Deep variational metric learning for transfer of expressivity in multispeaker text to Speech

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Deep variational metric learning for transfer of expressivity in multispeaker text to speech

Ajinkya Kulkarni, Vincent Colotte, and Denis Jouvet

Université de Lorraine,CNRS, Inria, LORIA, F-54000 Nancy, France
{firstname.lastname}@loria.fr

Abstract. In this paper, we propose to use the deep metric learning based multi-class N-pair loss, for text-to-speech (TTS) synthesis. We use the proposed loss function in a recurrent conditional variational autoencoder (RCVAE) for transferring expressivity in a French multispeaker TTS system. We extracted the speaker embeddings from the x-vector based speaker recognition model trained on speech data from many speakers to represent the speaker identity. We use mean of the latent variables to transfer expressivity for each emotion to generate expressive speech in the desired speaker's voice. In contrast to the commonly used loss functions such as triplet loss or contrastive loss, multi-class N-pair loss considers all the negative examples which make each class of emotion distinguished from one another. Furthermore, the presented approach assists in creating a robust representation of expressivity irrespective of speaker identities. Our proposed approach demonstrates the improved performance for transfer of expressivity in the target speaker's voice in a synthesized speech. To our knowledge, it is for the first time multi-class N-pair loss and x-vector based speaker embeddings are used in a TTS system.

Keywords: text-to-speech · variational autoencoder · deep metric learning · expressivity.

1 Introduction

Text-to-speech synthesis is basically the artificial production of human speech from text. The traditional formulation of a text-to-speech (TTS) system often leaves the expressivity contained in a text. Expressive speech synthesis aims at generating synthesized speech by adding expressivity such as emotion, speaking style, etc reflecting the complex emotional states from a textual sentence. To make the artificially produced speech more realistic, a system should be able to impute certain linguistic factors such as intonation, rhythm, stress, etc usually termed as prosody. In this paper, we have presented generating synthesized emotional speech for multiple speakers. We have considered six different emotions as expressivity to transfer into a multispeaker TTS system.

Deep neural network based expressive speech synthesis has made comparatively progressive improvement in their performances in the recent times [6,9,11].

Several approaches have been proposed to transfer expressivity either by controlling the prosody parameters in latent space for speech synthesis or by transferring the expressivity using interpolation of conditional embeddings of speaker identity and prosody embedding [5,6]. The work done by [11], proposed expressivity transplantation as an extension to speaker adaptation using Latent Hidden Unit Contribution (LHUC) units. In [1,4], the variational autoencoder (VAE) framework has been adapted within the end-to-end TTS systems such as voiceLoop, and tacotron. They transform parameterized speech into latent representation and disentangle the latent speech attributes such as prosody, and speaker. But, these approaches are limited to single speaker text-to-speech system. A limited number of work has focused on integrating expressiveness into the multispeaker TTS system [5,6]. In [7], for synthesizing clean speech with controllable speaking style the authors have used a two level conditional generative model based on variational autoencoder. Most of the works uses 'global style token' or GST which is basically a style embedding to learn the expressiveness for multispeaker TTS system [8]. They created the style token embedding considering variation in prosody as well as speaking style except emotion.

Although, emotion is an essential feature in human-computer interface, not much work has done on synthesizing emotional speech using different types of emotion such as joy, anger, etc. It is difficult to effectively model synthetic emotional speech using different emotions. Besides, there is unavailability of emotional corpora featuring the different types of emotion. Moreover, it is time consuming annotating and collecting a huge dataset of emotional speech is not feasible. In this paper, we propose multi-class N-pair loss [15], a novel loss function to transfer different emotions into a multispeaker expressive TTS system for French. Deep metric learning has gained wide recognition in the computer vision and image classification domain [14]. They have used it mainly for training discriminative models. Whereas, we exploit the idea of using N-pair loss for generative model. The proposed loss function is used in a recurrent conditional variational autoencoder (RCVAE) to produce speech with multiple emotions. The multi class N-pair loss is learning objective function of deep metric learning. Our proposed approach of using N-pair loss assists in creating a robust representation of emotion in latent space irrespective of speaker identities.

Additionally, to enable multispeaker setting in TTS, we use speaker embeddings as an explicit condition in RCVAE framework. We derived the speaker embedding from speaker encoder pretrained with x-vectors. The x-vector maps the variable-length utterances to text independent fixed dimensional embeddings which are trained using a deep neural network that discriminates between speakers [16].The speaker embeddings corresponds to the activations of the last hidden layer of speaker encoder network.

