EMOQ-TTS: EMOTION INTENSITY QUANTIZATION FOR FINE-GRAINED CONTROLLABLE EMOTIONAL TEXT-TO-SPEECH

Chae-Bin Im[1]*, Sang-Hoon Lee*[2]*, Seung-Bin Kim*[1]*, Seong-Whan Lee*[1,2]

[1]Department of Artificial Intelligence, Korea University, Seoul, Korea
[2]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea

ABSTRACT

Although recent advances in text-to-speech (TTS) have shown significant improvement, it is still limited to emotional speech synthesis. To produce emotional speech, most works utilize emotion information extracted from emotion labels or reference audio. However, they result in monotonous emotional expression due to the utterance-level emotion conditions. In this paper, we propose EmoQ-TTS, which synthesizes expressive emotional speech by conditioning phoneme-wise emotion information with fine-grained emotion intensity. Here, the intensity of emotion information is rendered by distance-based intensity quantization without human labeling. We can also control the emotional expression of synthesized speech by conditioning intensity labels manually. The experimental results demonstrate the superiority of EmoQ-TTS in emotional expressiveness and controllability.

Index Terms— Text-to-speech, expressive emotional speech synthesis, fine-grained control, emotion intensity modeling

1. INTRODUCTION

Recently, there has been significant improvement in end-to-end text-to-speech (TTS) systems [1, 2, 3, 4] due to the advancement of deep learning [5, 6]. Although synthesized speech from the current TTS model has already achieved outstanding performance, there still remains a limitation in synthesizing expressive speech with paralinguistic features such as pitch, tone, and tempo. In particular, emotional speech synthesis is a challenging task since emotion information is affected by various paralinguistic characteristics of speech.

For emotional speech synthesis, the common approach is to condition global emotion information extracted from reference audio [7, 8] or emotion labels [9, 10]. However, these methods have a disadvantage where the synthesized speech has monotonous expression since the whole sentence is regulated by only one global information. To generate expressive emotional speech similar to spontaneous human speech, fine-grained emotional expressions according to emotion intensity should be considered at the phoneme-level. Several studies have attempted to reflect fine-grained emotional expression by scaling [11, 12] or interpolating [13, 14] the representative emotion embedding. Nevertheless, they have a problem with unstable audio quality, and it is also difficult to find proper parameters for scaling or interpolation. In the case of [15], the model predicts phoneme-wise intensity scalar extracted from a learned ranking function [16]. However, this method tends to depend heavily on the global label, thus it is unstable to control the emotional expression based on intensity scalar.

To address the above problems, this paper proposes EmoQ-TTS, which synthesizes expressive emotional speech by conditioning phoneme-wise emotion information based on fine-grained emotion intensity. To reflect appropriate emotional expression, we utilize intensity pseudo-labels and via distance-based intensity quantization without human labeling. EmoQ-TTS synthesizes speech more expressively by predicting appropriate emotion intensity from the text only. Furthermore, we can control emotion expression easily by conditioning intensity labels manually. The experimental results show that our system successfully achieves better emotional expressiveness and controllability than conventional methods. The synthesized audio samples are available at `https://prml-lab-speech-team.github.io/demo/EmoQ-TTS/`

