applied sciences

MDPI

Article

Design of a Multi-Condition Emotional Speech Synthesizer

Sung-Woo Byun [1] and Seok-Pil Lee [2,*]

[1] Department of Computer Science, Graduate School, SangMyung University, Seoul 03016, Korea; 123234566@naver.com
[2] Department of Electronic Engineering, SangMyung University, Seoul 03016, Korea
* Correspondence: esprit@smu.ac.kr

Abstract: Recently, researchers have developed text-to-speech models based on deep learning, which have produced results superior to those of previous approaches. However, because those systems only mimic the generic speaking style of reference audio, it is difficult to assign user-defined emotional types to synthesized speech. This paper proposes an emotional speech synthesizer constructed by embedding not only speaking styles but also emotional styles. We extend speaker embedding to multi-condition embedding by adding emotional embedding in Tacotron, so that the synthesizer can generate emotional speech. An evaluation of the results showed the superiority of the proposed model to a previous model, in terms of emotional expressiveness.

Keywords: emotional speech synthesizer; Korean emotional speech; Tacotron

Citation: Byun, S.-W.; Lee, S.-P. Design of a Multi-Condition Emotional Speech Synthesizer. *Appl. Sci.* **2021**, *11*, 1144. https://doi.org/10.3390/app11031144

Academic Editor: José Luis Rojo-Álvarez
Received: 29 December 2020
Accepted: 23 January 2021
Published: 26 January 2021

1. Introduction

A text-to-speech system (TTS) synthesizes and output human-like speech signals so that linguistic, prosodic, and semantic information can be naturally delivered. Recently, due to advances in deep learning technologies, researchers have developed text-to-speech models based on deep learning, which have produced superior results to those of previous approaches. This work led to investigations into the issue of understanding the contextual meaning of synthesized speech [1]. Among the TTS models [1–6], Wang et al. [1] proposed an end-to-end TTS model, referred as Tacotron, that can be trained from scratch on <text, audio> pairs. Tacotron is a sequence-to-sequence (seq-to-seq) model with an attention mechanism. Tacotron has many advantages compared with other TTS models, such as WaveNet [4] and Deep Voice [5,6], while the speech signal generated by Tacotron achieved competitive mean opinion scores (MOS). As TTS models have been improved rapidly, the need for eclectic application programs, such as audio book narration, conversational assistants, or news readers, are increasing. Although the model has shown a potential to strongly integrate long vocal expressions, the research in this field still remains in the early stage [7].

In order for computers to deliver speech like human beings, TTS system has to learn how to model prosody. Prosody refers to a combination of various phenomena in speech, such as linguistic information, intonation, accent, and style. Research pertinent to this focuses on style modeling, and the goal is to provide a model with features to choose the proper speaking style relevant to a given context. Style contains information, such as intention and emotion, and affects speakers' choice of intonation and flow. Proper rendering of style influences overall recognition, which is important to applications such as audio books and news readers.

Deep learning-based TTS systems, such as global style tokens (GST) Tacotron, produce expressive speech by adding embedding vectors that implicitly provide prosody-related latent features [8]. However, because those systems only mimic the generic speaking style of reference audio, it is difficult to assign user-defined emotion types to synthesized speech [1]. Lee et al. [3] input emotional labels to the decoder of Tacotron by concatenating

the labels with the output of the pre-net. Training using this approach can make the model reflect the emotional state of a speaker. Even though the method showed feasibility when attaching emotions to synthesized speech, when the amount of data is small, there is still a problem with limited emotional representation in the speech.

The main objective of this work is to design an emotional speech synthesizer that can generate emotional speech clearly. To achieve this, we construct a Korean emotional speech database for emotional speech synthesis in a professional studio. The database consists of recordings using Korean scripts from dramas and movies, with four professional actors. Additionally, we propose a multi-condition embedding technique that controls not only speaking styles but also emotional styles by extending speaker embedding to multi-condition embedding. This approach facilitates training the attention of the model, by training using an emotional speech dataset with a large amount of speech data. The use of multiple speakers means a lack of emotional state in one speaker's speech data can be complemented with data from other speakers. Evaluation of the results showed the superiority of the proposed model to a previous model, in terms of emotional expressiveness.

This paper is organized as follows. Section 2 describes the related works. Section 3 introduces the emotional speech database used in this research. Section 4 discusses multi-condition emotional speech synthesizers, which are a key part of the proposed method. Section 5 presents the experiment description and results, after which Section 6 provides a discussion and a conclusion.