The rest of the paper is organized as follows: Section 2 presents the multispeaker expressive TTS approach which relies on RCVAE architecture and deep metric learning. In Section 2.1, we present the implementation of the acoustic model by a RCVAE architecture trained using the multi-class N-pair loss metric learning. Section 2.2 discusses the speaker embeddings created for French

speakers using a pre-trained speaker recognition model. Section 4.1 describes the pre-processing of speech and text data for the training of the multispeaker expressive TTS system. Section 4.2 describes the experimentation set up and Section 5 demonstrates the results obtained using the RCVAE model with and without using multi-class N-pair loss to show its impact on the transfer of expressivity. Section 6 presents the discussion on the obtained results and conclusion is presented in Section 7.

2 Multispeaker expressive TTS

We build our TTS system using parametric speech synthesis approach, which is divided into duration model and acoustic model. In this section, we present the implementation of the acoustic model with multi-class N-pair loss in a RCVAE architecture as it is described in Section 2.1. We used an explicit duration model to predict the number of acoustic feature frames required to synthesize the speech for a given text. Hence, for predicting duration for each phoneme, we used a bidirectional long short term memory (BLSTM) based neural network as explained in [27]. Section 2.2 illustrates the implementation of speaker embedding as a adaptation technique from pre-trained speaker recognition.

2.1 Model architecture

Variational autoencoders were introduced in 2013 by Kingma and Rezende independently [17, 18]. The main components of a variational autoencoder are an encoder, a decoder, and the loss function used for training. For the RCVAE architecture, we implemented a BLSTM based encoder network. The input of the encoder is a sequence of acoustic features, x, along with condition c. Here, the condition c corresponds to textual features, duration information, and speaker embedding. The activation of hidden states of BLSTM layer is given to feedforward layers to estimate both mean vector and variance vector. The mean and variance are further used to describe the encoder's latent variable, z. Similarly, the decoder network consists of BLSTM layers. The usage of BLSTM based recurrency allows the model to extract long term context from acoustic features. The input of the decoder network are a latent variable z and the condition c. The decoder generates the sequence of predicted acoustic features $\hat{x}$, as shown in Fig 1. During the inference, we sample z from the latent space distribution.

The loss function in VAE corresponds to the reconstruction loss plus a regularization term defined with the Kullback-Leibler (KL) divergence. The reconstruction loss represents the expectation over the reconstruction of acoustic features, $logP(x|z,c)$. The KL divergence measure indicates how close the learned distribution $Q(z|x,c)$ is to the true prior distribution $P(z|c)$. Recurrent network based VAE frameworks often leads to sudden drop in KL divergence [19]. To deal with this problem, we added variable weight, λ to KL divergence term as a KL annealing cost, as is is mentioned in Eq. 1. This assists to enhance the

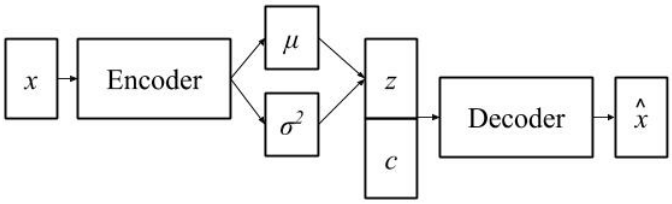

Fig. 1. RCVAE architecture used for training acoustic model. Here, x is a sequence of acoustic features to be reconstructed as $\hat{x}$, c is condition (textual features, speaker embedding) μ and σ are mean and variance parameters provided by the encoder network, and used to generate the latent variable z.

disentangled latent space representation with good interpretability of the latent variable.

$$\begin{aligned} Loss &= E_z[logP(x|z,c)] + \lambda KL[Q(z|x,c)\|P(z|c)] \\ &+ \quad log(1 + \sum_{i=1}^{N-1} exp(z^\top z_i^- - z^\top z^+)) \end{aligned} \tag{1}$$