2. EMOQ-TTS

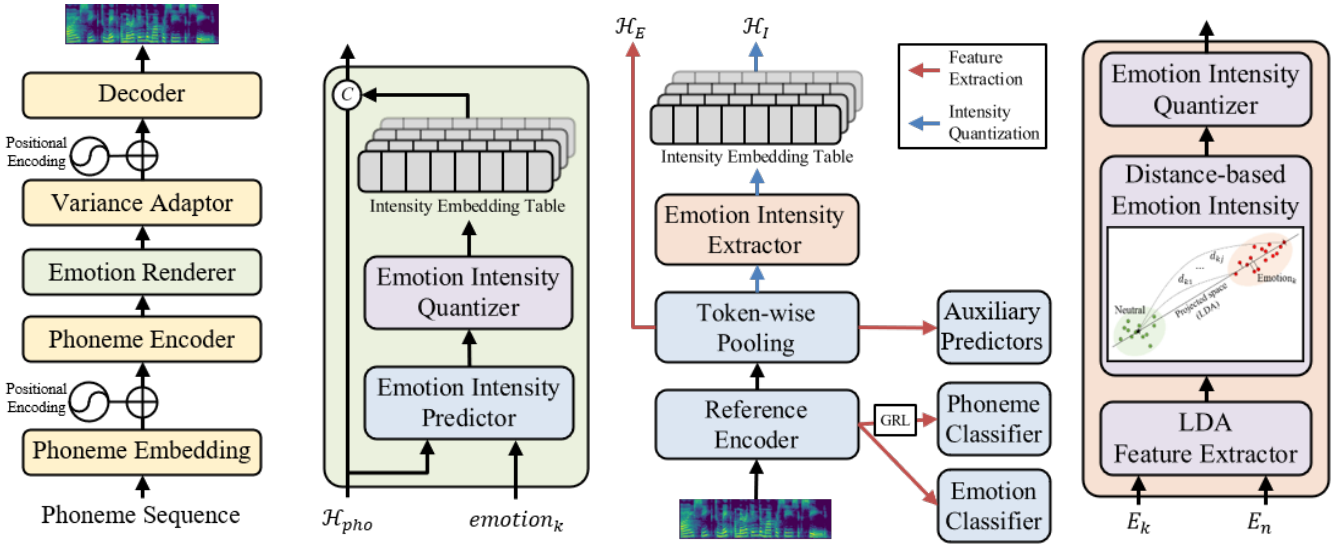
2.1. Model Architecture

The entire architecture of EmoQ-TTS is depicted in Fig.1a. EmoQ-TTS is based on FastSpeech2 [17] which consists of an encoder, a decoder, and a variance adaptor. To synthesize fine-grained emotional speech, we modify FastSpeech2 architecture as follows: First, we introduce an emotion renderer to provide phoneme-level emotion information according to fine-grained emotion intensity. This enables all variance information, including the pitch, energy, and duration to be affected by the fine-grained emotion intensity. Second, the duration predictor is moved to the end of the variance adaptor. This leads to all variance information to be processed at the phoneme-level, which has been proved better performance than the frame-level method in speech quality [18].

2.2. Emotion Renderer

The emotion renderer provides phoneme-level emotion information according to fine-grained emotion intensity. As shown in Fig.1b, the emotion renderer consists of an intensity predictor, intensity quantizer, and intensity embedding table. When the phoneme hidden sequence $\mathcal{H}_{pho}$ and the k-th emotion category $emotion_k$ are provided, the intensity predictor predicts the phoneme-wise emotion intensity scalar sequence suitable for $emotion_k$ as a value between 0 and 1. The intensity predictor is optimized by mean absolute error (MAE) [19] loss which minimizes the differences between the predicted intensity scalar sequence and ground-truth intensity scalar sequence.

For robust training, the intensity scalar is quantized into N_I-sized emotion intensity pseudo-labels at regular intervals through emotion intensity quantizer. Here, N_I denotes the total number of quantized intensity pseudo-labels. Moreover, we introduce an intensity embedding table. The quantized intensity pseudo-labels are the entry index of the embedding table for each emotion. Finally, the phoneme-wise intensity embedding sequence is concatenated to the phoneme hidden sequence. The 'C' in fig.1b represents concatenation. The ground-truth intensity scalar and the intensity embedding



(a) Overall structure (b) Emotion renderer (c) Emotion intensity modeling (d) Intensity extractor

Fig. 1: (a) The overall architecture of EmoQ-TTS. (b) The detailed structure of the emotion renderer. 'C' denotes concatenation. (c) The entire pipeline of the emotion intensity modeling. 'GRL' denotes gradient reversal layer. (d) The detailed structure of the intensity extractor.

table are designed through emotion intensity modeling. The detailed process of the intensity modeling will be described in the following section.