2. Related Works

2.1. Tacotron

There have been several TTS models [1–6]. Wang et al. [1] proposed an end-to-end TTS model, referred to as Tacotron. The Tacotron model has been improved, to make it more similar to humans in terms of sound quality than conventional speech synthesis studies. Tacotron consists of four modules. First, the Encoder module converts the input text into numbers which the computer can understand. Second, the Decoder module converts the numbers, converted from the input text, back into a form of spectrogram. The Decoders consist of recurrent neural networks (RNN), and creates the next spectrogram using the previous spectrogram through a decoding process. Third, the Attention module makes the appropriate connections between the Encoder and the Decoder. In the Tacotron, the Attention module plays the most important role, and the Decoder decides which part of the sentence to focus on to generate speech, using Attention values. With this approach, a model learns from various data, and emotional speech can be produced by imitating the accent or speech rate of a speaker. Lastly, the task of the Vocoder is to reconstruct the generated spectrogram into waveforms. The authors used the Griffin–Lim algorithm to synthesize waveforms from the predicted spectrogram [8]. Figure 1 describes the model, which includes an Encoder, an Attention-based Decoder, and a Vocoder [1].

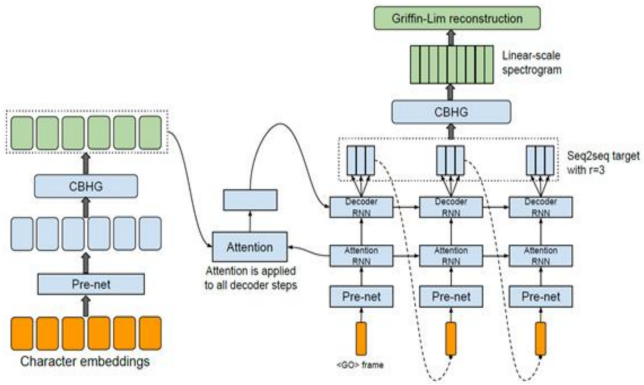

Figure 1. Structure of the Tacotron [1].

2.2. Emotional End-to-End Neural Speech Synthesizer

Lee et al. proposed a modified Tacotron as an emotional speech synthesizer [3]. The system takes a character sequence and a desired emotion as input, and generates the corresponding wave signal. The model was implemented by injecting a learned emotional embedding, e, and demonstrated the feasibility of adding emotions to synthesized speech. The authors implemented the emotional Tacotron by injecting a learned emotional embedding e as follows:

$$h_t^{att} = AttentionRNN\left(x_t, h_{t-1}^{att}, e\right), \quad h_t^{dec} = DecoderRNN\left(c_t, h_{t-1}^{dec}, e\right), \tag{1}$$

Following Equation (1), the style vector, e, can be injected into the *AttentionRNN* and *DecoderRNN* of a Tacotron. The model could successfully generate speech for given emotional labels.

2.3. GST Tacotron

The GST Tacotron flexibly changes the style of synthesized speech to have characteristics similar to reference audio by adding a global style token module to an end-to-end TTS [8]. The Tacotron network estimates the mel-spectrogram from input text, utilizing a sequence-to-sequence model with an Attention mechanism [9–14]. The style vector, referred to as style embedding, which determines the style of the reference audio, is generated by a token layer. The reference encoder of the token layer generates a fixed length embedding vector, referred to as reference embedding, from the input of the mel-spectrogram of the reference audio. Then, multi-head attention calculates the similarity between the style embedding and the reference embedding. The weighted sum of the style embedding is input into the attention module of the Tacotron, with the vector generated by the encoder of the Tacotron. Therefore, the GST model, illustrated in Figure 2, consists of a reference encoder, style attention, style embedding, and sequence-to-sequence (Tacotron) model.

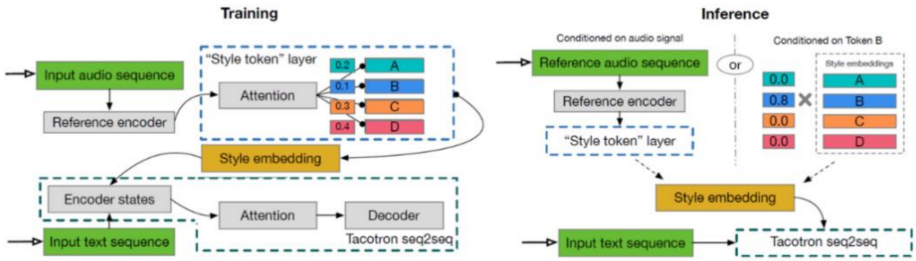

Figure 2. Structure of the global style tokens Tacotron [8].

