390/app13042225

Academic Editors: Chunsheng Zhu and Weidong Fang

Received: 30 November 2022
Revised: 3 February 2023
Accepted: 7 February 2023
Published: 9 February 2023

1. Introduction

People and things can be connected through the IoT [1–3], such as the answering function of smart speakers, etc. Speech synthesis is one of the key technologies [4,5]. Text-to-speech (TTS) is a kind of technology that can convert any input text into corresponding speech. It is one of the indispensable modules in human–computer speech interaction. It is not only used in audiobooks, voice guidance of public service facilities, and other fields, but is also an important part of intelligent driving and robotics. At present, speech synthesis has been able to achieve excellent synthesis effects; the naturalness of speech can be close to the real human. As a classical algorithm in neural network-based sequence-to-sequence (Seq2Seq) models, Tacotron2 [6] uses a modified WaveNet [7] as a vocoder on the basis of Tacotron's framework [8]. On the same dataset, the MOS value of Tacotron2 reached 4.53 compared with 4.58 for human speech. However, the autoregressive model uses spectrum sequences to synthesize speech, which cannot avoid the operation process of cyclic network, which leads to slow training and prediction in the speech synthesis process. Non-autoregressive TTS models use parallel structures that can effectively speed up synthesis, such as FastSpeech [9], FastSpeech 2 [10], and TransformerTTS [11].

Although the synthesized speech of the current TTS model has achieved excellent performance, it is still difficult for people to control the emotion and style of speech well. Therefore, it is still difficult to express emotions flexibly without affecting the text expression itself in current speech synthesis. The reason is that the emotion of speech is affected by many aspects, such as pitch, tone, and prosody, among which there are so many details that it is difficult for people to completely decouple various paralinguistic features. So far, emotional embedding in speech synthesis is still very challenging work. In addition, emotional information is also affected by various paralanguage features, which are difficult to extract and decouple from other paralanguage features. Some affective speech synthesis methods rely on artificially annotated affective speech datasets [12,13], but artificially annotated datasets are costly and difficult to obtain, and it is difficult to popularize this training method. Therefore, more and more researchers have studied unsupervised methods.

The classical unsupervised method is to extract global emotional information from reference audio [14–16], such as Global Style Tokens (GST) based on the Tacotron2 emotion speech synthesis model. The basic idea of emotion embedding is to learn an embedding vector for each emotion and introduce a global style token to obtain utterance-level embedding. In recent studies, Variational AutoEncoder (VAE) [17] showed stronger capabilities in disentanglement, scaling, and interpolation for style control, and VAE for explicit modeling of potential variables has become one of the most popular methods.

However, there are still some problems in the emotion embedding methods of speech synthesis. First of all, there is the problem of useful information loss in the two-stage transmission of parallel TTS. At present, the parallel speech synthesis algorithm is generally simplified into two stages. The first stage analyzes the input text information, generates the corresponding linguistic specification input, and predicts the spectrum waveform. In the second stage, according to the spectrum waveform provided by the speech analysis section, the corresponding audio is generated to realize the sound. In general, the modules of the two-stage model are trained separately, and there will be a certain gap between the spectrum generated in the first stage and the real spectrum. Therefore, when the spectrum generated in the first stage is applied to the second stage, this gap will be preserved, and the effective information will be lost in transmission, thus reducing the final audio synthesis effect. On the other hand, the training of the two-stage model is separate, so the connection relies on pre-defined intermediate features, which makes it difficult to learn the hidden representation directly from the previous stage to further improve performance [18]. A complete end-to-end model, on the other hand, does not require separate training, so it avoids these problems to some extent.

Second, if the control features of synthesized speech have only one granularity, the result is likely to be boring. In expressive TTS, variation information can be modeled at different granularity. It can be roughly divided into coarse-grained levels of information that reflect global features and fine-grained levels of information that reflect local features. For example, speaker level, paragraph level, and utterance level information are coarse-grained information, while word level, phoneme level, and frame level are fine-grained information [19]. The emotions of human speech are notoriously complex. When an intense emotion is extracted and applied to a new text, the veracity of the voice is often reduced because of the sharp intonation. Selecting a single emotion feature to control will make the emotion expression of synthesized speech decline.

In this paper, to solve the above problems, we put forward the Emo-VITS system. Emo-VITS is based on the Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) emotional speech synthesis method. VITS is a completely parallel end-to-end TTS method proposed by Kim et al. [18], which connects two modules of the TTS system through latent variables and generates sound audio that is more natural than that generated by the two-stage model. VITS has wider application prospects in the field of emotion speech synthesis because the structure, which is based on VAE, is more controllable than other structures, such as Tacotron. In this paper, Emo-VITS is proposed based on VITS. Emo-VITS has two emotion encoders to encode the extracted reference speech emotion features and transform the synthesized speech emotion features from both global and local perspectives. At the same time, an emotion feature fusion layer was used to integrate the extracted global and local emotion features, following the example of Attention Feature Fusion (AFF) [20]. In the prediction process, when given a reference audio with a given emotion, the Emo-VITS model is able to extract its features and synthesize a speech with that emotional feature, while achieving intensity control of that emotional feature. At the same time, we also cluster the emotional feature vectors of the training data so that the desired type of features can be synthesized without reference to the audio. Emo-VITS also supports the fusion of the global features of one reference voice with the local features of the other reference voice, and then generates the emotional voice with two speech features.

