Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.026234
Article

Tech Science Press

Emotional Vietnamese Speech Synthesis Using Style-Transfer Learning

Thanh X. Le, An T. Le and Quang H. Nguyen*

School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, 10000, Vietnam
*Corresponding Author: Quang H. Nguyen. Email: quangnh@soict.hust.edu.vn
Received: 19 December 2021; Accepted: 21 February 2022

Abstract: In recent years, speech synthesis systems have allowed for the production of very high-quality voices. Therefore, research in this domain is now turning to the problem of integrating emotions into speech. However, the method of constructing a speech synthesizer for each emotion has some limitations. First, this method often requires an emotional-speech data set with many sentences. Such data sets are very time-intensive and labor-intensive to complete. Second, training each of these models requires computers with large computational capabilities and a lot of effort and time for model tuning. In addition, each model for each emotion failed to take advantage of data sets of other emotions. In this paper, we propose a new method to synthesize emotional speech in which the latent expressions of emotions are learned from a small data set of professional actors through a Flowtron model. In addition, we provide a new method to build a speech corpus that is scalable and whose quality is easy to control. Next, to produce a high-quality speech synthesis model, we used this data set to train the Tacotron 2 model. We used it as a pre-trained model to train the Flowtron model. We applied this method to synthesize Vietnamese speech with sadness and happiness. Mean opinion score (MOS) assessment results show that MOS is 3.61 for sadness and 3.95 for happiness. In conclusion, the proposed method proves to be more effective for a high degree of automation and fast emotional sentence generation, using a small emotional-speech data set.

Keywords: Emotional speech synthesis; flowtron; speech synthesis; style transfer; vietnamese speech

1 Introduction

Speech is a means of communication through language, the most basic human tool to help us communicate, express emotions, and thoughts as well as exchange experiences and information. As society and technology develop, machines are increasingly invented to replace human labor, increasing the need for communication between humans and machines. Speech processing hence has become one of the most important and concerning domains.

Text-to-speech or speech synthesis (TTS) is a technique for converting input text into a time-domain speech signal to change a given text into natural, intelligible speech. This topic has been researched for quite some time. The voice processing community is very large, regularly releasing high-quality

publications. Participating in these studies requires knowledge of the language and how the human voice is generated and involves many areas, including linguistics, acoustics, digital signal processing, and machine learning. Currently, researchers in the field of speech processing in general and speech synthesis systems have made many remarkable achievements in most languages. Along with that, many useful products are research results have been widely applied in practice, such as Apple's Siri system, Amazon's Alexa, and Microsoft's Narrator.

Speech is effective when it expresses the speaker's content and emotions. Therefore, the inclusion of prosody and emotions in synthetic speech will help improve the efficiency of human-machine communication. Emotions are a complex human phenomenon and are difficult to distinguish directly. In a sentence, there are many ways to express emotions: through gestures, facial expressions, eyes, intonation, etc. Within the same sentence, the speaker can also express distinctive emotions. Moreover, human emotions are often mixed and difficult to distinguish clearly.

In the past few decades, popular speech synthesis methods have included speech synthesis by concatenation [1,2], speech apparatus simulation [3], and synthesis based on Hidden Markov Model (HMM) statistical parameters [4,5]. Today, with the development of artificial intelligence, deep learning has been applied to both emotion recognition [6,7] and emotional-speech synthesis, and neural network-based TTS has greatly improved the quality of natural synthesized speech (Tacotron [8], Deep Voice 3 [9], or FastSpeech [10]). Therefore, researchers are now attempting to integrate emotions into speech. The currently popular method is to use a large data set for each emotion [11–13]. Zhou et al. [13] used the IEMOCAP data set [14] to train the emotion recognition model and then used this model to extract parameters for each emotion, and Wasserstein's generative adversarial network (VAW-GAN) was trained to convert neutral-state sentences to emotional-speech sentences. Kwon et al. [12] used an emotional-speech corpus (Korean corpus) to train the GST model (global style tokens). The output of this model was used as emotion embedding, combined with the transcript embedding fed into the Tacotron 2 model to synthesize speech with corresponding emotions, the same method Wu et al. [11] used with an emotional Chinese corpus. However, the method of building a speech synthesizer for each emotion has some limitations when applied to new languages. First, this method often requires a large emotional-speech data set, and such data sets are time-and labor-intensive to complete [14]. Second, training each of these models requires computers with large computational capabilities and a lot of effort and time for model tuning. In addition, training each model for each emotion fails to take advantage of data sets of other emotions. Therefore, this method is very difficult to implement for new emotions in new languages.

In this paper, we propose a new method to synthesize emotional speech in which the latent expressions of emotions are learned from a small data set produced by professional actors through a Flowtron model. In addition, we provide a new method to build a speech corpus that is scalable and whose quality of corpus is easy to control. Next, to produce a high-quality speech synthesis model, we used this data set to train the Tacotron 2 model, which was used as a pre-trained model to train the Flowtron model. The proposed method proves to be more effective for a high degree of automation and fast emotional sentence generation, and it uses a small emotional-speech data set. In Vietnam, language processing has received a lot of attention, research, and development, including Vietnamese synthesis. In recent years, Vietnamese synthesis systems have achieved many remarkable achievements [4–6]. However, researchers have conducted few studies on integrating emotions into synthetic speech. For this reason, we applied this method to synthesize Vietnamese speech with sadness and happiness.

We organize the rest of the paper as follows. Section 2 describes related work. We present the emotional corpora used in this paper, emotional Vietnamese speech synthesis system based on Flowtron, in Section 3. We provide the experimental results and discussions in Section 4. Finally, Section 5 presents the conclusions and perspectives.

2 Related Work

It can be said that the initial successes in speech synthesis are based on hidden Markov models [5–7]. Associated with the development of artificial intelligence, speech synthesis models using neural networks have been proposed, such as DeepVoice [15] and DeepVoice 2 [16]. These systems eventually completely replaced traditional studies based on hidden Markov models. Next is the end-to-end model-building phase. This is a complete system from start to finish, with almost no intermediaries. In this system, the input will be text content and the output will be a time-domain waveform audio file. Some prominent studies can be mentioned, such as Tacotron [8], Deep Voice 3 [9], and FastSpeech [10].