During the inference, the EmoQ-TTS synthesizes expressive emotional speech by conditioning quantized intensity embedding from the predicted intensity scalar. Furthermore, the EmoQ-TTS controls the emotional expression of synthesized speech by controlling intensity with custom labels manually.

3. EMOTION INTENSITY MODELING

In this section, we describe designing emotion intensity information via distance-based intensity quantization. As shown in Fig.1c, the emotion intensity modeling is conducted in two stages: i) emotion feature extraction stage and ii) emotion intensity quantization stage.

3.1. Emotion Feature Extraction

In the first stage, we train the reference encoder to extract clustered emotion embedding from mel-spectrogram. The reference encoder is comprised of three 1D convolution layers which maintain temporal information for each frame. To extract discriminative emotion embedding, we apply the emotion classifier and the phoneme classifier with gradient reversal layer (GRL) [20] as depicted in Fig.1c. These classifiers make feature vectors clustered by emotion without interruption of phoneme information. Both classifiers are optimized with a softmax layer followed by cross-entropy loss. In the case of the phoneme classifier, the gradient is reversed by multiplying a negative scalar during back-propagation through the gradient reversal layer.

Then, the token-wise pooling layer transforms frame-level sequence to phoneme-level sequence by averaging within range of each phoneme boundary. Here, we add two auxiliary predictors which predict pitch and energy respectively. These predictors allow the clustered embedding to reflect the emotion information well by predicting paralinguistic features which is directly related to emotion. The auxiliary predictors are optimized with mean squared error (MSE) [21] loss.

The components of the emotion feature extraction module are jointly trained with our TTS system. During training, the extracted phoneme-wise emotion embedding sequence $\mathcal{H}_E$ in fig.1c is directly passed to the emotion renderer, and concatenates to the phoneme hidden sequence in fig.1b instead of the intensity embedding sequence.

3.2. Distance-based Emotion Intensity Quantization

In the second stage, we generate emotion intensity pseudo-labels and intensity embedding table via intensity quantization. As shown in Fig.1d, two clusters of k-th emotion embeddings E_{kj} and neutral embeddings E_{nj} are fed to emotion intensity extractor where $j \in \{1, 2, \cdots, N_k\}$ and N_k is the total number of E_{kj}. For this work, we use the entire emotion embedding for each emotion extracted from the reference encoder.

For extracting proper intensity, we introduce the emotion distance, which represents how far the vector is relatively from the centroid of the neutral emotion. Here, we make two assumptions. At first, neutral emotion is the weakest intensity of other emotions. Secondly, the emotion intensity increases as it moves farther from neutral emotion. Since multi-dimensional space is too unstable to measure the distance, the emotion intensity extractor projects the provided emotion embedding onto a single vector. Among various vector projection methods, we choose the linear discriminant analysis (LDA) [22] method which is a class-sensitive projection method through experiments. The optimal projection vector w^* is obtained by maximizing objective function of binary-class LDA as follows:

$$\mathcal{L}_{LDA}(w) = \frac{(m_k - m_n)^2}{s_k^2 + s_n^2} = \frac{w^T S_B w}{w^T S_W w}. \quad (1)$$

Here, m_k and m_n are each mean of E_{kj} and E_{nj}, s_k^2 and s_n^2 are variances of E_{kj} and E_{nj}, w is projection vector, and S_B and S_W

Table 1: Evaluation results. The MOS scores are presented with 95% confidence intervals. Higher is better for MOS and emotion accuracy, and lower is better for the other metrics. 'FS2' denotes that the model is based on FastSpeech2, and Emo. Acc. denotes the emotion accuracy.