2. VITS

VITS is a high-performance speech synthesis model that combines variational inference augmented with normalizing flows and an adversarial training process. VITS uses Variational AutoEncoder (VAE) and applies normalized flows [21] to conditional prior distribution and adversarial training on waveform domains to improve the quality of the speech waveform. By connecting the two modules of the TTS system with latent variables, complete end-to-end learning is achieved, which can produce a more natural sound audio than the current two-stage model. Through random modeling on hidden variables and using a random duration predictor, the diversity of synthesized speech is improved. The same text input can synthesize speech with different tones and prosody.

VITS can be regarded as a conditioned VAE for a maximized variational Lower Bound, namely ELBO (Evidence Lower Bound):

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z) - \log\frac{q_\phi(z|x)}{p_\theta(z|c)}\right] \tag{1}$$

In the formula, $p_\theta(z|c)$ and $q_\phi(z|x)$, respectively, represent the prior distribution of the potential variable z and an approximate posterior distribution, and $\log p_\theta(x|z)$ is the likelihood function of a data point x. The sum of reconstructed loss, $-\log p_\theta(x|z)$, and kl divergence, $\log q_\phi(z|x) - \log p_\theta(z|c)$, can be used as the loss of training, namely -ELBO.

VITS use mel-spectrogram instead of the original waveforms, represented by x_{mel}. $L1$ losses are calculated between the predicted and target mel-spectrogram as reconstruction losses in training:

$$L_{recon} = \| x_{mel} - \hat{x}_{mel} \|_1 \tag{2}$$

The kl divergence is as follows, where x_{lin} is the linear-scale spectrogram of target speech:

$$L_{kl} = \log q_\phi(z|x_{lin}) - \log p_\theta(z|c_{text}, A), \tag{3}$$

$$z \sim q_\phi(z|x_{lin}) = N(z; \mu_\phi(x_{lin}), \sigma_\phi(x_{lin}))$$

To estimate the alignment between the input text and the target speech A, VITS adopts the Monotonic Alignment Search (MAS), which maximizes the likelihood of data parameterized by a normalizing flow, f:

$$\begin{aligned} A &= \underset{\hat{A}}{\operatorname{argmax}} \log p(x|c_{text}, \hat{A}) \\ &= \underset{\hat{A}}{\operatorname{argmax}} \log N(f(x); \mu(c_{text}, \hat{A}), \sigma(c_{text}, \hat{A})) \end{aligned} \tag{4}$$

VITS can calculate the duration of each input tag, d, by adding all the columns in each of the estimated aligned rows. In order to achieve a more realistic rhythm of human speech, VITS adds a stochastic duration predictor and introduces two random variables, u and v, which have the same time resolution and dimension as the duration series, d, and are used for variational dequatization and variational data augmentation, respectively. The two variables are sampled by an approximate posterior distribution. The resulting goal is a variational lower bound on the logarithmic likelihood value of the phoneme duration. Training loss, L_{dur}, is the lower bound of negative variation:

$$\log p_\theta(d|c_{text}) \geq \mathbb{E}_{q_\phi(u,v|d,c_{text})}\left[\log\frac{p_\theta(d-u, v|c_{text})}{q_\phi(u, v|d, c_{text})}\right] \tag{5}$$

VITS uses adversarial training, with a discriminator, D, distinguishing between the output generated by the decoder, G, and the ground-truth waveform, y. Below is their loss, where T represents the total number of layers of the discriminator and D^l outputs the feature map of the l layer of the discriminator with N_l feature numbers:

$$Ladv(G) = \mathbb{E}_z\left[(D(G(z)) - 1)^2\right] \tag{6}$$

$$Lfm(G) = \mathbb{E}_{(y,z)}\left[\sum\nolimits_{l=1}^{T} \frac{1}{N_l} \parallel D^l(y) - D^l(G(z)) \parallel_1\right] \tag{7}$$

With the combination of VAE and GAN training, the total loss for training can be expressed as follows:

$$Lvae = Lrecon + Lkl + Ldur + Ladv(G) + Lfm(G) \tag{8}$$

The overall architecture of the model includes a posterior encoder, prior encoder, decoder, discriminator, and stochastic duration predictor, as shown in Figures 1 and 2.

The prior encoder consists of two parts, the text encoder and the normalized stream f_θ. The text encoder is a Transformer [22] encoder. The normalized flow can improve the variation range of prior distribution, so as to improve the final speech synthesis effect. A posteriori encoder uses the non-causal WaveNet [7] residuals block used in Glow-TTS [23]. The decoder is an improved HIFI-GAN [24] generator, which consists of multiple transposed convolutions with Multi-Receptive field Fusion (MRF). The random duration predictor estimates the distribution of phoneme duration based on conditional inputs. The discriminator uses the method of adversarial training, reconstructs the one-dimensional sequence into a two-dimensional plane through the sub-discriminator, and carries out two-dimensional convolution operation to judge the true probability of the sample, thus realizing the function of the discriminator.