We use mean of latent variables as representation of emotion for expressivity transfer. Hence, the desired latent space should have well separated cluster's corresponding to the various emotions. This indicates better clustering of emotion may lead to improved performance of expressivity transfer in TTS system. Therefore, we proposed to use multi-class N-pair loss in variational inference as deep variational metric learning. Multi-class N-pair loss has shown superior performance compared to triplet loss or contrastive loss by considering one positive sample and $N-1$ negative samples for N classes [15]. This loss criteria increases the intercluster distance from $N-1$ negative samples and decreases the intracluster distance between positive samples and training examples. We employed mean of latent variables of emotion for mining the positive and the negative samples. In our case, positive samples refer to latent variables from the same emotion class and negative samples correspond to examples of different emotion classes. For N classes, z^+ is a positive sample, and $\{z_i^-\}_1^{N-1}$ samples from negative classes as stated in Eq. 1. This usage of multiple negative samples in training leads to faster convergence of the model creating a robust representation of emotion.

The RCVAE acoustic model's t-distributed stochastic neighbor embedding (t-SNE) plot shows the overlap of clusters of emotions, as illustrated in Fig 2. We used the mean of latent variables of emotions to transfer the expressivity. If latent space has an unclustered representation of emotion, it may lead to poor transfer of expressivity. The t-SNE plot of the RCVAE N-pair acoustic model shows well-clustered emotion in latent space. The orange cluster in the t-SNE plot represents neutral speech. It is undesirable to synthesize expressive speech with

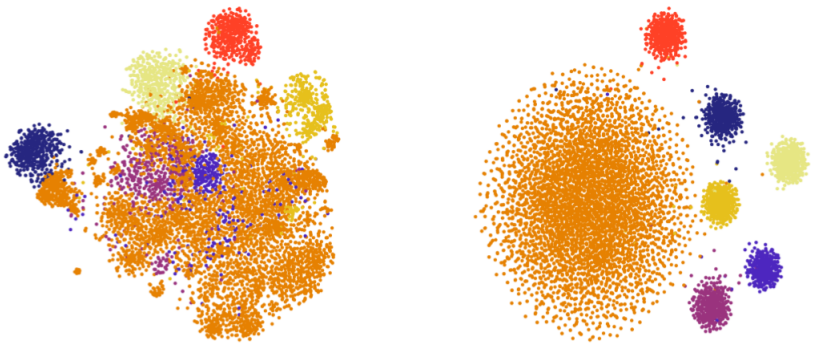

Fig. 2. t-SNE plot of latent representation of RCVAE acoustic model (left side) and RCVAE acoustic model with N-pair loss (right side). Each color represents the emotion.

modification in the target speaker's voice. This clustering of neutral speech for multiple speakers reflects the improvement in preserving the speaker's identity while transferring the expressivity. We build a RCVAE acoustic model without N-pair loss as a baseline system to evaluate the improvement in expressivity using deep metric learning.

2.2 Speaker embedding

The RCVAE encoder-decoder network is explicitly conditioned on the speaker embedding. We created speaker embeddings from pretrained speaker recognition model to capture the speaker's information. These embeddings should represent speaker characteristics irrespective of the textual content. For generating such embeddings, we develop a speaker encoder network from speaker recognition model trained on French speech synthesis corpora. Later, we use this speaker encoder to derive the speaker embedding.

To derive the speaker embeddings, we used x-vectors to train a feedforward neural network based speaker recognition model for discriminating between the speakers of our French speech synthesis corpora. The x-vector are deep neural network based embeddings trained on time-delay neural networks with a statistical pooling layer trained for the speaker recognition task [16]. We extracted x-vectors from the pretrained speaker recognition model trained on the voxceleb corpus available in the Kaldi tool [20, 21]. Finally, we obtained the speaker embeddings as an output of the last hidden layer of feedforward neural networks in the French speaker recognition model. The separation of speaker encoder and RCVAE framework results in lowering the complexity of network as well as requirement for multispeaker training data.

3 Experimentation

3.1 Data preparation

We used 4 speech corpora, namely Lisa [12], a French female neutral corpus (approx. 3 hrs), Caroline [26], a French female expressive corpus (approx. 9hrs), Siwis [23], a French female neutral corpus (approx. 3 hrs), and Tundra [24], a French male neutral corpus (approx. 2hrs). Caroline's expressive speech corpus consists of several emotions, namely joy, surprise, fear, anger, sadness, and disgust (approx. 1hr for each emotion and 3hrs for neutral). For each emotion, there are approximately 500 utterances for a total of 1hr duration. All the speech signals were used at a sampling rate of 16 kHz. Each speech corpus is divided into train, validation, and test sets in the ratio of 80%, 10%, 10% respectively.

