# Improving Latent Representation For End To End Multispeaker Expressive Text To Speech System

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

# IMPROVING LATENT REPRESENTATION FOR END TO END MULTISPEAKER EXPRESSIVE TEXT TO SPEECH SYSTEM

*Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet*

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France.

## ABSTRACT

The main goal of this work is to generate expressive speech in different speaker's voices for which no expressive speech data is available. To do that, we propose to use multiclass N-pair loss in end-to-end multispeaker expressive Text-To-Speech (TTS) for improving the transfer of expressivity to the target speaker's voice. This augmentation of the loss function during training paves the way to enhance the latent space representation of emotions. The presented approach condition tacotron based end-to-end system with latent representation extracted from the expressivity encoder. We have jointly trained the end-to-end (E2E) TTS with multiclass N-pair loss to discriminate between various emotions.

We experimented with two neural network architectures for expressivity encoder namely global style token (GST) and variational autoencoder (VAE). We transferred the expressivity using the mean of latent representation extracted from the expressivity encoder for each emotion. The obtained results show that adding multiclass N-pair loss based deep metric learning in training process improves expressivity in the desired speaker's voice.

***Index Terms***— End-to-end TTS, metric learning, expressivity, transfer learning

## 1. INTRODUCTION

With recent advancements in computational power, end-to-end TTS systems can generate highly intelligible and natural voice. The term expressivity in speech usually refers to the characteristics of speech, such as emotions, speaking style, the relationship of speech with gestures, and facial expression. Throughout this paper, we considered only the emotional characteristics of expressivity in speech. Current end-to-end TTS system heavily relied on a large amount of speech corpus used for training the system [1]. Therefore, to build expressive speech synthesis for a new speaker, we have to create a speech corpus with various emotions. It is inconvenient to record an expressive speech corpus every time we want to build an expressive speech synthesis system for a new speaker's voice. Furthermore, creating an expressive speech corpus is laborious and expensive in terms of workload. This involves speech acquisition, labeling, alignment, and evaluation of expressive speech corpus. It is inconvenient to record an expressive speech corpus for the new speaker's voice.

Many frameworks have been proposed for the implementation of expressivity transfer by interpolation either of latent space of prosody or of prosody embedding [2, 3, 4, 5, 6, 7]. For controlling expressivity, these approaches enhance the tacotron based TTS system by the addition of advanced deep neural network architecture such as variational autoencoder (VAE) [5], Gaussian mixture VAE [2], Global style token [3], FLOW [8], etc.

The systems mentioned above have shown significant performance in controlling expressivity but limited to emotive storytelling style instead of emotions such as joy, sad, happy, fear, anger, etc. For the transfer of expressivity for emotions, very few approaches have addressed the usage of multiple emotions. For instance, Deep convolutional TTS (DCTTS) was trained with multiple emotions along with variational inference [9]. Recently metric learning framework was introduced in a parametric TTS system [10, 11] for transfer of expressivity. In [10], the author proposed to build a recurrent conditional variational autoencoder based acoustic model with multiclass N-pair [12]. This work was further extended by the addition of Inverse Autoregressive Flow (IAF) for implementing an encoder network of the acoustic model [11].

The above approaches indicate that the addition of multiclass N-pair loss increases the perceived expressivity in the target speaker's voice. The above approaches have shown promising results for acoustic modeling but still dependent on the bottleneck step of the duration model. Also, synthesizing various emotions need a change in phoneme duration. In this paper, we present a novel metric learning framework [13] jointly trained with a tacotron 2 based end-to-end TTS system. This results in enhancing the latent representation of expressivity for better transfer learning performance. The representation of expressvity learned by encoder influences the alignment of synthesized speech generated through attention.

The paper is organized as follows, Section 2 describes multi-speaker expressive end-to-end TTS; Section 3 presents details about data processing before training and speech corpora used; Section 4 presents experimentation setup; results are presented in Section 5,, Section 6 discussion and Section 7 conclusion.