3. Korean Emotional Speech Database

The study of emotion analysis has rapidly developed over the last decade with broad interest from researchers in neuroscience, computer science, psychiatry, psychology, audiology, and computer science. The key of these studies is to secure the availability of validated and reliable expressions of emotion. Most emotion datasets include either facial expressions or speech recordings. Among the datasets, few contain audio-visual recordings of speakers in Korean. This study constructs a Korean emotional speech database and re-ports validity and reliability of the data based on ratings from participants. The database was recorded with Korean utterances from professional actors. All recorded data were recorded in a professional studio, considering the sound quality of the data by eliminating any background noise.

The database was recorded using Korean scripts from dramas and movies, with four professional actors, two females and two males. The scripts were collected from actual scripts used in dramas and movies. The scripts consisted of 30-s-long conversations between a male and a female per emotional scene. We defined four emotions: Anger, Happiness, Neutrality, and Sadness. The scripts consisted of 120 scenes per emotion. As

suggested in [15], actors engaged in their role during recording may provide a more natural expression of the emotions, and emotions are incorporated into the scripts themselves, allowing an actor to record with strong emotions. This database relied on professional actors and senior students from the Drama Department. Actors had Korean as their first language, spoke with a neutral Seoul accent, and did not possess any distinctive features. The actors recorded their voices from around 30 cm away from the microphone (Figure 3). To allow the actors to understand which emotion was requested, a description and sample of each emotion were given. Actors were asked to maintain strong intensity during recording, and feedback was given if a research assistant considered a production to be ambiguous. Trials were organized by emotion, with low-intensity emotions followed by intense emotions. This ordering allowed actors to enter and remain within the desired state for all productions of that emotional category. Actors were then given time to prepare their emotional state using their desired induction technique. We focused on increasing a number of emotional speech data rather than the emotional quality, so we did not carry out the evaluation test, unlike general speech databases.

Figure 3. Two actors during recording, using conversation between a male and a female.

After recording, the data were manually segmented by dialog turn (speaker turn) from continuous segments in which the actors were actively speaking. The scripts were segmented into sentences in advance, and used as references with which to split the data. The final audio recordings were divided into approximately 3–10-s-long files. Each file was around two to three hours long per actor, producing a total of around 4000–5000 files. The total size of the database was 18,324 files. All speech emotion data was recorded at 48 kHz and downsampled with a 24 kHz sampling rate in PCM signed 16-bit format. We carried out post-processing, which normalized the amplitude of the signal and removed the front and back margin. The results of post-processing are shown as Figure 4.

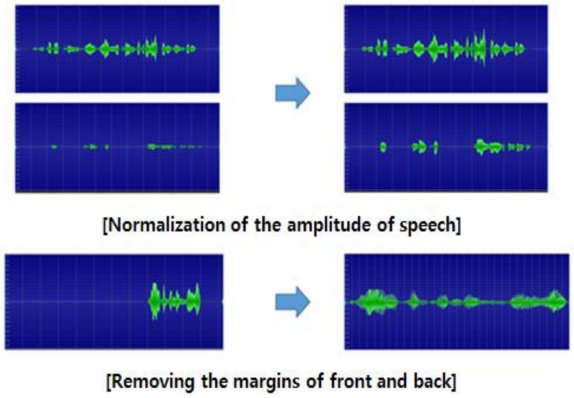

Figure 4. Results of post-processing.

4. Multi-Condition Emotional Speech Synthesizer

4.1. Model

Tacotron consists of four parts: an Encoder, which extracts numerical features representing the input text; a Decoder, which generates a mel-spectrogram based on the output

of the Encoder; an Attention module, which creates a relationship with the Encoder output value when creating the spectrogram in the Decoder; and a Vocoder, which generates speech waves from the generated spectrogram. In this paper, we propose multi-conditional emotional synthesizer by providing additional information to Tacotron when synthesizing speech. The objective of the proposed method was to train a TTS system using speech data from multiple speakers, so that the emotional states of different speakers complement each other.