Regarding Vocoder selection, WaveNet is the first model we considered to apply this new method. NVIDIA invented WaveNet and noticed that WaveNet can work efficiently on data with 10000 samples per second [17]. When evaluated by actual users, the system can produce sounds with characteristics of various voices and languages. In 2018, 1 year after the launch of WaveNet, NVIDIA introduced WaveGlow [18], which is a flow-based network capable of generating speech from the Mel spectrogram. WaveGlow is a combination of Glow and WaveNet to produce high-quality sound with less computation due to reduced network architecture and several model parameters. WaveGlow consists of a single network trained with only one cost function. It has improved computational and storage complexity.

To achieve a human-like sound, pitch, duration, emphasis, rhythm, and the speaker's style and emotions are extremely important factors. Previous models did not control for these factors, and Flowtron was created to address them [19]. Flowtron allows to transform the training speaker's style and emotion into the target speaker's style and emotion in the target speaker's conversation or to interpolate the training voice with the target speaker's emotions and style, saving a lot of money and resources and enhancing the system's scalability.

Therefore, the common feature of neural network-based systems is that the sound produced is of high quality and intelligibility and portrays a naturalness approximating that of a human voice. Pre-processing is also not as complex as other traditional speech synthesis methods.

Along with the development of deep learning, speech synthesis studies have shown that synthesized sounds are better than those produced by traditional Markov modeling methods in emotional transformation. A semi-supervised speech synthesis method encoding global style vectors (GST) has been proposed to tune emotional data [11,20–23]. These studies build on a combination of previous frameworks on speech synthesis, such as Tacotron or Tacotron 2, and Style Token representing emotions. Then, a method of representing stylistic characteristics was proposed and applied in the control and transformation of emotions [12]. To learn the latent representations of the speaker's style, the use of a variational autoencoder (VAE) recognition network model is proposed, after which these latent representations are passed into the Tacotron architecture via WaveNet to generate voice.

In practice, a simple text-only system input is not sufficient to produce a well-expressed emotional statement. Many factors affect emotion in a sentence that cannot be completely labeled, such as intonation, stress, rhythm, and the speaker's style; all these factors are collectively known as prosody. Otherwise, a prosody modeling solution that did not require specific labeling represented an architecture capable of extracting prosody from input audio to generate latent prosody representations [24]. Specifically, they proposed a coding architecture that exposes a latent prosody space and demonstrates the ability to capture speech variation in that space.

With the one-hot vector emotion-representation approach, the model will have some limitations, such as not being able to work with unseen emotions, such as emotions not included in the training data and therefore not defined during training. Another disadvantage is that one-hot vectors often do not carry much useful information for the model. Therefore, a model architecture uses a pre-trained model to represent emotional attributes in sentences in a different spatial domain [13].

Researchers studying the synthesis of Vietnamese with a neutral voice have made many remarkable achievements with a sound quality close to natural speech [25–27]. Meanwhile, systems used to synthesize Vietnamese speech with emotions have not appeared in many published works. Some of these publications are made in combination with external sources of information, such as facial expressions and gestures, notably an experiment to model Vietnamese intonation with multi-modal corpus to synthesize expressive Vietnamese [25] or research on system integration into virtual characters and human-machine interaction [27]. In this study, we collected data with a scenario consisting of 19 sentences containing five emotions, neutral, happy, sad, slightly angry, and very angry, with a male voice and a female voice.

Many parameters of speech directly affect the emotion of language, such as the spectral envelope, pronunciation duration, volume, spectral energy structure, and sound quality, for example for emotional Vietnamese speech [28]. In this study, we introduce the BKEmo corpus with diverse emotions and several speakers as well as characteristic parameters such as fundamental frequency F0 and speech energy affecting emotions in speech. We also statistically analyze the differences in emotions according to those characteristic parameters. We used the BKEmo data set to perform speech emotion recognition based on CapsNet [6].

3 Emotional Vietnamese Speech Synthesis Proposal Method

In this section, we will present a method to build an emotional Vietnamese speech synthesizer based on style transfer and the Flowtron model.

3.1 Dataset Building

To build a high-quality speech synthesizer, we first need a high-quality data set from a speaker. We have five motivations and five criteria for using a personal data set:

- Ensure stable quality, reduce noise
- Uniformity: The corpus has a similarity between the textual content and the corresponding audio sentence even in readings of different sentences.
- A single voice: A data set with only one voice in which the set from that voice is extensive instead of a data set with many voices in which the data for each voice is limited.
- Diverse vocabulary: The data set has a large number of words, focusing on common words in Vietnamese, including spoken and written languages.
- Scalability: When a word appears that the dictionary does not contain, we can quickly collect the speaker's sound to practice pronouncing that new word.

We propose a new method to build this data set automatically, which is depicted in Fig. 1. First, we collected Vietnamese writings online from the Internet. Each sentence ends with a period (.), a question mark (?), an exclamation mark (!), or a comma (,). From these sentences, we will remove abbreviations, foreign languages, misspellings, special characters, and numbers while still ensuring the sentences are clear. Next, we will correct the spaces between words to ensure that we include only one space. We will convert sentences to lowercase letters. Finally, we will review the content to make sure the sentences make sense.

To create the voice, we used a free text-to-speech tool (from https://ttsfree.com/), which allows us to create audio sentences corresponding to the pre-existing text in various languages. The sound is made with a male voice. The synthesized sentences' emotion, speed, and pitch are all at a normal level. For each audio file, we recorded the sentence with a sampling frequency of 22050 Hz and 16 bits per sample. We will remove silences from the beginning and end of each sentence so that the speech synthesis model can be focused on the important part of the sentence and to reduce the amount of memory required.