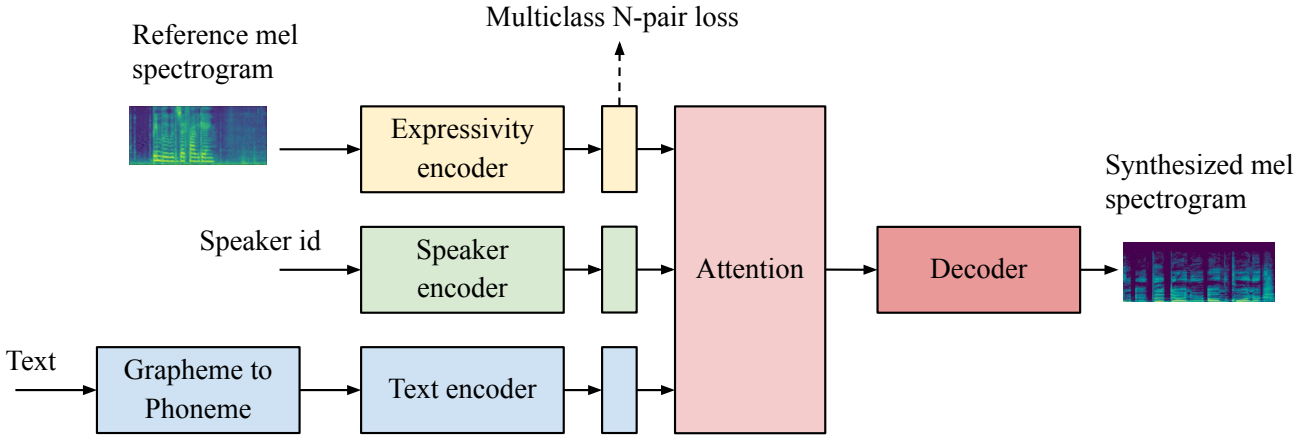


**Fig. 1**. End to End multispeaker expressive text to speech system

## 2. PROPOSED ARCHITECTURE

We extend a state-of-art Tacotron 2 model [14] based on sequence to sequence with attention module to implement end-to-end multispeaker expressive TTS system. To work with multiple speakers and expressvity, we extended Tacotron 2 approach by adding an expressivity encoder and a speaker encoder. For a more detailed explanation of Tacotron 2 architecture, refer to the work presented in [14].

### 2.1. General Framework

The proposed architecture takes input as text, which is then converted to a sequence of phonemes. The text encoder process this sequence of phonemes after passing through convolutional layers proceeded by BLSTM based recurrent neural network to get $z_t$ as a latent representation of text. The reference mel spectrogram is given to expressivity encoder to extract the latent emotional information as expressive embedding $z_e$. For enabling a multispeaker setting, we provided speaker identity to speaker encoder to create embedding, $z_s$. The speaker encoder network maps speaker index to non-linear fixed dimensional speaker embedding.

Afterwards, $z_t$, $z_e$, and $z_s$ are concatenated and given as input to location sensitive attention module, as illustrated in Fig. 1. This assists end-to-end TTS to learn the alignment between sequence of phonemes and desired Mel spectrogram. The decoder network is composed of prenet, BLSTM based recurrent network and convolutional layer based postnet. The decoder takes concatenated latent representation of text, expressivity and speaker along with attention vector to predict the Mel spectrogram frame by frame. This predicted mel spectrogram from prenet and recurrent network is further passed through postnet, which improves the overall reconstruction performance of Mel spectrogram.

### 2.2. Expressivity encoder

In this paper, we experimented with two neural network architecture for the implementation of expressivity encoder namely Global Style Token (GST) [3] and Variational Autoencoder (VAE) [5] as shown in Fig 1. The GST based expressivity encoder consists of a reference encoder, style attention, and style embedding. The reference encoder maps prosody of variable length mel spectrogram into a fixed-length vector, which is passed to style attention layer. This layer applies a multihead attention module to extract the similarity between reference embedding and each token in style embedding as an output of expressivity encoder. In this work, style embedding $z_e$ weights represents the expressiveness of each emotion as a stylistic factor to learn from reference embedding.

The second architecture for expressivity encoder is VAE based composed of reference encoder and two feedforward layers to generate mean and standard deviation of latent variable $z_e$. The reference encoder in VAE, generates hidden output which is passed through feedforward layers to obtain latent variable $z_e$. This $z_e$ is obtained using a reparameterization trick applied with mean and standard deviation. VAE based framework suffer from Kublick Libler (KL) annealing problem in which reconstruction loss is suppressed by KL loss term [15]. To avoid this, additional weight (close to zero) is multiplied to KL loss and gradually increased over the training epoch.