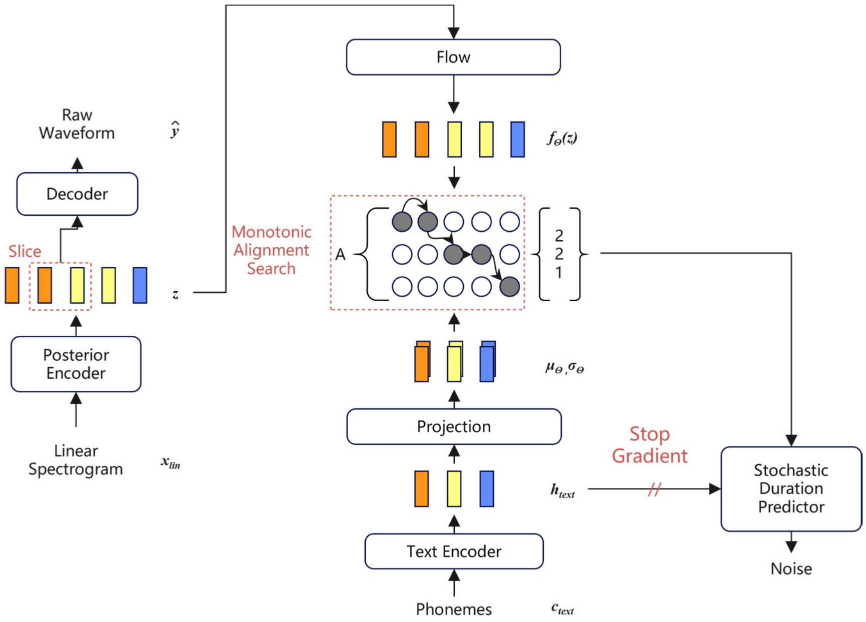

Figure 1. Training procedure of VITS. The slices of different colors represent different text embedding, and the number of slices of the same color represents the duration.

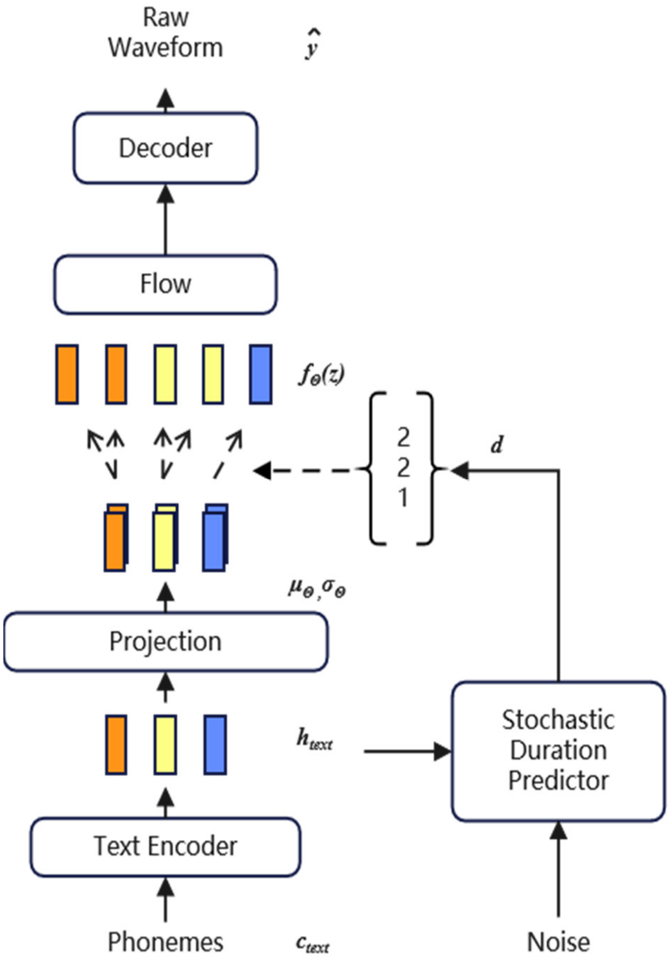

Figure 2. Inference procedure of VITS.

3. Emo-VITS

In order to synthesize emotion speech with high naturalness and taking into account different granularities of information, we incorporated the emotion network into VITS and propose the Emo-VITS system. Compared with other acoustic models, VITS has a very high synthetic naturalness [18]. The Emo-VITS model inherited the high performance of the VITS, and on this basis, the design of emotion network is concerned, including the emotion encoder network and emotion feature fusion module. The goal of the emotion network is to extract the emotion features of different receptive fields from the reference audio, integrate them based on the attention mechanism, and send them to the VITS module.

Figures 3 and 4 shows an overview of the Emo-VITS system. In the figure, Wav2Vec2.0 is used for feature extraction, a global emotion feature extractor and local emotion feature extractor is used for feature extraction of different granularities, and emotional feature fusion is used for feature fusion. In the training stage, the training audio as the raw audio is used as the emotion reference audio to extract the global and local emotion features. After feature fusion of the two feature vectors, they are added into the text encoder as emotion embedding and text embedding. In the prediction stage, emotional embedding is obtained according to the given reference audio and emotional speech is synthesized.