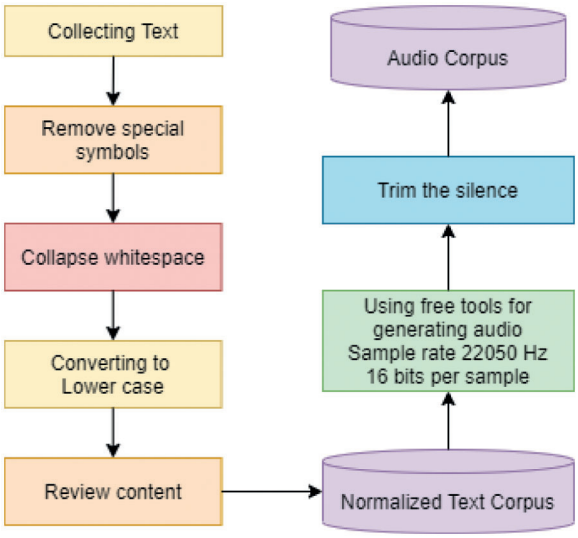

Figure 1: The method of building the AnSpeech data set

We labeled the data set AnSpeech. In AnSpeech, the lengths of the sentences varies from two to 80 words. Among them, about 3500 sentences contain fewer than 10 words. Sentences containing 30 words or more account for a small proportion, with about 1000 sentences. The AnSpeech data set includes 9796 sentences, corresponding to 15.97 h of speech and 2.4 GB of data.

3.2 Model of Neutral Speech Synthesis

Fig. 2 depicts our narrative Vietnamese speech synthesis model. The first speech synthesis model is Tacotron 2. This component's main task is to create a pre-trained model for the Flowtron model. The Tacotron 2 model consists of two main modules: the first module predicts the Mel spectrogram from the text, and the second module uses the Mel spectrogram to reconstruct the audio. In the first part, the input text is encoded into a string of characters, and the attention layer helps the model to learn important parts in the audio. The decoder block uses the sequence-to-sequence network to predict the Mel spectrogram. Finally, the WaveNet model acts as an encoder to generate a time-domain waveform signal from the Mel spectrogram.

The second component is the Flowtron model. Both the Tacotron 2 model and the Flowtron model generate a sequence of Mel spectrogram frames from the input character sequence. However, the Flowtron model allows for enhanced control over speech characteristics as well as the ability to transform between voices. To do so, the Flowtron model adds an Afin coupling layer, which includes three parts: Attention, Decoder, and Afin transform. With the Afin transform, the model can learn the reversibility of the function that maps the characteristic distribution of the speech (Mel spectrogram, text) to the z-parameter latent space by Gaussian distribution. In the latent space z, we can find and select various regions, from which we can generate corresponding voice feature samples (Mel spectrogram) with that selection.

The Afin coupling is performed K times. For odd values (K = 1, 3, 5 …), the network will perform forward propagation from beginning to end of the sentence and vice versa; for even values (K = 2, 4, 6…), the network performs back-propagation and learns from the end of the sentence to the beginning. The purpose of this approach is to control the model's learning ability, as the model can learn the attention of forwarding and back-propagation at the same time. Therefore, when we conducted the experiment, we performed the training process step by step, starting with K = 1 so that the model could converge more quickly, followed by K = 2 to improve the quality and accuracy. If the quality is still not good, we can increase K gradually until the produced sound meets the expectation.

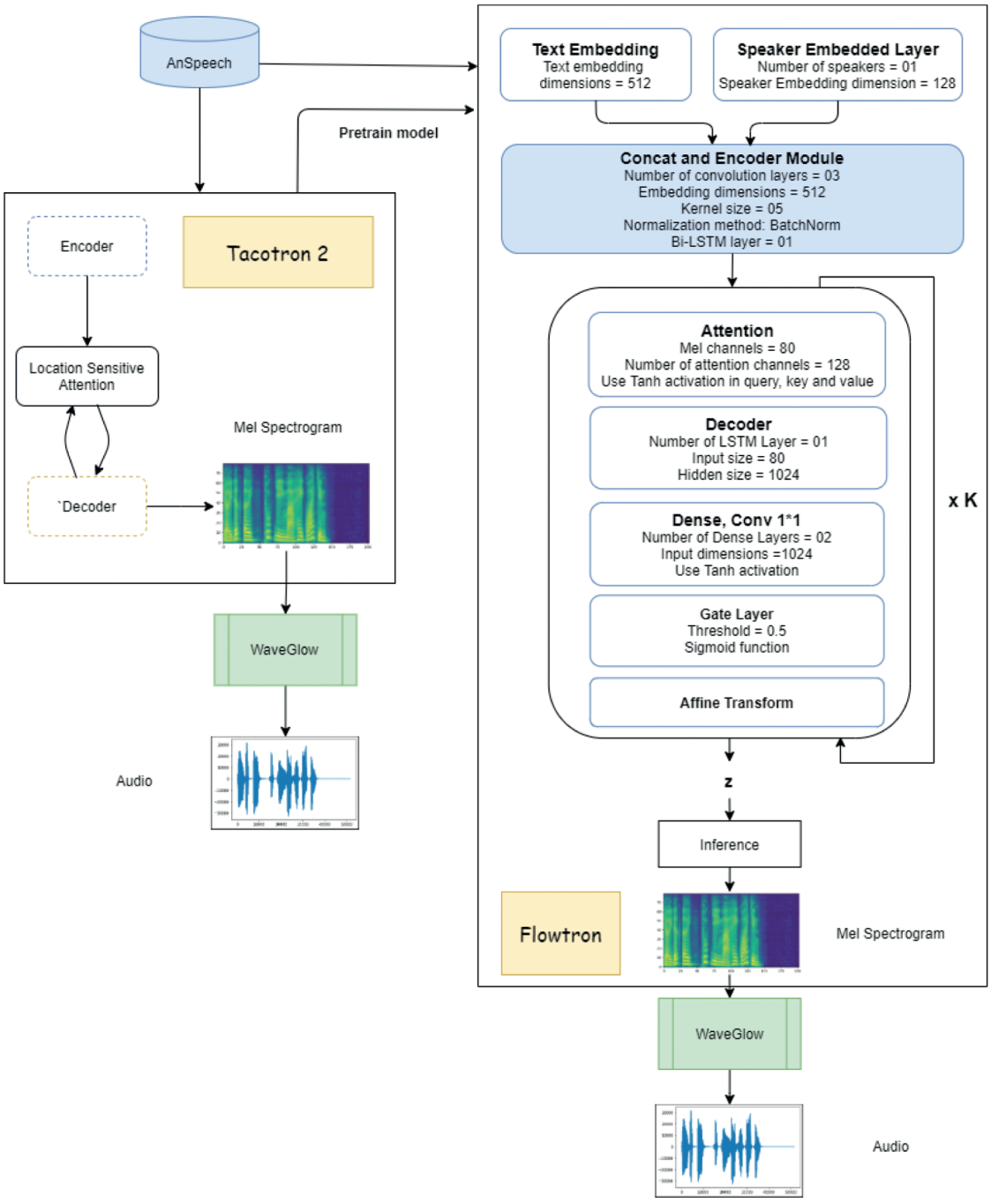

Figure 2: The model architecture of Vietnamese narrative speech synthesizer

3.3 Z-Space

Flowtron is a model capable of mapping from the spatial domain of the Mel spectrogram of speech patterns to the latent Z-space. We use a zero-mean spherical Gaussian distribution for the Z-space domain. In Fig. 3, the spatial domain consists of large circles, with the blue points representing various speech patterns, all of which are in a normal distribution with a mean of 0. Changing the variance helps