Unlike previous studies, we extended a single embedding vector, such as a speaker embedding vector or an emotional embedding vector, so that the model can reference multiple conditions. Each of the multi-condition embedding vectors were initialized randomly with a uniform distribution over $[-0.1, 0.1]$ and the network was trained using backpropagation:

$$z = concat(s, e), \tag{2}$$

$$Multicondition = \tanh\left(w_{f,n} \times z\right), \tag{3}$$

where s is a speaker embedding vector, and e is an emotional embedding vector. We extended the model by adding a multi-condition embedding vector (Equation (3)) to synthesize speech according to multiple conditions. Unlike previous studies, the information is stored in low-level vectors according to speakers and emotions. Each embedding vector is independent, but the input data can complement each other, although the model learns according to the embedding vector. For example, if there is an untrained pronunciation 'a' in emotion B of speaker A, the model can train the pronunciation of 'a' by training the model from the pronunciation of 'a' in emotion C of speaker B. Previous studies have shown that simply adding an embedding vector to the input layer is not effective [6]. Therefore, a multi-condition embedding vector was used at several locations, so that each speaker and emotion affected the model. The whole structure and the example of inputting the multi-condition embedding vector are shown in Figure 5.

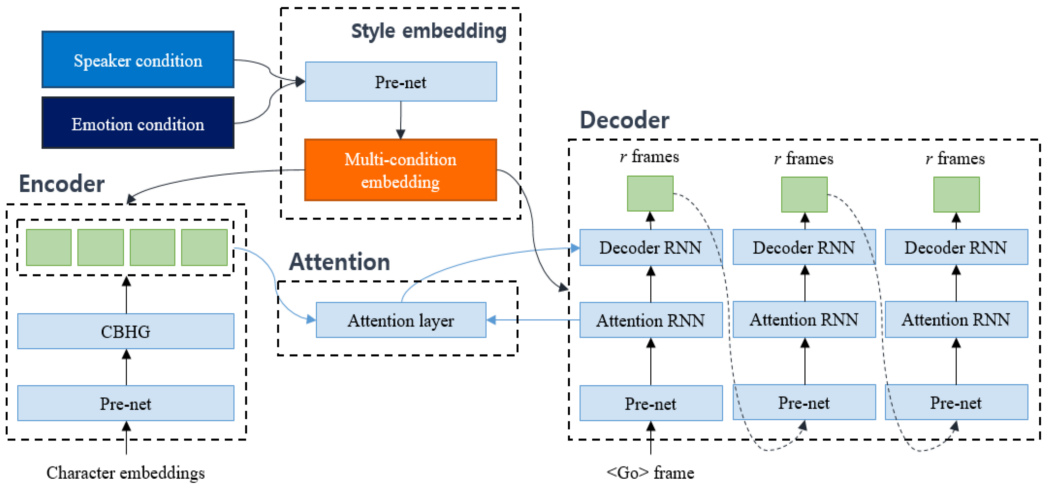

Figure 5. Structure and an example of inputting the multi-condition embedding vector.

4.2. Multi-Conditional Encoder

In the original Tacotron, a CBHG (Convolution Bank + High-way + bi-GRU) module generates encoding vectors from a text input, illustrated in Figure 6 [1]. The embedding vectors were combined with a CBHG module, a long short-term memory encoder, attention RNN, and a decoder RNN equivalent to Deep Voice 2 [6].

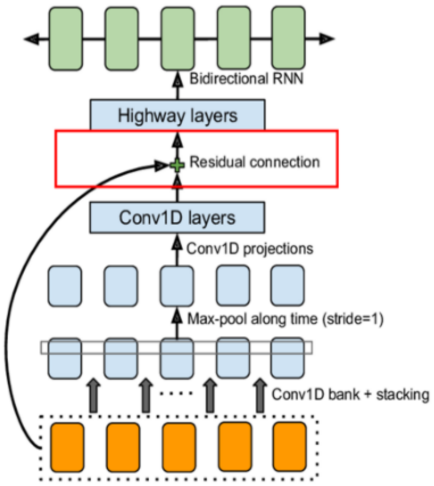

Figure 6. The location of the Residual connection in the CBHG module [1].

We injected the multi-condition embedding vector by adding the embedding vector to the Residual connection of the CBHG module. The location of the Residual connection in the CBHG module is shown in Figure 6.

In the CBHG module, the output of the convolution filter bank is projected by another convolution layer to match the size of the input vector, x. The output of the projection is connected to the input vector x, so that the gradient of the input vector x does not vanish. The Encoder generates the final output by adding multi-condition embedding vector to the Residual connection, taking into account the output of convolution filter bank, the input vector x, and the condition of multi-condition embedding vector. This procedure is shown in Equation (4):

$$Residual_{out} = Conv_{out} + x + vector_{emb}, \quad (4)$$

where $Conv_{out}$ is the output of the filter bank, x is the input vector, and $vector_{emb}$ is the multi-condition embedding vector.