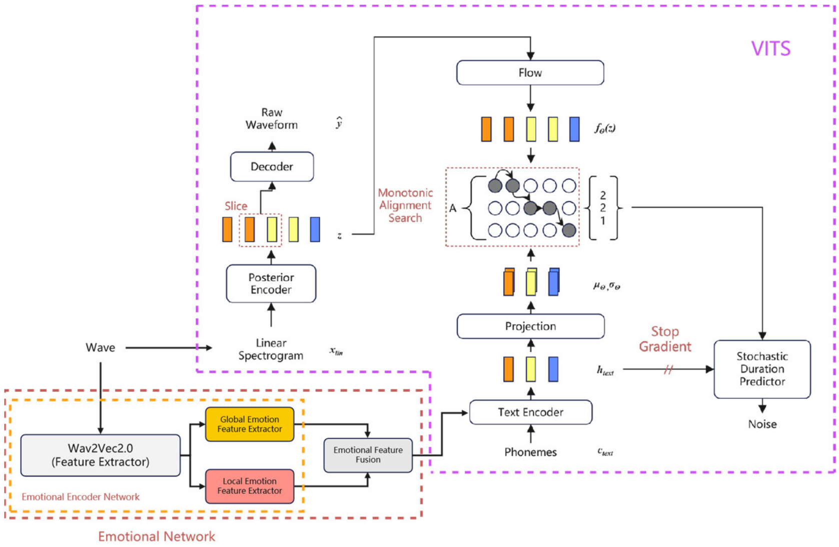

Figure 3. Training procedure of Emo-VITS.

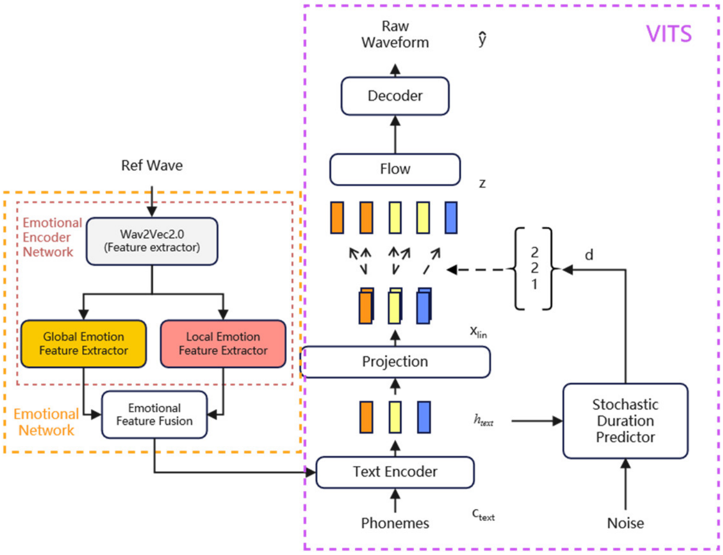
Raw Waveform
ŷ
VITS
Decoder
Flow
z
Ref Wave
Emotional Encoder Network
Wav2Vec2.0 (Feature extractor)
2
2
1
d
Global Emotion Feature Extractor
Local Emotion Feature Extractor
x_{lin}
Projection
Emotional Network
Emotional Feature Fusion
h_{text}
Stochastic Duration Predictor
Text Encoder
Phonemes
c_{text}
Noise

Figure 4. Inference procedure of Emo-VITS.

3.1. Emotional Encoder Network

The voice with emotion is extracted through wav2vec for emotion features. The global and local sentiment encoder networks of the features are carried out to obtain the global and local sentiment feature embedding, respectively. Figure 5 shows the structure of the emotional encoder network.

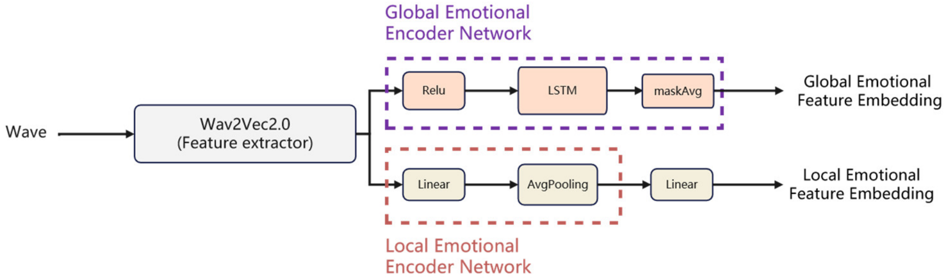
Global Emotional Encoder Network
Relu
LSTM
maskAvg
Global Emotional Feature Embedding
Wave
Wav2Vec2.0 (Feature extractor)
Linear
AvgPooling
Linear
Local Emotional Feature Embedding
Local Emotional Encoder Network

Figure 5. Emotional encoder network.

