Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis

Ajinkya Kulkarni
Université de Lorraine
CNRS, Inria, LORIA
F-54000 Nancy, France
ajinkya.kulkarni@loria.fr

Vincent Colotte
Université de Lorraine
CNRS, Inria, LORIA
F-54000 Nancy, France
vincent.colotte@loria.fr

Denis Jouvet
Université de Lorraine
CNRS, Inria, LORIA
F-54000 Nancy, France
denis.jouvet@loria.fr

***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. The presented approach conditions Tacotron 2 speech synthesis with latent representations extracted from text, speaker identity, and reference expressive Mel spectrogram. We propose to use multiclass N-pair loss in the end-to-end multispeaker expressive Text-To-Speech (TTS) for improving the transfer of expressivity to the target speaker's voice. We have jointly trained the end-to-end (E2E) TTS with multiclass N-pair loss to discriminate between various emotions. This augmentation of the loss function during training paves the way to enhance the latent space representation of emotions.**

We have experimented with two different neural network architectures for expressivity in the 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 the training process improves expressivity in the desired speaker's voice.

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

I. Introduction

The term expressivity in speech usually refers to the characteristics of speech, such as emotions, speaking style, relationship of speech with gestures, and facial expression. Throughout this paper, we consider only the emotional characteristics of expressivity in speech. The current end-to-end TTS systems heavily relies on a large amount of speech corpus used for training the system [1]. Therefore, to build expressive speech synthesis for a new speaker, one has to create a speech corpus with various emotions. It is inconvenient to record an expressive speech corpus every time somebody wants 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. Though, many approaches proposed to use audio-books, films, dialogues. to create expressive speech synthesis. However, labeling the expressions is not a trivial task due to a large number of possible variations in a single emotion [22]. This creates a bottleneck in the development of expressive speech synthesis in a new speaker's voice.

There are numerous frameworks that have been proposed for the implementation of expressivity transfer either by interpolation of latent representations of prosody or of prosody embedding [2]–[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].

The systems mentioned above have shown significant performance in controlling expressivity, but many approaches use audiobook emotive storytelling style. Besides, they don't work with emotions such as joy, sadness, happiness, fear, anger. However, few approaches addressed the usage of multiple emotions [9]–[11], [19] in TTS. 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 authors proposed to build a recurrent conditional variational autoencoder based acoustic model and is trained using multiclass N-pair loss as additional loss function [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 mentioned approaches indicates that the addition of multiclass N-pair loss increases the perceived expressivity in the target speaker's voice. They have shown promising results for parametric systems but still, depending on the bottleneck step of the duration model for each emotion. In this paper, we present a novel metric learning framework [13] jointly trained with a Tacotron 2 based end-to-end TTS system [14]. This results in enhancing the latent representation of expressivity for better transfer learning performance. The expressivity information learned by the encoder influences the alignment of synthesized speech generated through attention.

The paper is organized as follows, Section II introduces the proposed architecture and the loss function used for training the end-to-end TTS. Section III describes the speech corpora and the processing of data before training. Section IV presents the experimentation setup. This is followed by Section V, presents the results obtained. Finally, Section VI and Section VII concludes the paper and presents a discussion.