	KES (single speaker)				**ETOD** (multi speakers)			
Model	**MOS**	**MCD**$_{13}$	**RMSE**$_{f_0}$	**Emo. Acc.**	**MOS**	**MCD**$_{13}$	**RMSE**$_{f_0}$	**Emo. Acc.**
Ground Truth	4.29 ± 0.03	–	–	99.51%	4.54 ± 0.03	–	–	98.55%
Vocoded	3.98 ± 0.04	2.60	51.23	99.41%	4.09 ± 0.04	1.70	30.53	89.16%
TP-GST [7] + FS2 [17]	3.68 ± 0.04	4.89	54.43	98.16%	3.52 ± 0.05	3.69	31.55	81.97%
FEP [15] + FS2 [17]	3.66 ± 0.04	4.87	55.03	98.50%	3.21 ± 0.05	6.03	40.14	79.94%
EmoQ-TTS	$\mathbf{3.72 \pm 0.04}$	**4.81**	**53.15**	**99.39%**	$\mathbf{3.95 \pm 0.04}$	**2.94**	**30.61**	**86.85%**

are between-class variance and within-class variance of two clusters. After obtaining w^* by binary-class LDA, the emotion embeddings E_{kj} and E_{nj} are projected onto optimal projection vector w^* and become the projected emotion embeddings $E_{kj}^{'}$ and $E_{nj}^{'}$ as

$$E_{kj}^{'} = \frac{E_{kj} \cdot w^*}{||w^*||_2^2} w^*. \tag{2}$$

Then, the emotion distance d_{kj} is measured by the euclidean distance between the projected emotion embedding $E_{kj}^{'}$ and the centroid of projected neutral embeddings $E_{nj}^{'}$ as

$$d_{kj} = ||E_{kj}^{'} - M_n^{'}||_2 \quad \text{where} \quad M_n^{'} = \frac{1}{N_n} \sum_{j=1}^{N_n} E_{nj}^{'}. \tag{3}$$

Here, $M_n^{'}$ represents the centroid of projected neutral embedding. After extracting emotion distance, We remove outliers for each emotion using the interquartile range technique [23]. Through min-max normalization, the emotion distances turn to intensity scalar which has a value between 0 and 1. As we explained in the previous section, this intensity scalar is quantized into N_I-sized emotion intensity pseudo-labels at regular intervals.

Furthermore, we introduce the intensity embedding table which reflects the intensity pseudo-labels. In the intensity embedding table, the embedding of l-th intensity is set by the average of emotion embedding of reference encoder corresponding to intensity pseudo-label as follows:

$$I_{kl} = \frac{1}{N_{kl}} \sum_{i_{kj} \in C_{kl}} E_{kj} \tag{4}$$

where I_{kl} denotes l-th intensity embedding of $emotion_k$ in the intensity embedding table, i_{kj} denotes intensity pseudo-label of E_{kj}, C_{kl} refers to the group of i_{kj} corresponding to the l-th intensity, and N_{kl} represents the number of emotion embedding such that $i_{kj} \in C_{kl}$. The intensity labels are the entry index of the intensity embedding table for each emotion. This embedding table is used in the emotion renderer of EmoQ-TTS.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

We use the Korean Emotional Speech (KES) dataset [24] for single-speaker models. The KES dataset contains about 21,000 speeches of 30 hours recorded by a Korean professional female speaker with seven emotions (neutral, happy, sad, angry, surprised, fearful, and disgusted). In addition, we use the EmotionTTS Open DB (ETOD) dataset [25] for multi-speaker models. The ETOD dataset contains about 6,000 speeches of 10 hours recorded by 13 Korean speakers (five female and eight male) with four emotions (neutral, happy, sad, and angry). Both KES and ETOD datasets are split into train, validation, and test set, 100 sentences of each emotion category are reserved as validation and test set to evaluate the performance of emotional TTS and controllability. In the emotion intensity quantizer, the N_I as the total number of quantized intensity pseudo-labels is set to 16 for each emotion. In the case of multi-speaker EmoQ-TTS, we add a speaker classifier with GRL to the reference encoder to remove speaker information from emotion embedding.