to change the propagation around the center point of the region in the sampled Z-space. Setting the variance to zero means that no change in speech is produced. Sampling from specific speech styles is equivalent to sampling from a specific region in Flowtron's Z-space, and for all of that particular speech, styles correspond to a different Gaussian distribution. Red circles with red dots represent the target emotional and style distribution area.

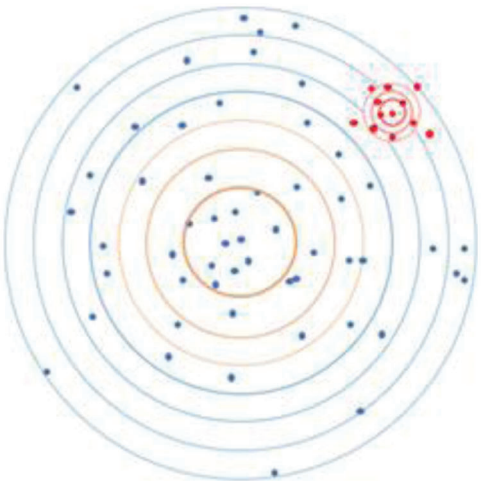

Figure 3: Example illustrating the transformation in the spatial domain Z

3.4 Methods of Synthesizing Emotional Speech

Emotional-speech synthesis becomes extremely difficult if emotional-speech data sets are required because emotions are temporary expressions depending on the speaker's mood. Therefore, in this study, we propose a new emotional-speech synthesis method, which has the main advantage of using a small set of emotional-speech data. This data set can be easily obtained by two methods. With the first method, voice data will be recorded in a real environment (maybe accompanied by a video); then we listen to these voices again to label emotions manually. This is also the method used to build the German emotional-speech data set (EmoDB data set) [29]. The biggest advantage of this method is its high degree of naturalness. However, the disadvantage is that the number of emotions is not equally distributed as well as the limited content of the text sentences. In the second method, artists express the emotion of the voice sentence. This method has the advantage of being able to generate emotional-speech sentences with selected textual content as well as being able to balance the data set by emotion. So we used the second approach to build a small set of emotional-speech data sets.

Fig. 4 shows the architecture of our proposed emotional-speech synthesizer. First, the Mel spectrogram of each emotional-speech sentence in the database is analyzed; then this Mel spectrogram, together with a transcript of sentences, is fed into the Flowtron model to obtain vectors in the hidden space **Z** domain. These values will be stored in the emotional-speech database. To synthesize a voice sentence with a specific emotion from the input text, we perform the following steps:

- Step 1. Extract the **Z** vectors of the sentences in the data set that have the same emotion as that of the sentence being synthesized.
- Step 2. Pass the text input into the Flowtron model and perform the forward step. This step will create a Z vector in the hidden space corresponding to the input text.
- Step 3. Combine this **Z** vector (obtained in step 2) with the **Z** vectors obtained in step 1 to produce a set of **Z** vectors. Calculate the **Z**-means of this set.

- Step 4. Sampling the new **Z** vector from a normal distribution with **Z** mean, the value of variance is chosen implicitly. Thus, we have mapped the **Z** value of the input text sentence in the original distribution (the distribution corresponding to the neutral sentences) to the new **Z** value in the new distribution (the distribution corresponding to the emotions to be synthesized).
- Step 5. Pass the new **Z** value and text input into the Flowtron model and perform the inference step. Then we will get the Mel Spectrogram of the emotional-speech sentence to be synthesized.
- Step 6. Passing this Mel Spectrogram through the WaveGlow model, we get the audio of the emotional-speech sentence to be synthesized.

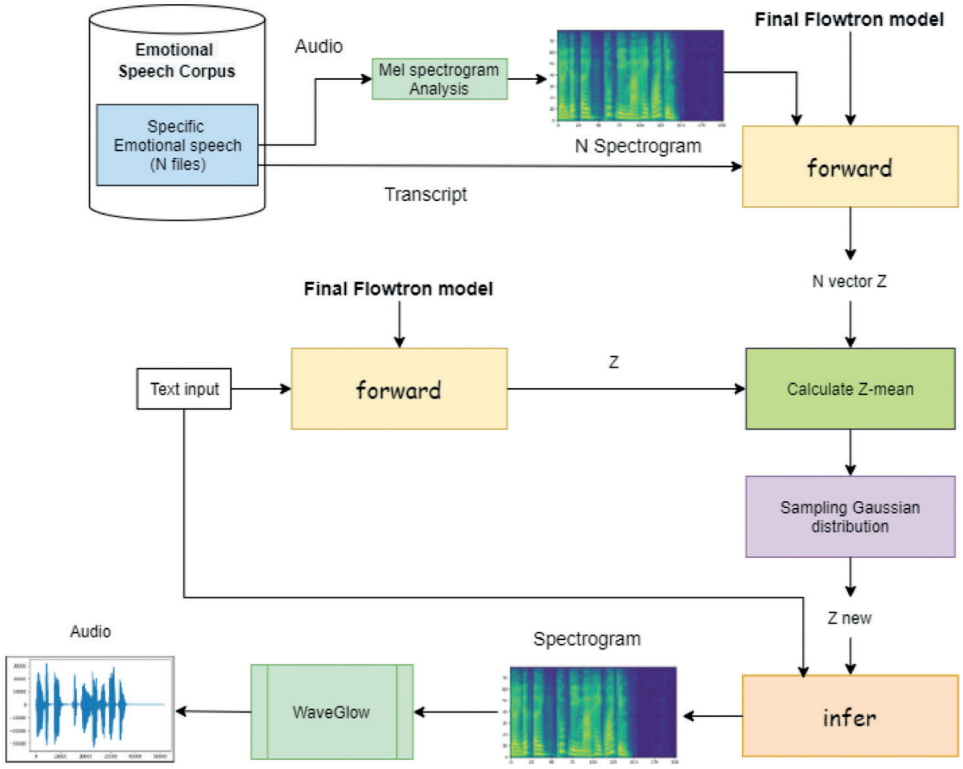

Figure 4: The model architecture of Vietnamese speech synthesizer with emotions

4 Results and Discussions

4.1 Experiment Environment

In association with the development of machine learning, models have an increasingly complex structure and storage capacity, and they require several calculations. The construction of the experimental environment is also an important factor. Using supercomputers will save a lot of time in research.