Wav2vec [25] in the figure is composed of an encoder and prediction model transformer. Wav2vec can be trained in an unsupervised manner to learn the representation vector of the audio and use the vector in the training of improved acoustic models. Wav2vec can also be used as a general feature extractor for speech. The resulting representation is used to improve acoustic model training and can be used as a general feature extractor for speech. Compared with traditional spectrum-based features, wav2vec 2.0 [26], as a transformer model, can extract features that are suitable for attention-based fusion. In addition, compared with the original waveform or spectrum features, features extracted from the network with wav2vec 2.0 make it more difficult to restore the original audio content and reconstruct the original reference audio. Therefore, using wav2vec 2.0 as a feature extractor can reduce the influence of text information of the reference audio [27].

Here we use the wav2vec pre-training model from [28]. This model was created by fine-tuning the pre-trained wav2vec2-large-robust model on the MSP-Podcast (v1.7). The pre-trained model was pruned from 24 to 12 transformer layers before fine-tuning.

The global emotion encoder network in the figure processes the output features of wav2vec 2.0 through a linear layer with ReLU activation. Then, through the single-layer LSTM [29], the culture is further improved. Finally, the feature sequences are aggregated into the mean vector representation after a certain step length by the MaskAvg module, and 192-dimensional global affective feature embedding is formed.

Different from the global emotion encoder network, the local emotion encoder network needs to keep the time information of feature extraction of Wav2vec 2.0. Specifically, the output characteristics of wav2vec2.0 are first changed through a linear layer, and then the representation is smoothed using average pooling. While expanding the Kernel field of perception, the local significant features are effectively preserved. Local emotional features contain time dimension, so the final extracted local emotional features and global emotional features will be matched and fused through the broadcast mechanism.

Wav2vec 2.0 extracts the emotion feature vectors of reference audio files, generates corresponding feature vectors through the local emotion encoder and global emotion encoder, respectively, and then enters the emotion feature fusion module below.

3.2. Emotional Feature Fusion Module

The purpose of this module is to fuse the global and local emotion features to facilitate the introduction of emotion features into the text encoder, as shown in Figure 6.

The calculation formula of feature fusion is as follows:

$$Z = \alpha \bigotimes X + (1 - \alpha) \bigotimes Y \tag{9}$$

where, X and Y are global and local emotion feature vectors, respectively. α is the new feature weight obtained by vectors X and Y through the feature fusion network, and the magnitude is between 0 and 1. It is the result obtained by the sum of X and Y, the sum of output obtained by the global and local attention modules, respectively, and then the result is obtained by Sigmoid.

In the training process, the local emotion embedding and the global emotion embedding are added to obtain a new vector, and then enter the local attention module and the global attention module, respectively. The partial attention module first passes through the convolution layer, then through the Batch Normolization layer, Relu layer, and finally through a convolution layer. The global attention module has one more global average pooling layer than the local attention module, so as to extract the global salient features. The sum of the results obtained by the two modules can be obtained by a sigmoid activation function to obtain the weight reassigned by the two parts of the features, and then the feature reweighted summation is carried out. The two parts of the network will determine their respective network weights in the training. In the process of prediction, the local and global emotional features of a single reference audio can be selected. We also provide local and global emotional signatures using different reference audio for a more three-dimensional effect.

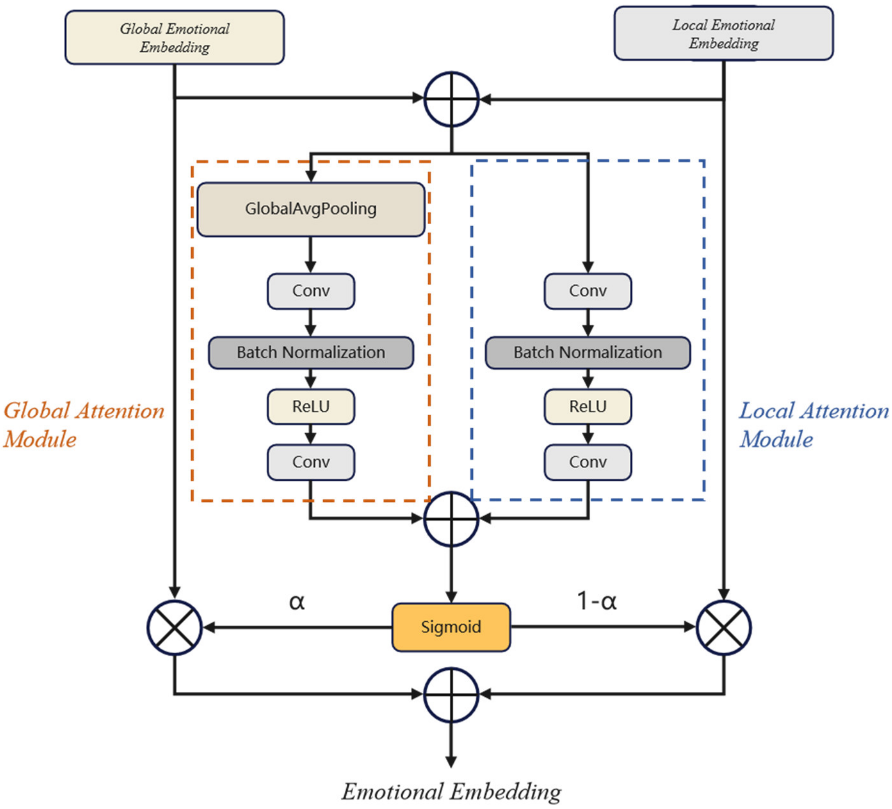

Figure 6. Architecture of the emotion feature fusion module.