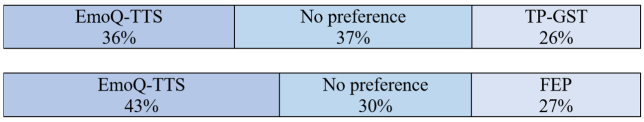


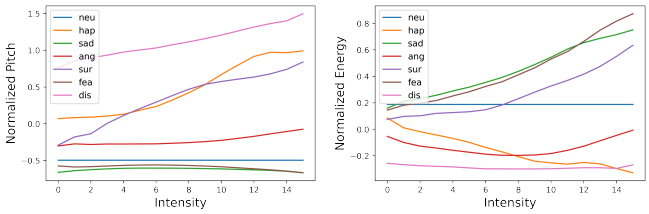
Fig. 2: A/B preference test

4.2. Model Performance

To evaluate naturalness, we conducted the mean opinion score (MOS) test [26]. In the test, 20 subjects are asked to rate naturalness on a 5-scale for 105 generated speeches which are sampled uniformly for each emotion. For the baseline, we use two methods: TP-GST [8] which predicts global style embedding from text, and fine-grained emotion prediction model (FEP) [15] which also predicts phoneme-level intensity scalar from text. For a fair comparison, we implemented these two methods based on FastSpeech2. We use Fre-GAN [27] as our neural vocoder to generate waveforms from mel-spectrogram.

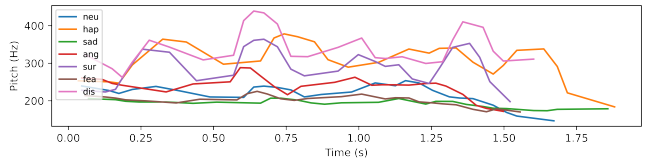
Table 1 shows evaluation results of the EmoQ-TTS compared to ground truth, vocoded speech, and two competitive baselines. The MOS score of EmoQ-TTS outperforms baseline models for both single-speaker and multi-speaker datasets. Additionally, we computed mel cepstral distortion (MCD) [28], root mean squared error (RMSE) of log f_0 [29] metrics for objective evaluation. Furthermore, we calculated emotion classification accuracy through external speech emotion recognition (SER) model [30]. For comparison, we use dynamic time warping (DTW) [31] which matches the duration between predicted features and ground truth features. The objective evaluation shows the same result with the MOS test as projected in Table 1. Especially in the multi-speaker model, FEP shows low performance. We found that the ranking function of FEP is unable to decompose speaker information and intensity information well.

Moreover, we conducted an A/B preference test on the single-speaker dataset to evaluate emotional expressiveness. In the test, The subjects are asked to choose which of the two speeches generated in the same sentence by different models is perceptually more expressive. As shown in Fig.2, EmoQ-TTS got more preference than

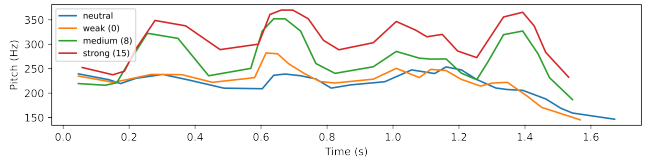


(a) Pitch (b) Energy

Fig. 3: The feature tendency according to emotion and intensity



(a) Emotion control



(b) Intensity control (surprised)

Fig. 4: Pitch tracks of one sample

both TP-GST and FEP. Additionally, we have conducted a 1-sample t-test on the result. As the result of the t-test, the preference test with TP-GST got p-value=10^{-9} and the preference test with FEP got p-value=10^{-20}. In both cases, the results demonstrate the difference is statistically significant under the significance level which is typically less than 0.05. It demonstrates EmoQ-TTS generates more expressive and robust speech than other baseline methods.

4.3. Emotion Controllability

To demonstrate the ability to control emotional expression, we visualized the tendency of pitch and energy as shown in Fig.3. These values were computed by averaging the pitch and energy from the synthesized speech which is conditioned by all combinations of emotion and intensity labels from 105 test sentences. For example, in the "happy" emotional speech, the pitch tends to increase as the intensity increases, but energy has a tendency to decrease. This indicates that synthesized speech from EmoQ-TTS reflects each intensity of emotions in a versatile manner. Moreover, Fig.4 shows pitch tracks of one sample. It can be seen that pitch and duration are reflected according to the emotion and intensity well even in one sample.