We performed the training process on a supercomputer with the following configuration:

- CPU: Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)
- RAM: 40 GB
- GPU: NVIDIA A100
- Hard drive: SSD 1 TB

We performed the process of synthesizing speech sentences for testing purposes on a Google Cloud service with the following configuration:

- CPU: N2 General, 8vCPU
- RAM: 8 GB
- GPU: NVIDIA Tesla P4
- Hard drive: SSD 7.5 GB

These environments operate with the Ubuntu 20.04 LTS operating system. We developed the models using the PyTorch platform.

4.2 Training Vietnamese Neutral Synthesizer

Fig. 5 depicts the process of training the narrative Vietnamese synthetic model. First, we selected 100 random sentences containing 10 to 30 words from the AnSpeech data set to make up the Validation set. We used the remaining sentences in the AnSpeech data set as the training set. The process of training each model continues until the error on the Validation set decreases to a small and stable value. When the error value converged (the difference between adjacent errors is not greater than 0.001), we selected the model obtained at that time for evaluation. We performed this evaluation process by listening to several randomly synthesized sentences, usually selected from daily conversations and common sentences in online articles.

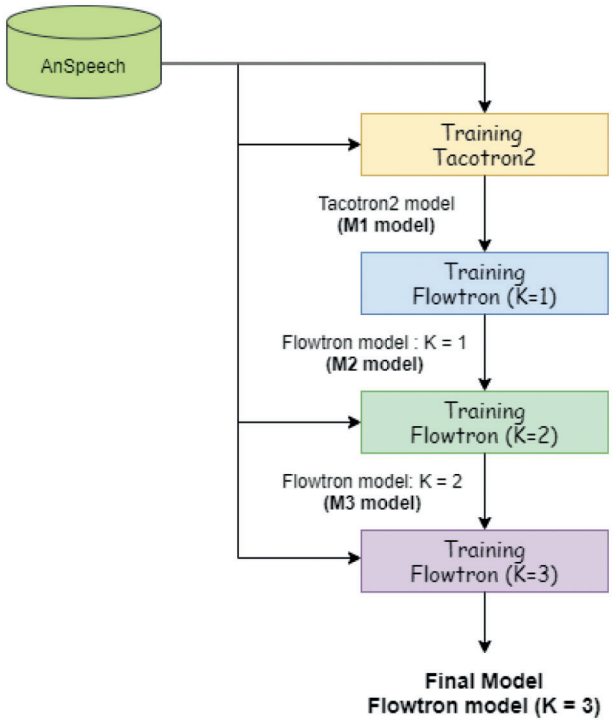

Figure 5: The process of training the narrative Vietnamese speech synthesis model

The model-training process includes the following steps:

- We first trained the Tacotron 2 model using the AnSpeech data set. We conducted the training without using pre-trained models; learning rate = 1e-3, weight decay = 1e-6, with size for each batch of 12.

Using this batch size ensures maximum utilization of the computer's 40 GB GPU memory capacity while still ensuring that the training time after each iteration is reasonable. The system automatically stores the model at every 1000 iterations. The training process ended after 300,000 iterations. The training time for this model was almost five days. We labeled this model the M1 model.

- Next, we trained the Flowtron model with K = 1. We used the M1 model as the starting point for the learning process to help the Attention layer of the model learn better and the learning process converge more quickly. The training process ended at 200,000 iterations. The training time for this model was almost two days. We labeled this model the M2 model.
- We then trained the Flowtron model with K = 2 with the M2 model we used as the pre-trained model. The training process ended after 39,000 iterations, with a training time of 450 min. We labeled this model the M3 model.
- We recognized that the speech samples we synthesized from the M3 model are natural and intelligible but lack acceptable breaks. The effect of punctuation appeared in some samples but was not obvious in the other samples, so we processed transcripts of sentences with "," for breaks and "." at the end of all sentences because the AnSpeech data set is not natural data, so punctuation such as "?" "!" and ";" will pose no difference in the pronunciation of the sound. Therefore, we removed all these punctuation marks. We limited sentence length from five to 25 words. We obtained the best results after 7,000 iterations, corresponding to 28 min of training. The resulting speech sentences have acceptable breaks after the above processing steps. This is the final model we used for the next step of synthesizing emotional-speech sentences.

4.3 Building a Module to Synthesize Vietnamese with Emotions

We describe the method to build an emotional-speech synthesis module in Section 3.4. In this study, we implemented speech synthesis with three emotions: neutral, happy, sad. First, we built a database of emotional voices by recording artists' voices; that is, artists expressed emotions when given a transcript of the sentence. Here, the recording process was completed with two famous Vietnamese artists, Ngo Ngoc Trung (as Speaker 1) and Nguyen Dunc Tam (as Speaker 2). They recorded 30 sentences. Each artist performed each sentence three times with three emotions. We completed the recording process many times and saved the recording instance in which the artist expressed the best emotion in the data set. Therefore, the data set contained 180 sentences, 90 for each artist. When we listened to and assessed these two artists' emotional expression, we found that Speaker 2 better expressed happiness and Speaker 1 better expressed sadness. Therefore, we used 30 sentences from Speaker 2 to synthesize happy sentences and 30 sentences from Speaker 1 to synthesize sad sentences. To synthesize emotionally neutral sentences, we passed the text directly into the Flowtron model, from which the model generated a Mel Spectrogram to be passed through the WaveGlow model to produce the final audio.

