# IMPROVING EMOTIONAL SPEECH SYNTHESIS BY USING SUS-CONSTRAINED VAE AND TEXT ENCODER AGGREGATION

*Fengyu Yang, Jian Luan, Yujun Wang*

Xiaomi Corporation, Beijing, China

## ABSTRACT

Learning emotion embedding from reference audio is a straightforward approach for multi-emotion speech synthesis in encoder-decoder systems. But how to get better emotion embedding and how to inject it into TTS acoustic model more effectively are still under investigation. In this paper, we propose an innovative constraint to help VAE extract emotion embedding with better cluster cohesion. Besides, the obtained emotion embedding is used as query to aggregate latent representations of all encoder layers via attention. Moreover, the queries from encoder layers themselves are also helpful. Experiments prove the proposed methods can enhance the encoding of comprehensive syntactic and semantic information and produce more expressive emotional speech.

***Index Terms***— emotional TTS, variational autoencoder, emotion embedding, encoder aggregation

## 1. INTRODUCTION

Emotional speech synthesis is widely applied in various scenarios, such as voice assistant, audio customer service, audio book and etc. Generally, a group of emotion types are defined based on product requirements. For each emotion type, hundreds of sentences are designed and recording lines are collected accordingly. Due to the limited data of each type, usually a multi-emotion TTS model training is implemented by leveraging all the data. To distinguish the data from different emotion types, two questions emerge:1) how to get the emotion embedding from category label; 2) how to inject the emotion embedding into TTS acoustic model. Based on the popular encoder-decoder TTS frameworks, many attempts have been reported to address these two questions.

For the emotion embedding, the basic idea is to learn an embedding vector for each emotion type [1, 2], which is usually called as one-hot embedding or look-up table. However, emotion expression always has slight vibration in intensity or status among utterances, even if the voice talent is guided to keep consistent. Global Style Token(GST) [3] is then introduced to get utterance-level embedding. At the same time, an emotion classifier can be added to restrict embedding vectors to be clustered by category [4]. During inference, the average embedding vector of target emotion type may be employed [5], while [6] even tries to control the emotion intensity and inter-emotion transition by interpolating the embedding. Also, GST can generate fine-grained embedding at phoneme level [7]. In recent studies, Variational AutoEncoder(VAE) [8] shows stronger capabilities in disentanglement, scaling and interpolation for expression modeling [9] and style control[10]. However, VAE training is not robust and usually suffers from posterior collapse.

For the injection of emotion embedding, mostly popular method is to concatenate the embedding vector into decoder input [1, 3, 5, 6, 4, 7, 9, 10], while some studies inject emotion embedding to both attention and decoder RNN layers of Tacotron framework [2]. Both of them influence the decoder merely, not considering the effects of emotion embedding to the textual emotion.

In this paper, we propose a framework with innovative solutions for both of the above questions. Firstly, VAE is applied for the emotion embedding. Instead of conventional KL-divergence regularizer, the new constraint expects the means of the embedding vectors are on the surface of the unit sphere while all dimensions have a uniform standard deviation. Secondly for the injection of emotion embedding, this paper uses it as one resource of queries in the attention-based text encoder aggregation to enable emotion-specific sentential information encoding. Another resource of queries are from text encoder themselves as we did in previous study [11]. In summary, the multi-query attention is designed to capture the syntactic and semantic information better for emotional speech generation.

## 2. RELATED WORK

Some previous works also employ VAE embedding as query to attend encoder output. In BVAE-TTS [12], reference audio is encoded by VAE to get frame-level latent variables. These variables work as query to attend text encoder output for better decoder input. In VARA-TTS [13], however, the output of VAE intermediate layers (called hierarchical latent variables) are all used as query. Different from them, the proposed method encodes reference audio into one vector, i.e. a utterance level emotion embedding, rather than a frame level sequence. Moreover, both BVAE-TTS and VARA-TTS only

use the output from the last layer of text encoder as attention memory and the attention is implemented along the time axis. In our framework, the outputs from all intermediate encoder layers are leveraged and the attention is implemented along the stacked layers. In our previous work[11], we utilize the contexts extracted form the stacked layers to do self-learned multi-query attention over an expressive corpus. In this paper, we propose to introduce acoustic emotion information to the multi-query for better emotion modeling over a multi-emotion corpus.