We parameterized speech using the WORLD vocoder [22] with 187 acoustic features computed every 5 milliseconds, namely 180 spectral features as Mel generalized cepstrum coefficients (mgc), 3 log fundamental frequencies (lf0), 3 band-aperiodicities (bap) and 1 value for voiced-unvoiced information (vuv). Based on the mean and standard deviation values, the acoustic features extracted from the WORLD vocoder were z-normalized. We used the front-end text processor from SOJA-TTS (developed internally in our team) for converting French text to linguistic features also known as context labels (dimension 180) which include pentaphone information.

3.2 Experimentation setup

The RCVAE architecture consists of 2 BLSTM layers of 256 hidden units for both encoder network and decoder, The latent variable is of dimension 50. The training is done using a learning rate of 0.0001. The Adam optimizer initialized with default parameters, a batch size of 10 and a lambda factor of 0.001. The model was trained until the 100^{th} epoch. To ensure better convergence of model parameters, the multi-class N-pair loss was activated only after the first 5 epochs. In the training phase, we used precomputed means of latent variables for each emotion from the previous epoch. These precomputed means are used in multi-class N-pair loss as positive and negative samples. For the baseline model, we trained the RCVAE acoustic model without multi-class N-pair loss with the same configuration as described above.

In the inference phase, we used the mean of latent variables computed for each emotion as a latent variable to synthesize each particular emotion. As mentioned before, we implemented a duration model explicitly for each speaker using a BLSTM network of 512 hidden units with the same configuration of batch size, learning rate, and optimizer as for the RCVAE architecture.

For speaker embeddings, for all speech samples in corpora, we extracted 512-dimensional x-vector using the speaker recognition model trained on the voxceleb corpus [20]. Then, we implemented a 5 layer of feedforward neural network, and trained it to classify 4 French speakers (corresponding to our speech synthesis corpora) with (512-256-128-64-16) hidden units, using cross-entropy loss criteria,

Adam optimizer, and 50 epochs of training. We extracted speaker embedding for each speech sample by taking the output of activations of the last hidden layer of dimension 16.

4 Results

We first computed Mel cepstrum distortion (MCD) on test data between reference acoustic features and those generated by acoustic models. The obtained results are presented in Table 1. One of the challenge we encountered was the fact that there is no reference emotional acoustic features available for Lisa, Tundra, and Siwis. Therefore, we evaluated the performance of transfer of expressivity using a subjective evaluation.

Table 1. Objective evaluation using MCD results

Model	**MCD**
RCVAE	5.795
RCVAE+N-pair	**5.472**

4.1 Evaluation of multispeaker TTS

We carried out a Mean opinion score (MOS) [25] perception test for evaluating our multispeaker text-to-speech synthesis system. For the perception test, each listener had to score the synthesized speech stimuli from 1 to 5, where 1 is bad and 5 is excellent, considering intelligibility, naturalness, and quality of the speech stimuli. 12 French listeners participated in the perception test; each listener had to score 5 stimuli for each speaker-emotion pair randomly chosen from the test set. The results of the test are shown in Table 2 with an associated 95% confidence interval. The presented score for Caroline speaker in Table 2 represents the average score obtained for Caroline's neutral voice and all Caroline emotions (with associated confidence interval). The scores for all others speakers have comparably similar results, in which Lisa speaker received the highest score for both the models trained with and without deep metric learning. Due to limited training data (1hr) for each emotion for Caroline's voice, performance of Caroline's speech synthesis for emotion is lower compared to Caroline's neutral speech synthesis. The results presented in Table 2 show that deep variational learning approach leads to better results compared to RCVAE acoustic model without N-pair loss. This is in line with the better separation between emotions in the latent space, as observed from the t-SNE plots in Fig 2.