4.4 Evaluation of the Quality of the Vietnamese Synthesizer with Emotions

To evaluate the emotional-speech synthesizer, we randomly selected five sentences from the set of 30 sentences the artists recorded and another 10 sentences we did not record. The synthesizer generated each sentence with three emotions: neutral, happy, and sad. Therefore, the total number of sentences to be evaluated was 45. We used the mean opinion score (MOS) scale, which is the arithmetic mean of individual ratings, to assess the quality of the synthesized sentences. The MOS is a commonly used scale to assess the quality of synthesized speech. The assessment involved 60 people, 30 men and 30 women, ranging in age from 22 to 25. They all received training to use the assessment tools proficiently and understand the assessment methods. After listening to each sentence, participants evaluated the sound quality and emotional expression level. Listeners rated each sentence's emotional expression level on a scale from 1 to 5, corresponding with very poor, poor, pass, good, and very good. We used a similar set

of scales to assess the synthetic voices' naturalness and intelligibility. During the evaluation process, we displayed the sentences randomly to enhance the assessment's objectivity. Tabs. 1 and 2 present the results of these assessments. The values in the tables include the mean and standard deviation of assessment scores the listeners provided.

Table 1: MOS for naturalness and intelligibility

Evaluator gender	Emotion	Synthesized speech		Artist impression
		10 sentences not from the training set	5 sentences from the training set	5 sentences from the training set
Male	Neutral	4.46 ± 0.18	4.43 ± 0.30	4.16 ± 0.13
	Sad	4.29 ± 0.16	4.21 ± 0.31	4.34 ± 0.18
	Happy	4.17 ± 0.26	4.25 ± 0.39	4.34 ± 0.12
Female	Neutral	4.54 ± 0.26	4.45 ± 0.31	4.39 ± 0.13
	Sad	4.47 ± 0.21	4.37 ± 0.27	4.61 ± 0.21
	Happy	4.39 ± 0.29	4.37 ± 0.39	4.48 ± 0.15
Both	Neutral	4.50 ± 0.20	4.44 ± 0.29	4.28 ± 0.12
	Sad	4.38 ± 0.17	4.29 ± 0.27	4.47 ± 0.13
	Happy	4.28 ± 0.26	4.31 ± 0.39	4.41 ± 0.12

Table 2: MOS for the level of emotion expressed in speech

Evaluator gender	Emotion	Synthesized speech		Artist impression
		10 sentences not from the training set	5 sentences from the training set	5 sentences from the training set
Male	Neutral	4.19 ± 0.18	4.07 ± 0.34	4.16 ± 0.13
	Sad	3.68 ± 0.25	3.28 ± 0.30	4.34 ± 0.18
	Happy	3.75 ± 0.19	3.89 ± 0.12	4.34 ± 0.12
Female	Neutral	4.29 ± 0.15	4.17 ± 0.22	4.39 ± 0.13
	Sad	3.85 ± 0.27	3.62 ± 0.26	4.61 ± 0.21
	Happy	4.03 ± 0.21	4.13 ± 0.15	4.48 ± 0.15
Both	Neutral	4.24 ± 0.14	4.12 ± 0.28	4.28 ± 0.12
	Sad	3.76 ± 0.25	3.45 ± 0.28	4.47 ± 0.13
	Happy	3.89 ± 0.18	4.01 ± 0.11	4.41 ± 0.12

The results in Tab. 1 show that the synthesized voice is very natural and easy to understand. Most synthetic voice assessment results were higher than 4 points. The results also show that there is not much difference in the rating results based on the listeners' gender, but female assessors tended to give higher scores than male listeners. In addition, the difference in assessment results between the assessors is not large, showing that the evaluation results are reliable. The results in Tab. 2 show that the average MOS is

always between "pass" and "excellent." For sadness, the MOS is 3.61, and for happiness, it is 3.95. The results from Tabs. 1 and 2 show that there is not much difference between the evaluation results for the sentences included in the emotional training set and those outside the emotional training set, proving the proposed method's efficient scalability. However, in terms of emotional expression, the expression level of sadness is not as high as that of happiness. In addition, a gap in scores remains to achieve emotional normality.

To evaluate the emotional transition in the Z-space domain, we used the t-SNE method [30], which represents the Z-values of the training speech data and sound synthesis in a 2-dimensional space (Fig. 6). T-distributed stochastic neighbor embedding (t-SNE) is a method of directing multidimensional data that converts the similarity between data points into a probability. Using t-SNE for analysis, we aim to demonstrate the division of sentences in various emotions. The results in Fig. 6 show that emotions have good concentration in a certain area, so we can distinguish each expression area in each emotional area. Hence, we can also see the benefits of our proposed method. Furthermore, information about the speaker is also guaranteed, which is shown by the fact that each person's sentences are still concentrated in one area.

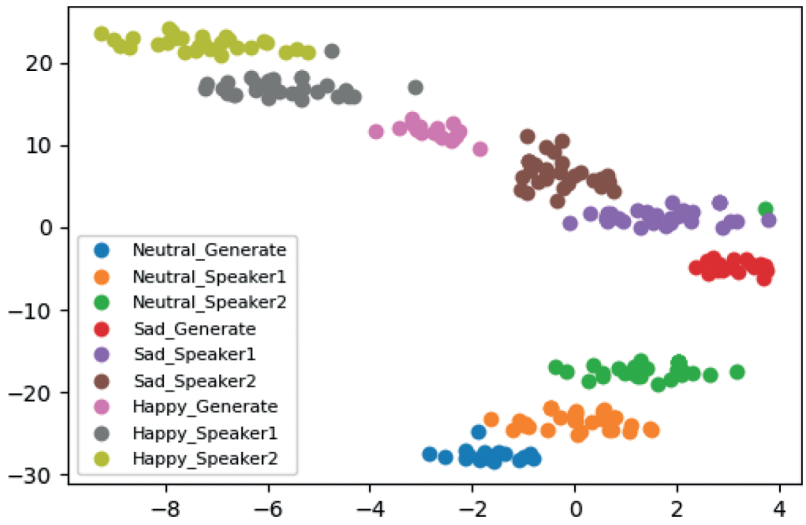

Figure 6: The t-SNE chart shows the distribution in the Z-space of all synthesized sentences, the sounds of two artists according to three different emotions