4.4. Ablation Study

To validate the effectiveness of each component of EmoQ-TTS, we conducted an ablation study and the result is in Table 2. In the case of phoneme classifier and auxiliary predictor, the MOS score decreased after removing either one or both. It demonstrates that both components are effective in reflecting emotional expression to embedding.

Table 3 represents an ablation study of the projection method for feature extraction. We compare the LDA method to L_1 distance without projection and the principal component analysis (PCA) [32] which is one of the other projection-based methods. As a result, the proposed model with LDA outperforms the other ablation models. Specifically, the L_1 model was not trained well since the intensity was not properly modeled.

Table 4 shows an ablation study of intensity quantization. The ablation model uses continuous emotion distance before pseudo-labeling as emotion intensity. The result shows that the ablation model attained much lower performance due to unstable synthesized speech because of noise. It can be seen that the intensity quantization contributes to synthesizing the speech robust.

Table 2: Ablation study of reference encoder.

Setting	MOS	MCD_{13}
EmoQ-TTS	**3.94 ± 0.04**	**4.81**
w/o Phoneme Cls.	3.90 ± 0.04	4.82
w/o Auxiliary Pred.	3.84 ± 0.04	4.82
w/o Both Cls. and Pred.	3.81 ± 0.04	4.84

Table 3: Ablation study of projection method.

Setting	MOS	MCD_{13}
LDA proj. (Proposed)	**3.88 ± 0.04**	**4.81**
PCA proj.	3.85 ± 0.04	4.86
L_1 distance w/o proj.	3.36 ± 0.04	4.93

Table 4: Ablation study of intensity quantization.

Setting	MOS	MCD_{13}
EmoQ-TTS	**3.84 ± 0.04**	**4.81**
w/o Intensity Quantization	3.48 ± 0.04	5.45

5. CONCLUSION AND FUTURE WORKS

This paper proposed the EmoQ-TTS, which synthesizes expressive emotional speech by predicting phoneme-wise emotion information with fine-grained emotion intensity. To reflect expressive emotional expression without human labeling, we proposed the emotion intensity modeling. Moreover, we achieved robustness of the synthesized speech through the distance-based quantized intensity pseudo-labels and the intensity embedding table. Experimental results show the superiority of EmoQ-TTS in emotional expressiveness. In addition, we demonstrate that controlling emotion expression easily by manual intensity labels. For future work, we will apply the proposed methods to other tasks such as predicting the aging factors.

6. ACKNOWLEDGEMENT

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Department of Artificial Intelligence, Korea University), and the Magellan Division of Netmarble Corporation.

This study (or Project) used on open Speech database as the result of research supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (No. 10080667, Development of conversational speech synthesis technology to express emotion and personality of robots through sound source diversification).

7. REFERENCES

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards End-to-end Speech Synthesis," in *Proc. INTERSPEECH*, 2017.

[2] Sang-Hoon Lee, Hyun-Wook Yoon, Hyeong-Rae Noh, Ji-Hoon Kim, and Seong-Whan Lee, "Multi-SpectroGAN: High-Diversity and High-Fidelity Spectrogram Generation with Adversarial Style Combination for Speech Synthesis," in *Proc. the AAAI Conference on Artificial Intelligence*, 2021.

[3] Hyunseung Chung, Sang-Hoon Lee, and Seong-Whan Lee, "Reinforce-Aligner: Reinforcement Alignment Search for Robust End-to-End Text-to-Speech," in *Proc. INTERSPEECH*, 2021.

[4] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A Survey on Neural Speech Synthesis," *arXiv:2106.15561*, 2021.

[5] Heiga Ze, Andrew Senior, and Mike Schuster, "Statistical Parametric Speech Synthesis using Deep Neural Networks," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *Advances in neural information processing systems*, 2017.