## 3. PROPOSED MODEL

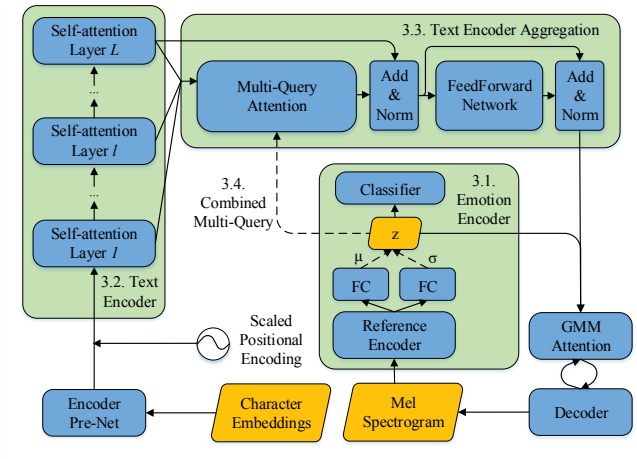



**Fig. 1**. Proposed architecture with multi-query attention.

Figure1 illustrates our proposed approach with text encoder aggregation on exploiting emotional contexts for emotional speech synthesis. It contains a self-attention-based text encoder, an RNN-based auto-regressive decoder, a GMM-based attention[14] bridging them, a VAE-based emotion encoder and an emotion classifier. WaveRNN[15] is adopted to convert mel spectrogram to waveforms. The augmented encoder with a context aggregation module will be described in detail.

### 3.1. SUS-constrained VAE

VAE does not generate the latent vector directly. Instead, it generates Gaussian distributions each represented by a mean and a standard deviation. During inference, a latent vector is sampled from these distributions. If no additional constraints are employed, the standard deviation will trend to be 0 and sampled latent vectors will always be the mean. Therefore, the desired sampling mechanism becomes invalid. However if a Kullback-Leibler (KL) divergence regularizer is added, it is usually found the generated distributions become normal Gaussian distribution independently of the inputs. Both these two phenomena can be regarded as posterior collapse. Many attempts have been made to address this puzzle, such as annealing strategy in [10].

The critical problem here, we think, is the distances between the means should be in the similar order with their standard deviations. If the distance between means is much bigger than their standard deviation, latent vectors will collapse to the means. Conversely, if the distance between means is much smaller, latent vectors will collapse to be independent on the input. Inspired by it, this paper restricts the means approaching to the Surface of the Unit Sphere (SUS) while set the standard deviations to be an appropriate constant for all dimensions, such as 1. In this way, the distributions of latent vectors will finally have appropriate overlapping proportions, which guarantees VAE's advantages of disentanglement, scaling and interpolation.

Formally, sampling $z$ from distribution $N(\mu, \sigma^2 I)$ is decomposed to first sampling $\epsilon \in (0, I)$ and then computing $z = \mu + \sigma \cdot \epsilon$, where $\cdot$ represents an element-wise product. We restrict the means $\mu$ approaching to the surface of the unit sphere through L2 distance:

$$loss_{SUS} = (\sqrt{\sum(\mu^2)} - 1)^2. \quad (1)$$

Meanwhile, we set the standard deviations $\sigma$ as a constant.

### 3.2. Self-attention based Encoder

Self-attention based sequence-to-sequence framework has been successfully applied to speech synthesis, such as Transformer-TTS[16] and Fastspeech[17]. We also adopt self-attention networks(SAN) as our based text encoder following [16]. Formally, from the previous self-attention block output $H^{l-1}$, the multi-head attention $C^l$ and the followed feed forward network $H^l$ can be computed by:

$$C^l = \mathrm{LN}(\mathrm{MH}(head_1^l, \ldots, head_H^l) + H^{l-1}), \quad (2)$$

$$H^l = \mathrm{LN}(\mathrm{FFN}(C^l) + C^l), \quad (3)$$

where MH$(\cdot)$, FFN$(\cdot)$ and LN$(\cdot)$ represent multi-head attention, feed forward network and layer normalization respectively. And in multi-head attention, each head split from the previous self-attention block is calculated as:

$$head_h = \mathrm{softmax}(\frac{Q_h K_h^T}{\sqrt{d}}) \cdot V_h, \quad (4)$$

where $\{Q, K, V\}$ are queries, keys and values, $d$ represents the hidden state's dimension.