### 2.3. Multiclass N-pair loss

The expressivity encoder is jointly trained with end to end TTS system. This assists in predicting Mel spectrogram output in desired emotions. For the transfer of expressivity, we use precomputed means of expressivity embeddings for each emotion. Thus, during the inference phase, for a given mean

of latent variables of emotion, the system transfer expressive attributes to the target speaker's voice. The latent space representation of unclustered emotion may lead to poor transfer of expressivity. For better performance of expressivity transfer, we need the tightly bounded representation of latent variables of emotions. Therefore, we propose a novel metric learning framework implemented using multiclass N-pair loss to further enhance the expressivity representation.

The deep metric learning has gained popularity for solving discriminative tasks in computer vision and image processing domain [16]. The deep metric learning framework assists in the clustering of embeddings by reducing the distance between embeddings of positive class and increasing distance between each negative classes. The multiclass N-pair loss shown finer performance than triplet loss or contrastive loss by considering embeddings of multiple negative classes [12]. In the training phase, model needs to reduce the multiclass N-pair loss function is given in Eq.1.

$$log(1 + \sum_{i=1}^{N-1} exp(z_e^\top z_i^- - z_e^\top z^+)) \quad (1)$$

In our approach, multiclass N-pair loss reduces the distance between latent variables of the same emotion class. This loss criteria increases the intercluster distance from $N1$ negative samples and decreases the intracluster distance between positive samples and training examples [10]. The positive example refers to latent variables from the same emotion class and negative samples correspond to examples of various emotion classes. We provided the mean of latent variables of emotion for sampling the positive example and the negative examples. For $N$ classes, $z^+$ is a positive example and $z_i^-$ examples from negative classes as stated in Eq.1.

## 3. DATA PREPARATION

In this paper, we used 4 French Female speech synthesis corpora for implementing end-to-end multispeaker expressive TTS system. The speech corpora used are Lisa neutral speech corpus (approx. 3hrs, in house speech synthesis corpus), SIWIS, neutral speech corpus (approx. 5hrs) [17], Synpaflex speech corpus (approx. 7hrs) [18], and Caroline expressive speech corpus [19]. Caroline's expressive speech corpus consists of 6 emotions namely joy, surprise, fear, anger, sadness, and disgust (approx. 1hr for each emotion). Besides expressive speech, Caroline speech corpus also has neutral speech recorded for approximately 3hrs. Each speech corpus is split into train, validation, and test sets in 80:10:10 ratio respectively. In Synpaflex corpus, expressive speech samples are available but due to an insufficient number of speech samples for each emotion as well as the unbalanced distribution of emotional speech samples, we used only the neutral voice of Synpaflex in our work.

We used a sampling rate of 16000 Hz and extracted mel spectrograms as acoustic features to be predicted by the end-to-end TTS system. We applied STFT with an FFT length of 1024, hop length of 256, a window size of 1024, and extracted Mel spectrograms using 80 Mel filters. As input features to end-to-end TTS, we used a sequence of phonemes extracted from the text. For French grapheme to phoneme conversion SOJA-TTS tool (developed internally in the team) is used as a front end text processor.

## 4. EXPERIMENTAL SETUP

For training end-to-end TTS system, we used the same model parameters as explained in [3, 5, 14] for implementing the Tacotron 2 system and expressivity encoders based on GST and VAE. We used a 128 dimensional latent variable of expressivity for both GST and VAE. To avoid Kullback Leibler (KL) annealing problem, before 150K training steps, we applied the weight of 0.0001 in every 200 steps. Afterward, weight is increased by 0.00001 after every 500 steps. We adopted a similar technique for fine-tuning with multiclass N-pair loss, for which till 150K training steps a weight of 0 is applied on multiclass N-pair loss, and afterward weight is increased by 0.001 after every 200 steps.

We incorporated Waveglow [20] based neural vocoder for synthesizing speech waveform and trained it on 4 French speech synthesis corpora mentioned in Section 3. For evaluating performance improvement obtained using the addition of multiclass N-pair loss, we used an end-to-end TTS model with GST and with VAE as baseline models. We compared the baseline models with end to end TTS trained along with multiclass N-pair loss for both expressivity encoders, GST, and VAE.

## 5. RESULTS

### 5.1. Objective evaluation

We conducted an objective evaluation using Mel Cepstrum Distortion (MCD), F0 Root Mean Squared Error (F0 RMSE), and Voiced-Unvoiced error (VUV) between reference speech samples and proposed end-to-end TTS systems. The objective evaluation results are presented in Table 1.