[7] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," in *Proc. International Conference on Machine Learning (ICML)*, 2018, pp. 5180–5189.

[8] Daisy Stanton, Yuxuan Wang, and RJ Skerry-Ryan, "Predicting Expressive Speaking Style from Text in End-to-End Speech Synthesis," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 595–602.

[9] Younggun Lee, Azam Rabiee, and Soo-Young Lee, "Emotional End-to-End Neural Speech Synthesizer," *arXiv:1711.05447*, 2017.

[10] Paarth Neekhara, Shehzeen Hussain, Shlomo Dubnov, Farinaz Koushanfar, and Julian McAuley, "Expressive Neural Voice Cloning," *arXiv:2102.00151*, 2021.

[11] Tao Li, Shan Yang, Liumeng Xue, and Lei Xie, "Controllable Emotion Transfer for End-to-End Speech Synthesis," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1–5.

[12] Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei Xie, "Controllable Cross-Speaker Emotion Transfer for End-to-End Speech Synthesis," *arXiv:2109.06733*, 2021.

[13] Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang, "Emotional Speech Synthesis with Rich and Granularized Control," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7254–7258.

[14] Oytun Turk, Marc Schröder, Baris Bozkurt, and Levent M Arslan, "Voice Quality Interpolation for Emotional Text-to-Speech Synthesis," in *European Conference on Speech Communication and Technology*, 2005.

[15] Yi Lei, Shan Yang, and Lei Xie, "Fine-grained Emotion Strength Transfer, Control and Prediction for Emotional Speech Synthesis," in *Proc. Spoken Language Technology Workshop (SLT)*, 2021, pp. 423–430.

[16] Devi Parikh and Kristen Grauman, "Relative Attributes," in *Proc. International Conference on Computer Vision (ICCV)*, 2011, pp. 503–510.

[17] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and High-quality End-to-End Text to Speech," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[18] Adrian Lańcucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6588–6592.

[19] Cort J Willmott and Kenji Matsuura, "Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance," *Climate research*, pp. 79–82, 2005.

[20] Yaroslav Ganin and Victor Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proc. International Conference on Machine Learning (ICML)*, 2015, pp. 1180–1189.

[21] Zhou Wang and Alan C Bovik, "Mean Squared Error: Love It or Leave It? A New Look at Signal Fidelity Measures," *IEEE signal processing magazine*, pp. 98–117, 2009.

[22] Suresh Balakrishnama and Aravind Ganapathiraju, "Linear Discriminant Analysis-a Brief Tutorial," *Institute for Signal and information Processing*, pp. 1–8, 1998.

[23] Steven Walfish, "A Review of Statistical Outlier Methods," *Pharmaceutical technology*, p. 82, 2006.

[24] AIHub, "Korean Emotional Speech Dataset," 2019.

[25] SelvasAI, "EmotionTTS-open-DB Dataset," 2019.

[26] Robert C Streijl, Stefan Winkler, and David S Hands, "Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives," *Multimedia Systems*, 2016.

[27] Ji-Hoon Kim, Sang-Hoon Lee, Ji-Hyun Lee, and Seong-Whan Lee, "Fre-GAN: Adversarial Frequency-consistent Audio Synthesis," in *Proc. INTERSPEECH*, 2021.

[28] Robert Kubichek, "Mel-cepstral Distance Measure for Objective Speech Quality Assessment," in *IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PACRIM)*, 1993.

[29] Tianfeng Chai and Roland R Draxler, "Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)?–Arguments against Avoiding RMSE in the Literature," *Geoscientific model development*, pp. 1247–1250, 2014.

[30] Amir Shirian and Tanaya Guha, "Compact Graph Architecture for Speech Emotion Recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6284–6288.

[31] Donald J Berndt and James Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," in *KDD workshop*, 1994, pp. 359–370.

[32] Svante Wold, Kim Esbensen, and Paul Geladi, "Principal Component Analysis," *Chemometrics and intelligent laboratory systems*, pp. 37–52, 1987.