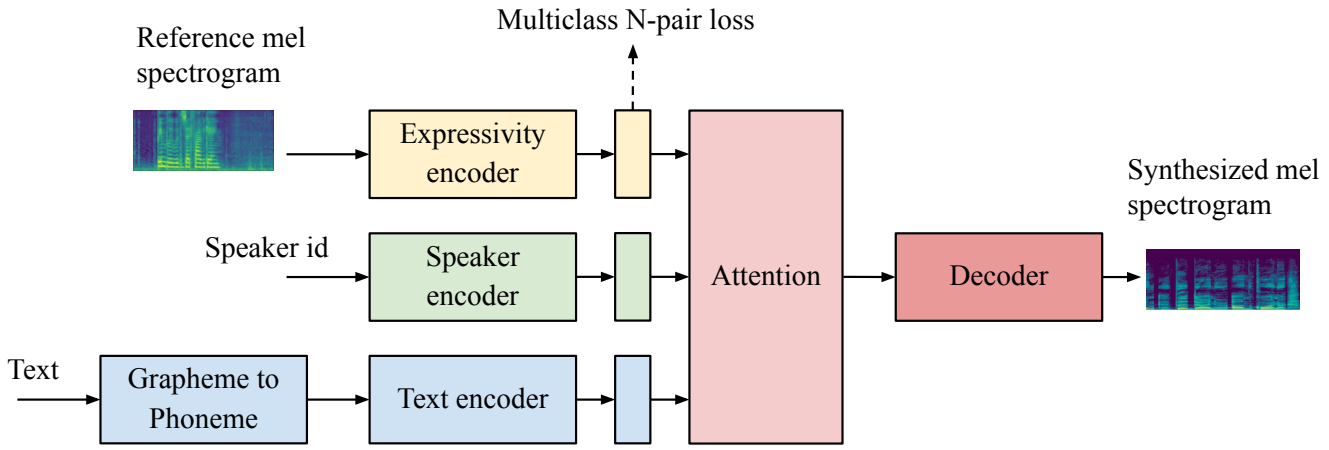

Fig. 1. End-to-end multispeaker expressive text to speech system

II. Proposed Architecture

The Tacotron 2 consists of a recurrent sequence-to-sequence feature prediction network that maps character embeddings to mel-scale spectrograms. This recurrent sequence-to-sequence replaces the convolutional block highway gates as proposed in Tacotron system [1]. A more detailed explanation of Tacotron 2 architecture is presented in [14].

We extend the state-of-the-art Tacotron 2 model [14] based on sequence to sequence with attention module to implement an end-to-end multispeaker expressive TTS system. In order to work with multiple speakers and expressivity, we modify the Tacotron 2 approach by adding an expressivity encoder and a speaker encoder.

A. General Framework

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

Afterwards, z_t, z_e, and z_s are concatenated and given as input to the location sensitive attention module, as illustrated in Fig. 1. This assists the end-to-end TTS to learn the alignment between the sequence of phonemes and the desired Mel spectrogram. The decoder network is composed of pre-net, BLSTM based recurrent network, and convolutional layer based post-net.

The decoder takes encoder outputs with attention vector as input to predict the Mel spectrogram frame by frame. The output from the previous frame is first passed through the pre-net network. The pre-net network consists of fully connected layers with the ReLU activation function. This predicted mel spectrogram from the pre-net and recurrent network is further passed through the post-net network. The post-net network composes of 5 layers of a convolutional network. The post-net improves the overall reconstruction performance of Mel spectrogram by predicting residual to add to the predicted Mel spectrogram.

B. Expressivity encoder

We have experimented with two neural network architectures for the implementation of expressivity encoder namely, Global Style Token (GST) [3] and Variational Autoencoder (VAE) [5]. The GST based expressivity encoder consists of a reference encoder, style attention, and style embedding. The reference encoder maps expressivity of variable length mel spectrogram into a fixed-length vector, which is passed to the style attention layer. This layer applies a multi-head 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 represents the expressivity as a stylistic factor to learn from the reference embedding.

The second architecture for the expressivity encoder is VAE based which is composed of a reference encoder and two feed-forward layers to generate mean and standard deviation of latent variable z_e. The reference encoder in VAE generates a hidden output which is passed through the feed-forward layers to obtain latent variable z_e. The z_e is obtained by using a reparameterization trick applied with mean and standard deviation. VAE based framework suffers from Kublack Leibler (KL) annealing problem in which reconstruction loss is suppressed by KL loss term [15]. In KL annealing problem, after few initial epochs, KL divergence term suddenly goes close to zero.

To avoid this, additional weight (close to zero) is multiplied to KL loss and gradually increased over the training epoch.

C. Multiclass N-pair loss

The end-to-end TTS is jointly trained with the multiclass N-pair loss function. This assists in predicting Mel spectrogram output in desired emotions. For the transfer of expressivity, we have used pre-computed means of latent variables of 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 emotions may lead to the poor transfer of expressivity. And for better performance of expressivity transfer, we need the tightly bounded representation of the latent variables of emotions. Therefore, we propose a novel deep metric learning framework implemented using multiclass N-pair loss to further enhance the expressivity representation.

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 classes and increasing the distance between each negative class. The multiclass N-pair loss has shown better performance than triplet loss or contrastive loss by considering embeddings of multiple negative classes [12]. In the training phase, the model needs to reduce the multiclass N-pair loss function as stated in Eq.1, in addition to the reconstruction loss and attention loss.

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

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

III. Data preparation

We have used 4 French Female speech synthesis corpora for implementing an 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 have used only the neutral voice of Synpaflex in our work.

We have 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 have 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 the end-to-end TTS, we have 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.

IV. Experimental setup

For training the end-to-end TTS system, we have 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 have used a 128 dimensional latent variable of expressivity for both GST and VAE. Before 150K training steps, we have applied the weight of 0.0001 in every 200 steps in order to reduce the KL annealing effect. This weight is increased by 0.00001 after every 500 steps.We have adopted a similar technique for fine-tuning with multiclass N-pair loss, for which until 150K training steps a weight of 0 is applied on the multiclass N-pair loss, and afterwards the weight is increased by 0.001 after every 200 steps.

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

V. Results

In this section, Table I and Table II includes proposed end-to-end TTS (E2E) models, along with evaluation scores obtained using the parametric multispeaker expressive TTS as stated in [10], [11]. We have presented the evaluation scores for parametric TTS as recurrent conditional variational autoencoder (RCVAE), RCVAE N-pair, and inverse autoregressive flow (IAF) N-pair.