### 3.3. Weighted Aggregation

As different SAN layers extract different levels of prosodic-related sentential context information[18], we propose a text encoder aggregation module, aggregating them to learn a comprehensive sentence representation to enhance the emotion of the final generated speech. In detail, we utilize a multi-query attention to learn the contribution of each block across the stacked layers. Formally, given a sequence $X$ of

$N$ elements, the multi-query calculates the correlation of individual sentential contexts $\{H^0, \ldots, H^L\}$ in phoneme level, which are transposed as $\{head_1^g, \ldots, head_N^g\}$ for keys and values. We modify Eq. (2) to obtain the weighted contexts:

$$C^g = \mathrm{LN}(\mathrm{MH}(head_1^g, \ldots, head_N^g) + H^L), \quad (5)$$

$$H^g = \mathrm{LN}(\mathrm{FFN}(C^g) + C^g). \quad (6)$$

There are several choices for the multi-query. The first is as our previous work does[11], utilizing $\{H^0, \ldots, H^L\}$ to obtain the textual multi-query $Q^t$. This self-learned weighted aggregation module leverages the textual contexts information to learn the combination relationship across layers.

### 3.4. Combined multi-query

For the textual multi-query, the sententail contexts are totally extracted from the stacked textual encoder layers, which does not consider the proved important information from emotion embedding. Commonly the emotion embedding is directly concatenated to the encoder output, influencing the decoder merely. Assuming emotion embedding affects the textual emotion significantly, we propose to introduce contexts extracted from emotion embedding to out multi-query on the basis of direct concatenation.

In details, we employ the output of VAE $vae$ to learn a weighted matrix $Q^a$:

$$Q^a = \mathrm{DNN}(vae), \quad (7)$$

where DNN$(\cdot)$ represents a nonlinear transformation with tanh. Then, we combine the contexts information from text and emotion embedding with a learned coefficient to investigate the effectiveness of a comprehensive multi-query:

$$Q^c = Q^t + \mathrm{Sigmiod}(w)Q^a, \quad (8)$$

where Sigmoid$(\cdot)$ is a activation function.

## 4. EXPERIMENTS

### 4.1. Basic setups

To investigate the effectiveness of modeling emotion, we carried out experiments on a Mandarin corpora from a male speaker with 7 emotion categories (neutral, happy, sad, angry, shy, concerned and surprised), which contains about 4.4 hours and a total of 4371 utterances. Except for the neutral, each emotion categories has nearly 500 utterance with consistent emotional strength, separated to non-overlapping training and testing sets (with data ratio 9:1) respectively. For linguistic inputs, we use phones, tones, character segments and three levels of prosodic segments: prosodic word (PW), phonological phrase (PPH) and intonation phrase (IPH). And 80-band mel-spectrogram is extracted from 16KHz waveforms as acoustic targets. For objective evaluation, we conduct mel cepstral distortion (MCD) on test set. And we conduct A/B preference test on 30 randomly selected test set samples with 20 native Chinese listeners as subjective evaluation.

**Table 1**. MCD scores of different systems for parallel transfer.

| Emotion | BASE | BASE-SUS | SA-WA | SA-WAC |
|---|---|---|---|---|
| Neutral | 5.5 | 5.15 | 4.44 | **4.22** |

### 4.2. Model details

The decoder structure in Tacotron2[19] is used as our baseline. But the CBHG encoder and GMMv2 attention are adopted instead for superior naturalness and stability[14], where the output of VAE[10] is added to the encoder output. For the encoder using SAN, the input text embeddings with positional information are pushed to a 3-layer CNN firstly. Then each self-attention block includes an 8-head self-attention and a feed forward sub-network. In the text encoder, there are totally 6 self-attention blocks. As for the aggregation module, $H^L$ is double fed for the convenience of implementation. In VAE module [10], the reference encoder consists of six 2D-convolution layers and a GRU layer. Further, the plugged emotion classifier in all systems has a fully connected (FC) layer with ReLu activation and a 7-unit output layer. In our proposed SUS-constrained VAE, we set the standard deviation to 1. WaveRNN is used as vocoder totally following [15], trained using the neutral set about 16 hours with the same speaker. For comparison, we built the following different systems:

- **BASE**: Baseline system following VAE-Tacotron2 [10] with CBHG as text encoder and slightly modified GMMv2 attention.
- **BASE-SUS**: Baseline system with SUS constraint instead of KL divergence constraint for VAE training described in Section 3.1.
- **SA-WA**: SAN based encoder with the aggregation module using textual multi-query described in Section 3.3.
- **SA-WAC**: SAN based encoder with the aggregation module using combined multi-query described in Section 3.4. (The learned coefficient for combination over the multi-emotion corpus is 0.47.)