4.5 Discussion

First, we proposed a method to build a data set for the problem of synthesizing emotional Vietnamese speech. This data set ensures lexical diversity and sound quality before we include it in the training model. This database consists of two main components: the AnSpeech data set (narrative data set) and the artist-recorded emotion data set. The emotional-speech data sets that have been developed include Emo-DB [29] and BKEmo [28]. The common feature of these data sets is the complexity of construction. Therefore, we built data the set with specific criteria to ensure high automation and rapid construction. Regarding AnSpeech, we carefully prepared the recording scenarios in terms of content, which included essays and poems in the lower and upper secondary education systems; famous novels; prose articles; and blog posts sharing experiences about food, travel, music, etc. We used a free speech synthesis tool, which allowed us to create data of a single voice with a large capacity. The data we obtained consists of 9796 files, taking up 2.4 GB. The data preparation process can be completed at a relatively low cost and is not very time consuming. All content preparation, audio creation, downloading, storage, and filtering were done by five people in seven working days. It can be said that the construction of AnSpeech has

significance in terms of scalability, providing flexibility in building a dictionary. In addition, the emotion conversion system uses only 30 quality audio sentences for each emotion but is still rated "good" for sound quality and the ability to express emotions, proving the proposed method's effectiveness.

Next, we described a neutral Vietnamese speech synthesis architecture with good results. Currently, there are many successful speech synthesis systems in the world. Many complex architectures are applied, such as Deep Voice 3 [9], FastSpeech [10], and FastSpeech 2 [31]. These systems have achieved high-quality synthetic speech. However, with Vietnamese, the model with the best published results is the one based on the Tacotron architecture [32]. We proposed the new method to use Tacotron 2 as the first model and used it as a pre-trained model to train the model according to the Flowtron architecture. We made some changes from the study by Rafael Valle et al. [19], beginning to train the Flowtron model with $K = 1$ before gradually increasing the value of K (K is the number of flows of the model). We performed synthetic audio-quality control at each step during model training. This approach is slow but ensures good quality for synthesized sound, which is improved in each step of model training. At the same time, the use of the new vocoder, WaveGlow [18], which is a combination of Glow and WaveNet [17], produces high-quality sound with less computational cost thanks to the reduction in network architecture and model parameters. This proposed method helps optimize costs and reduce computational complexity and storage capacity.

Instead of building a model for each emotion (a very costly solution, time-and resource-consuming), we proposed a new method to transfer emotions for Vietnamese. The idea of the transformation is simple but effective. In fact, with the success of architectures applied to neutral voice, research groups can completely use the same architecture with data sets for each emotion. However, any such emotional-speech data set requires elaborate construction, which costs money and time. In addition, to improve the practical application, sentences need to carry many emotions or be able to switch quickly between emotions in the conversation. Therefore, the option of synthesizing voice by emotion with each set of emotional-speech data is not feasible. Our proposed method performs emotional transformation based on the hidden Z-space domain according to normal distributions. The extracted speech audio data will have many features and descriptions in the multidimensional spatial domain. However, when one performs data dimensionality reduction, there is sure to be a clear distribution in each emotion (Fig. 5), shifting the center of one distribution to another result in a wide range of stylistic and emotional variations. This transformation is many times faster than retraining each emotional model. In each emotion, we only need to use 30 audio sentences in the small emotional-speech data set. The number of training sentences is small, and the collection time is short, so the emotional transition takes place quickly, but the results obtained are still appreciated.

Finally, we assessed the quality of the synthesized speech sounds using the MOS scale, including a gender-balanced sample of participants. On the MOS scale, the quality and clarity of the synthesized sound were consistently rated as "good"; the levels of emotional expression were "passed" and "good." The audio sentences with neutral emotions were rated as the best, followed by happiness and finally the sentences that expressed sadness. We also performed a comparative analysis of the results between the synthesized audio sentences inside and outside the training set. The evaluation scores of these two sets of voices are not significantly different, showing that the model is not overfitting with the emotional training data. In addition, we synthesized the sentences in the hidden Z-space domain and compared them to the sentences the artists recorded in this domain. The results show that different emotions will appear in different clusters, and we can clearly distinguish each area representing each emotion, demonstrating our proposed method's efficiency.

We also compared the results of the proposed method with those from current state-of-the-art studies (Tab. 3). Tab. 3 shows that our method has better results than those of Zhou et al. [13] and Wu et al. [11]

even though they used more emotional-voice sentences. The results in Tab. 3 also show that when there is a large data set of emotional-speech sentences, the method Kwon et al. [12] used gave the best results.

Table 3: Comparison of our results with the state-of-the-art accuracy MOS

Methods	Number of utterances for each emotion	Number of speakers	Happy	Sad
Zhou et al. [13]	350	20	2.94	3.24
Wu et al. [11]	3057	1	2.65	2.61
Kwon et al. [12]	4135 (happy); 3986 (sad)	1	4.40	4.48
Our method	60	2	3.95	3.61

5 Conclusions

In summary, we propose a new method for emotional-speech synthesis in this study. First, we presented the method of building a Vietnamese corpus from scripts, poems, and songs, so the vocabulary is extremely diverse, the sound is not mixed with noise, and each word is pronounced. This method saves preparation time and construction costs but still ensures diversity in context, sentence length, and word count.

Next, we proposed a new method for emotional-speech synthesis. The neutral-speech generation model is a combination of the Tacotron 2 and Flowtron models. With a total training time of nearly two days, the voice produced still has good quality in terms of naturalness and intelligibility. Next, we proposed a new method to synthesize emotional-speech sentences with the style-transfer technique in the Flowtron model. With only 30 sad sentences and 30 happy sentences to make style changes and nearly 10000 sentences to build a common speech synthesis system, the average score on the MOS scale from 60 participants fell between "pass" and "good" in terms of emotional expression and between "good" and "very good" in terms of naturalness and intelligibility of sentences. However, the evaluation results show that the MOS of the expression of sadness is not high. Also, a gap remains to achieve neutral emotions. Moreover, the new study is limited to three emotions, neutral, happy, and sad, but humans usually express more emotions. Maybe the problem is that the proportion of emotional sounds (0.3%) is too small compared to the total number of narrative sentences. According to Wu et al. [11], 5% of training data were labeled with emotions as the ideal number.