4.3. Attention-Based Multi-Conditional Decoder

We injected the embedding vector into the Decoder to construct the Tacotron, which synthesizes speech according to multi-conditions. In the Tacotron [1], the Decoder consists of a stack of RNN layers with a content-based tanh attention mechanism. The Decoder generates the frames of a mel-spectrogram at every time-step, utilizing a pre-net layer, an attention RNN layer, and a decoder RNN layer. The Decoder starts with a "GO" frame. The generated frame is fed to the next step of the Decoder's pre-net for every time-step. We injected the multi-condition embedding vector into the attention RNN by initializing the attention RNN with the embedding vector:

$$h_t^{att} = AttentionRNN\left(x_t, h_{t-1}^{att}, vector_{emb}\right), \quad (5)$$

To reflect the characteristics of multi-condition embedding when generating a mel-spectrogram, we injected the embedding vector into the *DecoderRNN*, which is equivalent to the *AttentionRNN*.

$$h_t^{dec} = DecoderRNN\left(x_t, h_{t-1}^{dec}, vector_{emb}\right), \quad (6)$$

5. Experiments and Results

5.1. MOS Test Result

To evaluate the performance of the proposed method, we used the conversation data of males and females, described in Section 3. The Korean emotional speech database is a database of emotional speech recorded with Korean utterances. The database was

constructed using Korean scripts from dramas and movies, so that priority was given to the naturalness of speech. The categories of emotions were Anger, Happiness, Neutrality, and Sadness, and the total size of the database was 18,324. Each file was around two to three hours' long per actor, producing a total of around 4000–5000 files. Then, the proposed model was trained to synthesize emotional speech signals. For learning, the Attention module of the Tacotron requires about four hours of speech data for each speaker. The database was not large enough for the model to learn the characteristics of each speaker and each emotion. As a supplement to this database, the KSS database [16] for the TTS model was used. The KSS database does not have the emotional labels; however, the "Neutrality" emotion of the KSS speech dataset could be used to learn the model.

To demonstrate the clarity and expressiveness of the proposed model, a five-scale MOS test was carried out with 10 subjects. Each subject listened to each sample and was asked to decide the clarity on a scale of 1 to 5, according to whether they could understand the words in the sample. They were also asked to provide an emotion score from 1 to 5. If a subject felt that the emotion was clearly evident in the speech data, they gave a score of 5. However, if a subject felt as if the speech data did not include emotion, they give a score of 1.

As shown in Table 1, the proposed model achieved the average of the clarity of 4.03, and the standard deviation was 0.42, which outperforms the previous emotional speech synthesizer, which achieved 3.59 ± 0.51. The GST Tacotron had an average of 3.54 for the clarity, and the standard deviation was 0.33. For the GST Tacotron, the extraction of style embedding from the reference audio did not accurately reflect the emotion of the reference audio, so the speech was ambiguously synthesized. As a result, since the model synthesized speech of poor clarity and that was awkward, the performance of the GST Tacotron was poor. The proposed model achieved a better emotional score, namely, 4.06, with a standard deviation of 0.46, compared to the other models, which had scores of 3.82 ± 0.46 and 3.3 ± 0.33, respectively. For the same reason, the other models had low emotional scores. In the result of all models, the synthesized speech of the "Happiness" emotion had lower clarity and emotional scores, since it sometimes sounded like "Anger" or "Neutrality." In particular, in the case of the GST Tacotron, the style embedding vector "Happiness" of the reference audio was misrepresented as "Anger," so the model generated speech that was completely different from the intention of the reference audio. The reason the proposed method has a higher clarity and emotional score than the emotional speech synthesizer is because it could learn using the database for large TTS, by expanding multi-condition embedding, which includes speaker embedding and emotional embedding. The model can, therefore, learn by complementing each speaker's emotional data.

Table 1. Results of the MOS test.

		Anger	Happiness	Neutrality	Sadness
Real data	Clarity	5 ± 0	4.8 ± 0.33	5 ± 0	5 ± 0
	Emotion score	5 ± 0	5 ± 0	4.71 ± 0.2	4.75 ± 0.3
Proposed	Clarity	4.11 ± 0.36	3.95 ± 0.42	4.11 ± 0.45	3.95 ± 0.46
	Emotion score	4.1 ± 0.43	3.8 ± 0.55	4.21 ± 0.41	4.11 ± 0.45
Emotional ETE Speech Synthesizer	Clarity	3.75 ± 0.44	2.78 ± 0.51	4.01 ± 0.59	3.81 ± 0.52
	Emotion score	3.88 ± 0.45	2.8 ± 0.53	4.5 ± 0.38	4.08 ± 0.49
GST-Tacotron	Clarity	3.57 ± 0.33	3.51 ± 0.26	3.53 ± 0.4	3.55 ± 0.33
	Emotion score	3.53 ± 0.26	2.51 ± 0.34	3.6 ± 0.32	3.55 ± 0.39