4. Experiment and Result

4.1. Experimental Data and Training Configuration

In our experiment, we used two Chinese datasets. The first is a publicly available dataset of female speakers, which we call the Biaobei dataset. The dataset contains 10,000 Chinese female utterancess with a total duration of about 12 h in uncompressed PCM WAV format with a sampling rate of 48 kHz and 16 bit. Phonological proofreading, prosodic labeling, and Chinese vowel boundary segmentation were performed. It was divided into a training set of 9400 utterances, a test set of 500 utterances, and a verification set of 100 utterances. The dataset is mainly used to test the accuracy of synthesis. Since most of the expression of emotion in the open dataset was not strong enough, and the expression of emotion was often not obvious in the test of affective speech synthesis, we also used an emotionally intense Chinese dataset from online games, which was pronounced by female speakers, which we later called Pm. This dataset is for performance testing purposes only and will not be made public or commercially available. This dataset contains 2210 sentences of Chinese female voices, of which 2200 sentences are used, and the sampling rate is 22,050 Hz, 16 bit. It was divided into a training set of 1800 utterances, a test set of 300 utterances, and a verification set of 100 utterances. This dataset is derived from the actual use of life, so its emotional intensity is very strong. However, the amount of data in this dataset is not large, so it is more appropriate for us to use it for the test of emotional synthesis strength and subjective evaluation.

When training the model, the network training uses the AdamW optimizer [30], β1 = 0.8, β2 = 0.99, and weight attenuation λ = 0.01. The initial learning rate is 2×10^{-4}, and the decay time of the learning rate is $0.999^{1/8}$ factor. We trained on an NVIDIA RTX3060 GPU. The batch size was set to 16 and the model was trained to 200k steps. We also trained the LST-A [27] model for comparative experiments, not only because the author shared the source code, but also because the synthesis effect of this model is more excellent in recent years.

4.2. Objective Evaluation

Firstly, an accuracy test was conducted, which was mainly used to evaluate whether emotion embedding had an impact on the accuracy of speech synthesis itself. Generally speaking, compared with pure speech synthesis, the naturalness of emotion speech synthesis will decrease due to the addition of emotion factors. Therefore, what we want to evaluate with word error rate (WER) is whether the inclusion of emotional embedment will lead to the decrease of naturalness and accuracy of speech [27]. The random audio within the training dataset was used as the reference audio, with 100 synthesized sentences of audio, and the WER was tested through a speech recognition (ASR) code. The speech recognition code here is provided by Baidu AI, and the accuracy rate has reached more than 98.4%. The results are shown in Table 1.

Table 1. Objective evaluation of Biaobei and Pm datasets.

	Biaobei			**Pm**		
	VITS	Emo-VITS	LST-A	VITS	Emo-VITS	LST-A
WER (%)	8.1%	9.2%	14%	15.9%	19.6%	25.1%

As can be seen from the table, after the emotional network is added to the acoustic model VITS, WER increases, but the amplitude of increase is acceptable. In our test, WER did not increase much after the addition of the emotion module to the Biaobei dataset, but the increase in the Pm dataset was greater than that of Biaobei dataset. This may be because Pm is a homemade dataset. When the amount of data is small, the influence of adding an emotion module on the accuracy of synthesis will be amplified. It is worth mentioning that the LST-A structure is not effective in this accuracy experiment, and the error rate is very high. We believe that the reason is that LST-A's TransformerTTS is not robust enough, and its performance is not as stable as that of VITS in the case of small data. This is also an advantage of our use of the VITS model.

Next, we did an emotional classification test. First of all, we used the audio in the test dataset Pm as the reference audio for corresponding emotion synthesis, and then used the pre-trained emotion recognition model to detect whether our emotion speech synthesis based on reference audio can restore the emotion. The accuracy of the speech emotion recognition model used here is around 80% [31].

In emotion recognition, *WA* (Weighted Accuracy) and *UA* (Unweighted Accuracy) are generally used to measure the accuracy rate. The formula is as follows:

$$WA = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i} \times 100\% \tag{10}$$

where N represents the total number of categories. TP represents the sample with a positive true value and a positive predicted value, and FP represents the sample with a negative true value and a positive predicted value.

$$acc_i = \frac{TP_i}{TP_i + FP_i} \times 100\% \tag{11}$$

$$UA = \frac{1}{N} \sum_{i=1}^{N} acc_i \times 100\%$$

where, *acc* represents the accuracy rate of each category, and N represents the total number of categories.

A total of 100 sentences of audio were selected as reference audio in the dataset. We controlled the same text information. At the same time, in order to minimize the impact brought by the accuracy of the emotion recognition model itself, we took the result of the reference audio used for testing after passing the emotion recognition model as the correct emotion. The total accuracy of emotion synthesis and the accuracy of emotion transfer are shown in Table 2.