We opt for subjective evaluation to measure the performance of transfer of expressivity, due to the unavailability of reference emotional speech samples for Lisa, Siwis, and Synpaflex speech corpora.

### 5.2. Subjective evaluation

At first, we evaluated end-to-end (E2E) multispeaker expressive TTS systems using Mean Opinion Score (MOS) [21] based listening test. In this work, we used the absolute category ranking scale, ranging from 1 to 5. Each listener had

**Table 1**. Objective evaluation for End-to-End TTS system

| Model | MCD | F0 RMSE | VUV error |
|---|---|---|---|
| E2E GST | 5.12 | 24.10 | 10.41 |
| E2E VAE | 5.29 | 24.24 | 10.72 |
| E2E GST N-pair | **4.71** | **23.63** | **8.50** |
| E2E VAE N-pair | 4.82 | 23.70 | 9.10 |

**Table 2**. Subjective evaluation of End-to-End TTS system

| Model | MOS | Speaker MOS | Expressive MOS |
|---|---|---|---|
| RCVAE | $2.62 \pm 0.5$ | $2.40 \pm 0.3$ | $1.53 \pm 0.3$ |
| RCVAE N-pair | $2.97 \pm 0.4$ | $2.86 \pm 0.3$ | $1.93 \pm 0.2$ |
| IAF N-pair | $\mathbf{3.02 \pm 0.4}$ | $\mathbf{2.93 \pm 0.4}$ | $\mathbf{2.03 \pm 0.3}$ |
| E2E GST | $3.51 \pm 0.3$ | $2.57 \pm 0.2$ | $3.05 \pm 0.2$ |
| E2E VAE | $3.38 \pm 0.4$ | $2.71 \pm 0.3$ | $3.12 \pm 0.2$ |
| E2E GST N-pair | $\mathbf{3.72 \pm 0.4}$ | $2.65 \pm 0.2$ | $3.15 \pm 0.4$ |
| E2E VAE N-pair | $3.47 \pm 0.3$ | $\mathbf{2.83 \pm 0.3}$ | $\mathbf{3.33 \pm 0.3}$ |

to assign the score for synthesized speech utterances from a scale of 1 as bad to 5 as excellent, considering intelligibility, naturalness, and quality of speech utterance. Each listening test consists of 10 randomly selected (from the test set) speech utterances for each model. 14 French listeners participated in this MOS test and results are displayed in Table 2 with an associated 95% confidence interval.

The main goal of this work is to transfer the emotion as expressive attributes to the target speaker's voice without altering the speaker's voice characteristics. As there is no possible way to extract quantitative results for evaluation of transfer of expressivity without reference to expressive speech samples, we opt for speaker MOS and expressive MOS as a qualitative measure for expressivity transfer.

In speaker MOS, we instructed listeners to assign the score between 1 (bad) and 5 (excellent) based on speaker similarity between reference speaker speech and synthesized expressive speech. Likewise, for expressive MOS, listeners are directed to provide scores between 1 (bad) to 5 (excellent) depending on how synthesized speech utterance resembles the expressivity given in the reference speech utterance. A total of 14 French listeners performed both listening tests mentioned above, where each listener scored 18 speech utterances for each speaker-emotion pair and model. The results obtained through expressive MOS and speaker MOS are presented in Table 2 with associated 95% confidence intervals.

Apart from the presented E2E models, Table 2 also includes subjective scores obtained using the parametric multispeaker expressive TTS [10, 11], where RCVAE, RCVAE N-pair and IAF N-pair are parameteric TTS systems.

## 6. DISCUSSION

We investigated the transfer of expressivity without the explicit need of reference mel spectrogram given to expressivity encoder. From Table 2., the obtained MOS scores for each E2E model are consistent with the objective evaluation score in Table 1. The E2E GST N-pair model outperformed all models, thus usage of multihead attention assists in speech synthesis. The E2E GST N-pair model performance shows that the addition of N-pair to expressivity encoder boosts the model performance which can also be seen with the performance of VAE based expressivity encoder.