5.2. Discussion of Attention Value Sequences

Attention plays an important role in the synthesis of speech by the Tacotron. To accurately learn and synthesize the Attention module of the Tacotron requires about four hours of speech data for each speaker. This study complemented the lack of data by

extending the embedding vector to multi-conditions, to enable the model to train on the constructed data with the additional TTS database.

Figure 7 shows the Attention value sequences when the model was trained using only the constructed data without additional training data from TTS. In Figure 7, the Attention value sequences have blurred parts in the middle of the sequence, within the red circle. As a result, the model was confused about which part of the output of the encoder it should focus on, causing broken generated speech. In the case of the GST Tacotron, it was unable to learn by adding data, and it was difficult for the Attention module to learn when the data are insufficient, because it relies on reference audio.

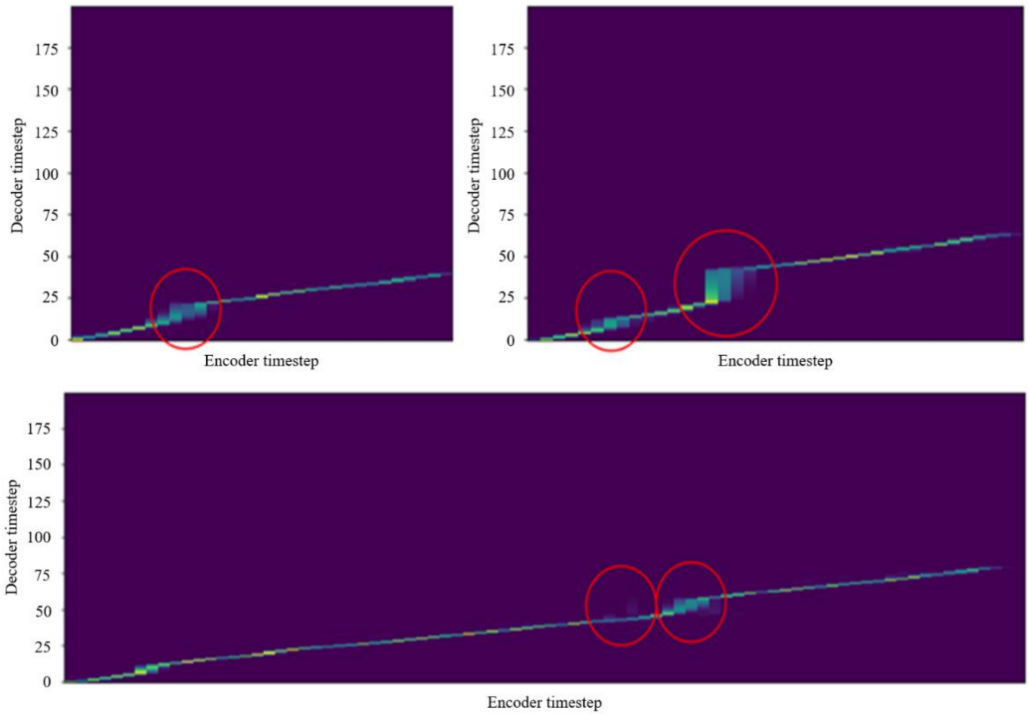

Figure 7. Attention value sequences when the model was trained using only the constructed data without additional training with the text-to-speech data.

Figure 8 shows the Attention value sequences of speech generated by the GST Tacotron. The sequences have gaps or blurred parts in the middle of the sequence, within the red circle.

Figure 9 shows the Attention value sequences of speech generated by the proposed model. We observed clearer and sharper Attention value sequences, unlike previous studies. This facilitated the learning of Attention using the database together with the database for TTS. Since the Attention value sequences were clear and sharp, clearer and cleaner emotional speech could be synthesized. Therefore, the result of the proposed model was superior to the results of previous studies with regards to the MOS test described in Section 5.1.