A. Objective evaluation

We have 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 E2E TTS systems. The objective evaluation results are presented in Table I. From Table I E2E

TABLE I
OBJECTIVE EVALUATION FOR AN 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

GST N-pair model has shown superior performance compared to other models.

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.

B. Subjective evaluation

At first, we have evaluated an end-to-end (E2E) multi-speaker expressive TTS systems using Mean Opinion Score (MOS) [21] based listening test. In this work, we used the absolute category ranking scale. Each listener had to assign a score for synthesized speech utterance between scale 1 to 5 considering the intelligibility, naturalness, and quality of speech utterance. Suppose the speech quality is bad the listener will then assign the score 1 and if the speech quality is excellent then the listener will assign the score 5. Each listening test consists of 10 randomly selected speech files from the test set for each model. A total of 14 French listeners participated in this MOS test and results are displayed in Table II 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.

Similarly, for speaker MOS as well, we instructed the listeners to assign the score between 1 (bad) and 5 (excellent) to the speech samples based on the speaker similarity between reference speaker speech and synthesized expressive speech. Likewise, for expressive MOS, listeners are directed to provide scores between 1 (bad) and 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 II with associated 95% confidence intervals.

VI. DISCUSSION

From Table II, the obtained MOS scores for each E2E model are consistent with the objective evaluation score in Table I. The E2E GST N-pair model outperformed all models, thus usage of multi-head 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

TABLE II
SUBJECTIVE EVALUATION OF AN 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	**3.02 $\pm$ 0.4**	**2.93 $\pm$ 0.4**	**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	**3.72 $\pm$ 0.4**	2.65 $\pm$ 0.2	3.15 $\pm$ 0.4
E2E VAE N-pair	3.47 $\pm$ 0.3	**2.83 $\pm$ 0.3**	**3.33 $\pm$ 0.3**

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 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 an expressivity encoder than the E2E GST N-pair model.

The proposed E2E models learn the duration information as alignment derived from attention vector as opposed to parametric TTS, where explicit duration model is required. From Table II, MOS scores for parametric TTS range between 2 and 3, while E2E models have greater than 3. The parametric TTS systems conduct expressivity transfer in acoustic space only, which lacks the interpolation in duration prediction. Thus, E2E TTS system not only influence the prosody of synthesized speech but also the alignment of synthesized speech for each emotion.

VII. CONCLUSION

We proposed to use multiclass N-pair loss on latent representation extracted using expressivity encoder to derive emotion as latent information. During the inference phase, we investigated the transfer of expressivity without the explicit need of reference Mel spectrogram. We used pre-computed means of latent variables of each emotion for expressivity transfer. 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. To our knowledge, the presented work is the first approach to use metric learning in an end-to-end multispeaker TTS system.

VIII. 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).

REFERENCES

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. 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, JonathanShen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis", ArXiv, vol. abs/1810.07217, 2019.

[3] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis", ArXiv, vol. abs/1803.09017, 2018.

[4] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron", ArXiv, vol. abs/1803.09047, 2018.

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

[6] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder", INTERSPEECH, pp.3067–3071, 2018.

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

[8] V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, "Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech", ICASSP, pp.6179–6183, 2020.

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

[10] A. Kulkarni, V. Colotte, and D. Jouvet, "Deep Variational Metric Learning For Transfer Of Expressivity In Multispeaker Text To Speech", SLSP, 2020.

[11] A. Kulkarni, V. Colotte, and D. Jouvet, "Transfer learning of the expressivity using FLOW metric learning in multispeaker text-to-speech synthesis", INTERSPEECH, 2020.

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

[13] M. Kaya and H. Bilge, "Deep metric learning: A survey", Symmetry, vol. 11, pp. 1066, 2019.

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

[15] S. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space", CoNLL, 2016.

[16] X. Lin, Y. Duan, Q. Dong, J. Lu, and J. Zhou, "Deep variational metric learning", ECCV,2018.

[17] J. Yamagishi, P. Honnet, P. Garner, and A. Lazaridis, "The siwis French speech synthesis database", 2017.

[18] A. Sini, D. Lolive, G. Vidal, M. Tahon, and E. Delais-Roussarie, "Synpaflex-corpus: An expressive French audiobooks corpus dedicated to expressive speech synthesis", LREC, 2018.

[19] S. Dahmani, V. Colotte, V. Girard, and S. Ouni, "Conditional variational autoencoder for text driven expressive audiovisual speech synthesis", INTERSPEECH, 2019.

[20] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow based generative network for speech synthesis", ICASSP, pp. 3617–3621, 2019.

[21] R. Streijl, S. Winkler, and D. Hands, "Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives", Multimedia Systems, pp. 213–227, 2014.

[22] M. Charfuelan and I. Steiner, "Expressive speech synthesis in MARY TTS using audiobook data and emotion", INTERSPEECH, 2013.