Table 2. Emotion objective evaluation of Emo-VITS.

Angry	Happy	Fear	Sad	Surprise
100%	72.7%	46.9%	28.2%	50%
Total Accuracy (*WA*) 61%				
Total Accuracy (*UA*) 59.6%				
Emotion objective evaluation of LSA-A				
Angry	Happy	Fear	Sad	Surprise
67%	34.6%	39.1%	75.0%	30%
Total Accuracy (*WA*) 48.5%				
Total Accuracy (*UA*) 49.1%				

It can be seen that under the classification test of five emotions, the model in this paper can finally achieve 61% accuracy of emotion synthesis. That is, after more than half of the audio is used as the reference audio, the output audio still retains the original emotional features and is classified into the same category in the emotional classification. Interestingly, the model in this paper has a high accuracy for Angry and Happy, but it is not accurate for Sad synthesis. We think it may be because when Sad is used as a reference audio mood, its synthesized audio is discriminated into the Fear category because of the obvious ups and downs of the sound, which leads to a decrease in accuracy. As for the LSA-A model, it seems that the synthetic effect of Happy and Surprised is not good.

Finally, we use the Emotional Speech Dataset (ESD) [32] to do the same experiment. This time, we used the test audio of five emotions for each of the eight speakers in the dataset, with 30 sentences for each emotion and a total of 1200 sentences. Similarly, we controlled the text information the same way, and in order to minimize the impact brought by the accuracy of the emotion recognition model itself, we used the reference audio for testing as the correct emotion after passing the emotion recognition model. Our overall classification accuracy for audio synthesis using the ESD dataset was 52.67%.

4.3. Subjective Evaluation

We evaluated the naturalness and emotional similarity of our system through a subjective evaluation test. Twenty groups, consisting of text and reference speech, were randomly selected from the test set, and a total of 40 pieces of audio were synthesized by the two emotion systems. The reference audio and synthetic speech were grouped, and the 13 workers who had never touched speech synthesis rated the naturalness and emotional similarity of the speech on a scale of 5, respectively. The naturalness was from 1 (bad) to 5 (excellent), and the emotional similarity was from 1 (not similar at all) to 5 (highly similar). Each audio was rated by different workers, and we reported an average opinion score (MOS) and a 95% confidence interval. In the table, Emo-VITS (Global) represents that only global vectors are used as the result of emotion embedding, while Emo-VITS (Local) represents that only local vectors are used as the result of emotion embedding. The results are shown in Table 3.

From this table, we can see that our model is obviously better than LST-A in terms of naturalness, although there is still a big gap with the ground truth. Through experiments, we found that our advantage over LST-A was reflected in long statement modeling. When the text information is too long, LST-A often has the phenomenon of repetition and missing words. The reason this happens is that the alignment of the TransformerTTS depends on the attention mechanism, and it is not stable enough. This kind of problem can greatly affect the naturalness of speech. In contrast, Emo-VITS is more stable. Even if there is a problem, it will only appear in the pronunciation of a certain word and will not cause a greater impact on the overall structure of the sentence, so the negative effect is limited.

Table 3. MOS with 95% confidence interval on Pm datasets.

	Naturalness	**Emotion Similarity**
Ground Truth	4.01 ± (0.14)	
Emo-VITS	3.24 ± (0.16)	3.26 ± (0.15)
LST-A	2.83 ± (0.14)	3.01 ± (0.15)
VITS + GST	3.02 ± (0.14)	2.60 ± (0.14)
Emo-VITS (Global)	3.01 ± (0.13)	2.10 ± (0.14)
Emo-VITS (Local)	2.71 ± (0.14)	2.71 ± (0.14)

However, under the same acoustic model of Emo-VITS and VITS + GST, Emo-VITS's emotional similarity is significantly better than VITS + GST's, which can show that the Emo-VITS embedding method is better. By comparing the separate embedding results of the global and local vectors of Emo-VITS, it can be found that the global emotion vector of Emo-VITS focuses on maintaining the naturalness of the whole statement, while the local emotion focuses on the expression of emotion.

5. Conclusions

We have developed a VITS-based emotional speech synthesis model, called Emo-VITS, to better serve the IoT. We evaluated the accuracy, naturalness, and emotional similarity of the synthesized speech, and the subjective and objective evaluations prove that the Emo-VITS system can achieve significant results in characterizing emotions. We took advantage of the VITS structure in synthetic speech. The global emotional feature embedding and local emotional feature embedding of the reference audio are extracted separately by designing an emotional encoder network, which is used to better express the emotional factors in the reference audio. The attention-based mechanism is used to design the emotional feature fusion module, which not only preserves the emotional information, but also facilitates the embedding into the text encoder. We plan to continue to improve the method of emotional features extraction and fusion and refine our dataset in the future based on the specific application scenarios of IoT.