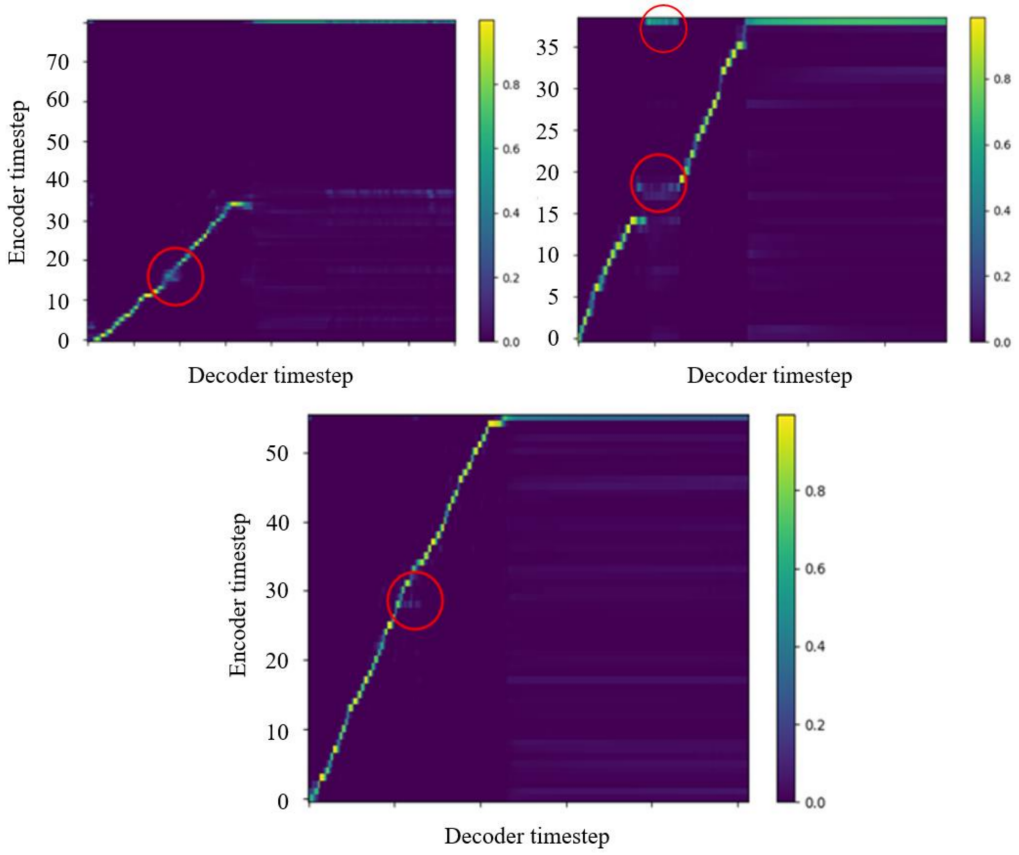

Figure 8. Attention value sequences of speech generated by the GST Tacotron.

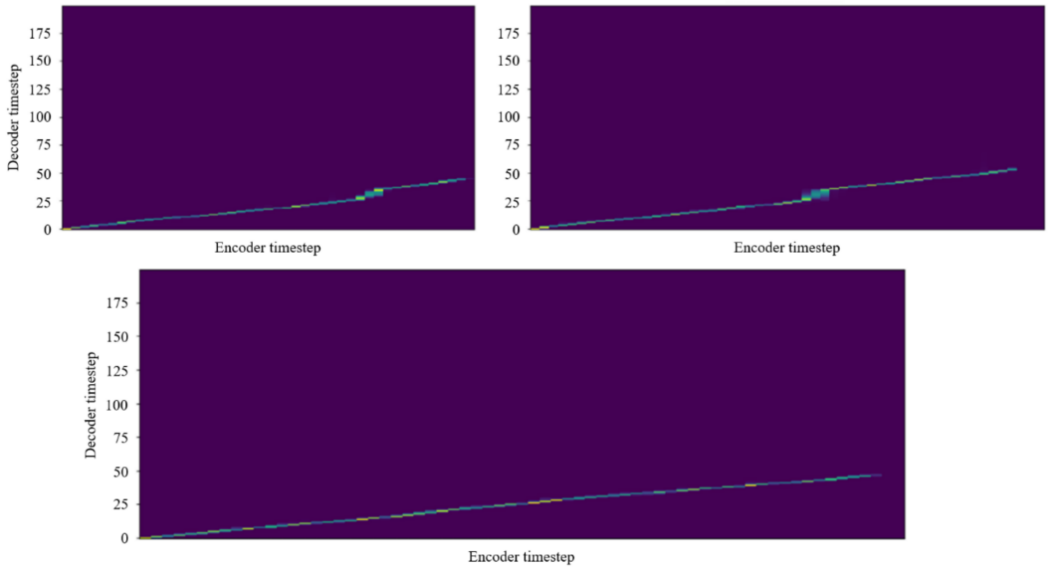

Figure 9. Attention value sequences of speech generated by the proposed model.