### 4.3. Objective Evaluation

The MCD results of different systems with parallel transfer are showed in Table 1. It shows that SUS-constrained VAE has lower MCD than KL constrained VAE, with better ability in generating more similar emotional speech for ground-truth. Meanwhile, it demonstrates that SAN based encoder with text encoder aggregation can improve the performance than the RNN based encoder, where combined multi-query is a better way than textual multi-query in text encoder aggregation to

extract emotional contexts. With the help of the injection of the emotion embedding to the textual multi-query, the synthesized speech samples turn into more similar ones to the real speech samples.

### 4.4. Subjective Evaluation

As Figure 2 shows, we conduct AB preference tests on the emotional test sets with non-parallel transfer, which include 7 emotion categories. The listeners are asked to select preferred audio according to the overall impression on the expressiveness of emotion in the testing samples[1]. Comparing the BASE system, we find that with the SUS-constrained VAE, our proposed BASE-SUS system obtains much more preferred, due to more expressive emotional speech. Meanwhile, both of the two systems using self-attention based encoder and text encoder aggregation achieve higher preference scores than the RNN based encoder one, which demonstrates that self-attention based aggregation module is a better strategy for generating more expressive emotional speech. Further, combined multi-query attention brings extra performance gain than simple textual multi-query attention according to the A/B preference test.

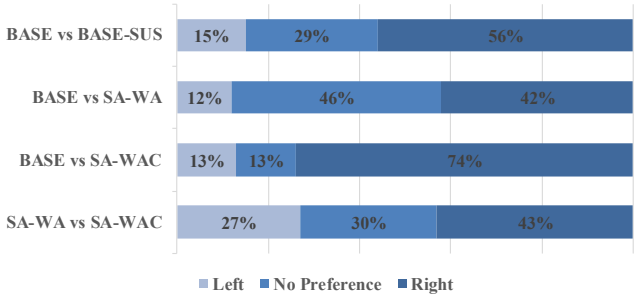


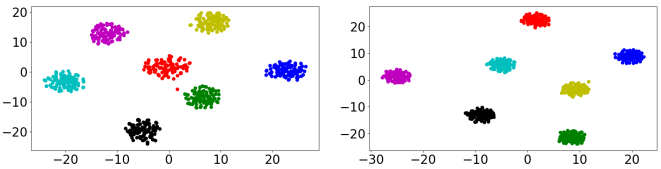

**Fig. 2**. A/B preference results for non-parallel transfer with confidence intervals of 95% and p-value<0.0001 from t-test.

### 4.5. Analysis

**Emotion Distortion** We visualize the two systems with parallel transfer for seven emotion categories in emotion embedding space by t-distributed stochastic neighbor embedding (t-SNE) plots[20]. Figure 3 shows that both the BASE system and the BASE-SUS system appear clear cluster separation, which demonstrates that both two systems have high classification accuracy. But the cluster cohesion of the BASE-SUS system is much better than that of the BASE system. It means that the proposed SUS-constrained VAE can extract emotion information more robustly with less disturbance, which finally helps TTS to generate emotional speech more accurately and expressively. They are also certified in above objective and subjective evaluations.

**Prosody Correlation** To further estimate the emotion with parallel transfer for statistical significance, phoneme-level intensity, rhythm and intonation of audio are selected.

[1]Samples can be found from: https://fyyang1996.github.io/emotion



(a) The BASE system (b) The BASE-SUS system

**Fig. 3**. Visualization of two systems using t-SNE for seven emotion categories.

**Table 2**. Correlation in relative energy, duration and F0 within a phoneme computed from different systems for parallel transfer.

| | BASE | BASE-SUS | SA-WA | SA-WAC |
|---|---|---|---|---|
| E | 0.542 | 0.56 | 0.582 | **0.595** |
| Dur. | 0.806 | 0.811 | 0.820 | **0.824** |
| F0 | 0.322 | 0.338 | 0.403 | **0.422** |

We extract three acoustic features commonly associated with emotion: relative energy within each phoneme (E), duration in ms (Dur.) and fundamental frequency in Hertz (F0) according to [21, 11]. Additional alignments are done to catch the three prosody attributes in phoneme level. The Pearson correlation coefficient between each system and the ground truth is calculated to evaluate these statistics, using 100 random samples in the test set. The higher Pearson correlation coefficient value demonstrates the higher accuracy of the predicted prosody attribute.