Therefore, to improve the quality of emotional-speech synthesizers, we will analyze in more detail the ability to express emotions in the Z hidden space domain. In addition, we will expand the data set to include about 500 sentences for sadness and happiness (5% of the training set). These sounds are expected to be recorded by professional artists who can express emotions well. The next research direction is the ability to combine emotions in sentences because we will encounter many cases in which a sentence carries many emotions. Another direction of research will be to learn more about the application of reinforcement learning [33] in emotional-speech synthesizers.

Funding Statement: This research is funded by the Hanoi University of Science and Technology (HUST) under grant number T2018-PC-210.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proc. ITRW*, Tokyo, Japan, pp. 1–6, 2000.

[2] T. S. Phan, T. C. Duong, A. T. Dinh, T. T. Vu and C. M. Luong, "Improvement of naturalness for a hmm-based Vietnamese speech synthesis using the prosodic information," in *Proc. RIVF*, Hanoi, Vietnam, pp. 276–281, 2013.

[3] T. T. Vu, M. C. Luong and S. Nakamura, "An hmm-based Vietnamese speech synthesis system," in *Proc. COCOSDA*, Urumqi, China, pp. 116–121, 2009.

[4] J. Yamagishi, K. Onishi, T. Masuko and T. Kobayashi, "Modeling of various speaking styles and emotions for hmm-based speech synthesis," in *Proc. EUROSPEECH*, Geneva, Switzerland, pp. 1–6, 2003.

[5] S. Kayte, M. Mundada and J. Gujrathi, "Hidden Markov model-based speech synthesis: A review," *International Journal of Computer Applications*, vol. 130, no. 3, pp. 35–39, 2015.

[6] V. L. Trinh, Q. H. Nguyen and T. L. T. Dao, "Emotion recognition with capsule neural network," *Computer Systems Science & Engineering*, vol. 41, no. 3, pp. 1083–1098, 2021.

[7] M. Mustaqeem and S. Kwon, "Clstm: Deep feature-based speech emotion recognition using the hierarchical convlstm Network," *Mathematics*, vol. 8, no. 12, pp. 1–19, 2020.

[8] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss *et al.,* "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Stockholm, Sweden, pp. 4006–4010, 2017.

[9] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan *et al.,* "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. ICLR*, Vancouver, Canada, pp. 1–6, 2018.

[10] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao *et al.,* "Fastspeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, Vancouver, Canada, pp. 3165–3174, 2019.

[11] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu *et al.,* "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *Proc. APSIPA ASC*, Lanzhou, China, pp. 623–627, 2019.

[12] O. Kwon, I. Jang, C. Ahn and H. G. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1383–1387, 2019.

[13] K. Zhou, B. Sisman, R. Liu and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *Proc. ICASSP*, Toronto, pp. 920–924, 2021.

[14] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh and E. Mower *et al.,* "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[15] S. O. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky *et al.,* "Deep voice: Real-time neural text-to-speech," in *Proc. PMLR*, Sydney, Australia, pp. 195–204, 2017.

[16] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng *et al.,* "Deep voice 2: Multi-speaker neural text-to-speech," in *Proc. NIPS*, California USA, pp. 2966–2974, 2017.

[17] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly *et al.,* "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, Calgary, Canada, pp. 4779–4783, 2018.

[18] R. Prenger, R. Valle and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, Brighton, UK, pp. 3617–3621, 2019.

[19] R. Valle, K. Shih, R. Prenger and B. Catanzaro, "Flowtron: An autoregressive flow-based generative network for text-to-speech synthesis," in *Proc. ICLR*, Vienna, Austria, pp. 1–17, 2021.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Lisbon, Portugal, pp. 1517–1520, 2005.

[21] Y. Wang, D. Stanton, Y. Zhang, R. S. Ryan, E. Battenberg *et al.,* "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, Stockholm, Sweden, pp. 5180–5189, 2018.

[22] Y. J. Zhang, S. Pan, L. He and Z. H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *Proc. ICASSP*, Brighton, UK, pp. 6945–6949, 2019.

[23] O. Kwon, I. Jang, C. Ahn and H. G. Kang, "Emotional speech synthesis based on style embedded tacotron2 framework," in *Proc. ITC-CSCC*, JeJu, Korea, pp. 1–4, 2019.

[24] W. Song, G. Xu, Z. Zhang, C. Zhang, X. He *et al.,* "Efficient waveglow: An improved waveglow vocoder with enhanced speed," in *Proc. INTERSPEECH*, Shanghai, China, pp. 225–229, 2020.

[25] T. T. T. Nguyen, C. D. Alessandro, A. Rilliard and D. D. Tran, "Hmm-based tts for Hanoi Vietnamese: Issues in design and evaluation," in *Proc. INTERSPEECH*, Lyon, France, pp. 2311–2315, 2013.

[26] T. V. Nguyen, B. Q. Nguyen, K. H. Phan and H. V. Do, "Development of Vietnamese speech synthesis system using deep neural networks," *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, pp. 349–363, 2018.

[27] M. T. Nguyen and X. N. Cao, "Vietnamese speech synthesis with end-to-end model and text normalization," in *Proc. NICS*, Ho Chi Minh City, Vietnam, pp. 179–183, 2020.

[28] T. L. T. Dao, V. L. Trinh, H. Q. Nguyen and X. T. Le, "Speech emotions and statistical analysis for Vietnamese emotions," *Journal of Vietnam Ministry of Information and Communication*, vol. 35, no. 5, pp. 86–98, 2016.

[29] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss, "A database of German emotional speech," in *Proc. Eurospeech*, Lisbon, Portugal, pp. 1517–1520, 2005.

[30] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[31] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao *et al.,* "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, Austria, pp. 1–4, 2021.

[32] D. C. Tran, "The first Vietnamese fosd-tacotron-2-based text-to-speech model dataset," in *Proc. VLSP*, Vietnam, pp. 1–5, 2020.

[33] R. Liu, B. Sisman and H. Li, "Reinforcement learning for emotional text-to-speech synthesis with improved emotion discriminability," in *Proc. INTERSPEECH*, Brno, Czechia, pp. 4648–4652, 2021.