Author Contributions: Conceptualization, W.Z.; methodology, Z.Y.; implementation, W.Z. and Z.Y.; supervision, W.Z.; writing—original draft, W.Z. and Z.Y.; writing—review and editing, W.Z. and Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This study is supported by the Fundamental Research Funds for the Central Universities (Grant No.CUC220B013 and CUC21GZ009), the National Key R&D Program of China (Grand No.2022YFC3302100) and the National Natural Science Foundation of China (Grant No.62207029).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

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

References

1. Fang, W.; Zhu, C.; Yu, F.R.; Wang, K.; Zhang, W. Towards Energy-Efficient and Secure Data Transmission in AI-Enabled Software Defined Industrial Networks. *IEEE Trans. Ind. Inform.* **2022**, *18*, 4265–4274. [CrossRef]
2. Fang, W.; Cui, N.; Chen, W.; Zhang, W.; Chen, Y. A Trust-based Security System for Data Collection in Smart City. *IEEE Trans. Ind. Inform.* **2021**, *17*, 4131–4140. [CrossRef]
3. Fang, W.; Zhang, W.; Yang, W.; Li, Z.; Gao, W.; Yang, Y. Trust management-based and energy efficient hierarchical routing protocol in wireless sensor networks. *Digit. Commun. Netw.* **2021**, *7*, 470–478. [CrossRef]
4. Bi, M.; Lu, H.; Zhang, S.; Lei, M.; Yan, Z. Deep Feed-Forward Sequential Memory Networks for Speech Synthesis. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018. [CrossRef]

5. Xu, G.; Song, W.; Zhang, Z.; Zhang, C.; He, X.; Zhou, B. Improving Prosody Modelling with Cross-Utterance BERT Embeddings for End-to-end Speech Synthesis. *arXiv* **2020**, arXiv:2011.05161. [CrossRef]
6. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R. Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.
7. Oord, A.V.D.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A. WaveNet: A Generative Model for Raw Audio. In Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop, Sunnyvale, CA, USA, 13–15 September 2016.
8. 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. In Proceedings of the Interspeech, Stockholm, Sweden, 20–24 August 2017. [CrossRef]
9. Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech: Fast, Robust and Controllable Text to Speech. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
10. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. FastSpeech 2: Fast and High-Quality End-to-End Text-to-Speech. *arXiv* **2020**, arXiv:2006.04558. [CrossRef]
11. Li, N.; Liu, S.; Liu, Y.; Zhao, S.; Liu, M. Neural Speech Synthesis with Transformer Network. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 6706–6713.
12. Liu, R.; Sisman, B.; Li, H. Reinforcement Learning for Emotional Text-to-Speech Synthesis with Improved Emotion Discriminability. *arXiv* **2021**, arXiv:2104.01408. [CrossRef]
13. Cai, X.; Dai, D.; Wu, Z.; Li, X.; Li, J.; Meng, H. Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition. *arXiv* **2020**, arXiv:2010.13350. [CrossRef]
14. Skerry-Ryan, R.J.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.; Clark, R.; Saurous, R.A. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *arXiv* **2018**, arXiv:1803.09047. [CrossRef]
15. Stanton, D.; Wang, Y.; Skerry-Ryan, R. Predicting Expressive Speaking Style from Text in End-to-End Speech Synthesis. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 18–21 December 2018; pp. 595–602.
16. Wang, Y.; Stanton, D.; Zhang, Y.; Ryan, R.S.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; Saurous, R.A. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv* **2018**, arXiv:1803.09017. [CrossRef]
17. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2014**, arXiv:1312.6114. [CrossRef]
18. Kim, J.; Kong, J.; Son, J. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *arXiv* **2021**, arXiv:2106.06103. [CrossRef]
19. Tan, X.; Qin, T.; Soong, F.; Liu, T.Y. A Survey on Neural Speech Synthesis. *arXiv* **2021**, arXiv:2106.15561. [CrossRef]
20. Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; Barnard, K. Attentional Feature Fusion. *arXiv* **2020**, arXiv:2009.14082. [CrossRef]
21. Rezende, D.J.; Mohamed, S. Variational Inference with Normalizing Flows. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1530–1538.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
23. Kim, J.; Kim, S.; Kong, J.; Yoon, S. Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 8067–8077. [CrossRef]
24. Kim, J.; Bae, J.; Kong, J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17022–17033.
25. Schneider, S.; Baevski, A.; Collobert, R.; Auli, M. wav2vec: Unsupervised Pre-training for Speech Recognition. *arXiv* **2019**, arXiv:1904.05862.
26. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *arXiv* **2020**, arXiv:2006.11477. [CrossRef]
27. Chen, L.W.; Rudnicky, A. Fine-grained style control in Transformer-based Text-to-speech Synthesis. *arXiv* **2021**, arXiv:2110.06306v2.
28. Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Eyben, F.; Schuller, B.W. Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *arXiv* **2022**, arXiv:2203.07378.
29. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting*; MIT Press: Cambridge, MA, USA, 2015.
30. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.
31. Available online: https://github.com/Renovamen/Speech-Emotion-Recognition (accessed on 1 October 2022).
32. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset. In Proceedings of the 2021 IEEE ICASSP International Conference on Acoustics, Speech and Signal Processing IEEE, Toronto, ON, Canada, 6–11 June 2021.

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.