From Table 2 we know that out proposed BASE-SUS achieves higher correlation scores than baseline in all three prosody attributes, which demonstrates that our SUS-constrained VAE has better reconstruction performance in phoneme-level intensity, rhythm and intonation. Meanwhile, in all three prosody attributes, our proposed both SA-WA system and SA-WAC system obtain higher correlation scores than baseline, and SA-WAC system acquires the highest scores. Consequently, we believe that the combined multi-query attention in text encoder aggregation has strong ability in modeling all the three emotional associated attributes.

## 5. CONCLUSION

In this paper, SUS-constrained VAE is proposed to extract emotion embedding with better cluster cohesion. Then, based on our previous work of the text encoder aggregation, we introduce the emotion embedding as one resource of queries for attention-based text encoder aggregation. It is with the belief that emotion type should affect the sentential information encoding. Experiments demonstrate that the proposed methods can enhance the encoding of syntactic and semantic information and produce more expressive emotional speech. Moreover, we believe that they can be easily scaled to other tasks, like multi-style or multi-speaker speech synthesis.

## 6. REFERENCES


[1] Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, , and Yuta Ochiai, "Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis," in *Speech Communication*, 2018, vol. 99, pp. 135–143.

[2] Younggun Lee, Azam Rabiee, , and Soo-Young Lee, "Emotional end-to-end neural speech synthesizer," in *arXiv:1711.05447*, 2017.

[3] 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. ICML*, 2018, p. 5180–5189.

[4] Pengfei Wu, Zhenhua Ling, Lijuan Liu, Yuan Jiang, Hongchuan Wu, , and Lirong Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *Proc. APSIPA ASC*, 2019, p. 623–627.

[5] Ohsung Kwon, Inseon Jang, ChungHyun Ahn, , and Hong-Goo Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE SPL*, vol. 26, no. 9, pp. 1383–1387, 2019.

[6] 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. ICASSP*, 2020, p. 7254–7258.

[7] Chunhui Lu, Xue Wen, Ruolan Liu, and Xiao Chen, "Multi-speaker emotional speech synthesis with fine-grained prosody modeling," in *Proc. ICASSP*, 2021, pp. 5714–5718.

[8] Diederik P Kingma and Max Welling, "Autoencoding variational bayes," in *Proc. ICLR*, 2014.

[9] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *arXiv:1804.02135*, 2018.

[10] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *Proc. ICASSP*, 2019, p. 6945–6949.

[11] Fengyu Yang, Shan Yang, Qinghua Wu, Yujun Wang, and Lei Xie, "Exploiting deep sentential context for expressive end-to-end speech synthesis," *arXiv preprint arXiv:2008.00613*, 2020.

[12] Yoonhyung Lee, Joongbo Shin, and Kyomin Jung, "Bidirectional variational inference for non-autoregressive text-to-speech," in *Proc. ICLR*, 2021.

[13] Peng Liu, Yuewen Cao, Songxiang Liu, Na Hu, Guangzhi Li, Chao Weng, and Dan Su, "Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention," in *arXiv:2102.06431*, 2021.

[14] Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. ICASSP*, 2020, pp. 6189–6193.

[15] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, and Koray Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, 2018, p. 2415–2424.

[16] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6706–6713.

[17] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, 2019, p. 3165–3174.

[18] Haohan Guo, Frank K Soong, Lei He, and Lei Xie, "Exploiting syntactic features in a parsed tree to improve end-to-end tts," in *2019 International Speech Communication Association (Interspeech)*. 2019, pp. 4460–4464, ISCA.

[19] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[20] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[21] Guangzhi Sun, Yu Zhang, Ron J Weiss, Yuan Cao, Heiga Zen, Andrew Rosenberg, Bhuvana Ramabhadran, and Yonghui Wu, "Generating diverse and natural text-to-speech samples using a quantized fine-grained vae and auto-regressive prosody prior," *arXiv preprint arXiv:2002.03788*, 2020.