6. Conclusions and Future Works

This paper describes an emotional speech synthesizer constructed by extending speaker embedding to multi-condition embedding by adding emotional embedding. To do this, we constructed a Korean emotional speech database for speech synthesis, which consisted of conversations between males and females. The model was trained using the emotional style of the speakers, and a lack of an emotion in one speaker was complemented by data from other speakers. To facilitate training the attention of the model, we trained

the model using both the emotional speech dataset and a large amount speech data for TTS. A five-scale MOS test was carried out using five subjects. In terms of emotional expressiveness, the proposed model achieved the clarity of 4.03, which outperforms the previous emotional speech synthesizers achieved 3.59 and 3.54.

In order to synthesize natural-sounding emotional speech, studies of Vocoders must be performed. Previous studies reconstructed mel-spectrograms to wave signals using the Griffin–Lim algorithm, but many recent studies involving speech synthesis have replaced the Vocoder with a deep learning model. Therefore, when synthesizing speech taking into account emotions, studies into Vocoders are required, to reconstruct speech while meeting specific conditions. Studies into synthesizing speech based on the intensity of emotion are also required, rather than simply synthesizing speech according to emotional labels.

Author Contributions: Conceptualization, S.-W.B.; methodology, S.-W.B. and S.-P.L.; investigation, S.-W.B.; writing—original draft preparation, S.-W.B.; writing—review and editing, S.-P.L.; project administration, S.-P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea.

Institutional Review Board Statement: Not applicable.

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

Data Availability Statement: Data sharing is not applicable to this article.

Acknowledgments: This work was supported by a grant from the Institute for Information & communications Technology Promotion (IITP) funded by the Korea government (MSIP) (No.2017-0-00515, Development of integraphy content generation technique for N-dimensional barcode application).

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

References

1. 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**, arXiv:1703.10135. Available online: arxiv.org/abs/1703.10135 (accessed on 29 December 2020).
2. Um, S.; Oh, S.; Byun, K.; Jang, I.; Ahn, C.; Kang, H. Emotional Speech Synthesis with Rich and Granularized Control. In Proceedings of the ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7254–7258.
3. Lee, Y.; Rabiee, A.; Lee, S. Emotional End-to-End Neural Speech Synthesizer. *arXiv* **2017**, arXiv:1711.05447. Available online: arxiv.org/abs/1711.05447 (accessed on 29 December 2020).
4. Oord, A.v.d.; 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**, arXiv:1609.03499. Available online: arxiv.org/abs/1609.03499 (accessed on 29 December 2020).
5. Arik, S.O.; Chrzanowski, M.; Coates, A.; Diamos, G.; Gibiansky, A.; Kang, Y.; Li, X.; Miller, J.; Ng, A.; Raiman, J. Deep Voice: Real-Time Neural Text-to-Speech. *arXiv* **2017**, arXiv:1702.07825. Available online: arxiv.org/abs/1803.09017 (accessed on 29 December 2020).
6. Gibiansky, A.; Arik, S.; Diamos, G.; Miller, J.; Peng, K.; Ping, W.; Raiman, J.; Zhou, Y. Deep Voice 2: Multi-Speaker Neural Text-to-Speech. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 2962–2970.
7. Wang, Y.; Stanton, D.; Zhang, Y.; Skerry-Ryan, R.; Battenberg, E.; Shor, J.; Xiao, Y.; Ren, F.; Jia, Y.; Saurous, R.A. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv* **2018**, arXiv:1803.09017. Available online: arxiv.org/abs/1803.09017 (accessed on 29 December 2020).
8. Griffin, D.; Lim, J. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
9. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-Based Models for Speech Recognition. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 577–585.
10. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 3104–3112.
11. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv* **2014**, arXiv:1409.0473. Available online: arxiv.org/abs/1409.0473 (accessed on 29 December 2020).
12. Luong, M.; Pham, H.; Manning, C.D. Effective Approaches to Attention-Based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025. Available online: arxiv.org/abs/1508.04025 (accessed on 29 December 2020).

13. Chan, W.; Jaitly, N.; Le, Q.; Vinyals, O. Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 4960–4964.
14. 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; pp. 4779–4783.
15. 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] [PubMed]
16. Korean Single Speaker Speech Dataset (KSS Dataset). Available online: https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset (accessed on 29 December 2020).