The speaker MOS score of the E2E VAE N-pair model is higher than with the other E2E models. The IAF N-pair based parametric TTS performed slightly better than the E2E VAE N-pair model for retaining speaker attributes. The IAF N-pair model uses x-vector based speaker embedding for creating speaker representation [11]. This results in better speaker representation than when only speaker identity is provided for creating speaker embeddings. The E2E VAE N-pair model obtains the highest expressive MOS score. This shows that the E2E VAE N-pair model can generalize a better emotion latent space using expressivity encoder than the E2E GST N-pair model. In parametric based approaches, expressivity transfer is conducted in acoustic space only, which lacks the interpolation in duration prediction. Thus, the E2E TTS system not only influences the prosody of synthesized speech but also the alignment of synthesized speech for each emotion.

## 7. CONCLUSION

We proposed to use multiclass N-pair loss on latent representation extracted using expressivity encoder to derive emotion as semantic information. The obtained results show that the performance of expressivity transfer is significantly improved "with" the addition of N-pair loss in comparison to "without" use of N-pair loss.

For transfer of expressivity VAE based expressivity encoder generalizes emotion representation better than GST. The results reported for speaker MOS show that providing speaker identity doesn't convey enough information for retaining speaker attributes. Therefore, in the future, we would like to improve the representation space of the speaker by providing a reference Mel spectrogram for estimating a speaker embedding.

## 8. ACKNOWLEDGEMENT

Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr).

## 9. REFERENCES


[1] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *ArXiv*, vol. abs/1703.10135, 2017.

[2] Wei-Ning Hsu, Y. Zhang, Ron J. Weiss, H. Zen, Y. Wu, Yuxuan Wang, Yuan Cao, Y. Jia, Z. Chen, Jonathan Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," *ArXiv*, vol. abs/1810.07217, 2019.

[3] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," 2018.

[4] R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *ArXiv*, pp. 4700–4709, 2018.

[5] 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," *ICASSP*, pp. 6945–6949, 2019.

[6] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Interspeech*. 2018, pp. 3067–3071, ISCA.

[7] Younggun Lee and Taesu Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," *ICASSP*, pp. 5911–5915, 2019.

[8] Vatsal Aggarwal, Marius Cotescu, Nishant Prateek, Jaime Lorenzo-Trueba, and Roberto Barra-Chicote, "Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech," *ICASSP*, pp. 6179–6183, 2020.

[9] Noé Tits, Fengna Wang, K. Haddad, V. Pagel, and T. Dutoit, "Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis," in *Interspeech*. 2018, ISCA.

[10] Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet, "Deep Variational Metric Learning For Transfer Of Expressivity In Multispeaker Text To Speech," in *SLSP*, 2020.

[11] Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet, "Transfer learning of the expressivity using FLOW metric learning in multispeaker text-to-speech synthesis," in *INTERSPEECH*, 2020.

[12] Kihyuk Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *NIPS*, 2016.

[13] Mahmut Kaya and Hasan Şakir Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, pp. 1066, 2019.

[14] Jonathan Shen, R. Pang, Ron J. Weiss, M. Schuster, Navdeep Jaitly, Z. Yang, Z. Chen, Yu Zhang, Yuxuan Wang, R. Skerry-Ryan, R. A. Saurous, Yannis Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *ICASSP*, pp. 4779–4783, 2018.

[15] Samuel R. Bowman, L. Vilnis, Oriol Vinyals, Andrew M. Dai, R. Józefowicz, and S. Bengio, "Generating sentences from a continuous space," in *CoNLL*, 2016.

[16] Xudong Lin, Yueqi Duan, Qiyuan Dong, Jiwen Lu, and Jie Zhou, "Deep variational metric learning," in *ECCV*, 2018.

[17] Junichi Yamagishi, Pierre-Edouard Honnet, Philip Neil Garner, and Alexandros Lazaridis, "The siwis french speech synthesis database," 2017.

[18] A. Sini, Damien Lolive, G. Vidal, M. Tahon, and Elisabeth Delais-Roussarie, "Synpaflex-corpus: An expressive french audiobooks corpus dedicated to expressive speech synthesis," in *LREC*, 2018.

[19] Sara Dahmani, Vincent Colotte, Valérian Girard, and Slim Ouni, "Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis," in *INTERSPEECH*, 2019.

[20] Ryan Prenger, Rafael Valle, and Bryan Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," *ICASSP*, pp. 3617–3621, 2019.

[21] Robert C. Streijl, Stefan Winkler, and David S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, pp. 213–227, 2014.