4.2 Evaluation of transfer of expressivity

We used speaker similarity score and expressive similarity to evaluate the performance of the proposed architecture transferring expressivity onto other speaker

Table 2. MOS score for evaluation of multispeaker TTS system

MOS	***Caroline neutral***	***Caroline emotion***	***Lisa neutral***	***Siwis neutral***	***Tundra neutral***
RCVAE	2.7 ± 0.4	2.1 ± 0.2	2.8 ± 0.7	2.6 ± 0.8	2.7 ± 0.2
RCVAE+N-pair	**3.2 ± 0.4**	**2.6 ± 0.2**	**3.1 ± 0.6**	**3.0 ± 0.5**	**2.9 ± 0.4**

Table 3. Speaker similarity scores when generating sentences with transfer of expressivity

Speaker similarity	**Lisa**	**Siwis**	**Tundra**
RCVAE	2.3 ± 0.2	2.2 ± 0.1	2.7 ± 0.3
RCVAE+N-pair	**3.0 ± 0.1**	**2.7 ± 0.3**	**2.9 ± 0.2**

Table 4. Expressive similarity scores when transferring expressivity

Expressive similarity	**Lisa**	**Siwis**	**Tundra**
RCVAE	1.4 ± 0.4	1.5 ± 0.3	1.7 ± 0.5
RCVAE+N-pair	**1.9 ± 0.3**	**1.9 ± 0.4**	**2.0 ± 0.2**

voices. The linguistic contents of the speech stimuli and reference stimuli are not the same during the evaluation. In the speaker similarity perception test, we instructed listeners to provide a score about the similarity between the original speaker speech stimuli and synthesized expressive speech in a range of 1 (bad speaker similarity) to 5 (excellent speaker similarity). Likewise, we also directed listeners to score expressivity observed in the synthesized expressive speech stimuli on a scale of 1 (bad similarity) to 5 (excellent similarity) depending on the closeness of expressive characteristics in speech stimuli compared to original expressive speech stimuli. 12 French listeners participated in a perception test, each listener scored 3 sets of stimuli for each target speaker-emotion pair. The results of expressive similarity and speaker similarity are shown in Tables 3, 4, with associated 95% confidence interval. Figs 3, 4 display respectively speaker similarity and expressive similarity scores for each emotion and each speaker. Figs 3, 4 show that similar results for all emotion are observed for the three speaker's voices, Siwis speaker got slightly lower score compared to other speakers.

The obtained results show that the addition of deep metric learning (multi-class N-pair loss) certainly improves the representation of expressivity, which leads to better transfer of expressivity. Furthermore, speaker similarity showed that while transferring expressive knowledge, addition of N-pair loss to architecture improve retainment of the speaker characteristics. Also results from Table 3 show that the system is able to equally transfer the expressivity not only from female (Caroline) to female (Lisa, Siwis) speakers but also from female (Caroline) to male (Tundra) speaker. Our proposed approach shows better performance for Lisa speaker than previous layer adaptation [12] approach for both speaker sim-

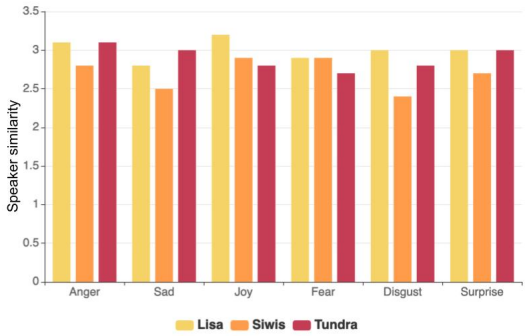

Fig. 3. Speaker similarity scores per emotion and speaker's voice, using RCVAE model trained with N-pair loss

ilarity and expressive similarity. From Figs 3 and 4, Tundra speaker shows that sad and surprise are the emotions perceived as close to expressive characteristics with respect to the original reference speech provided in evaluation. While anger is the least perceive emotion for all speakers. In Fig 3., we can observe that transferring anger emotion to target speakers received higher speaker similarity scores.

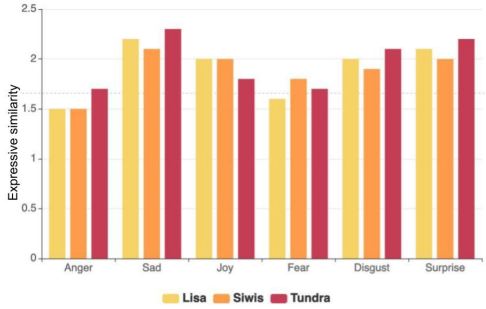

Fig. 4. Expressive similarity scores per emotion and speaker's voice, using the RCVAE model trained with N-pair loss

5 Discussion

We investigated the transfer of expressivity considering the emotional aspect of speech. The Caroline expressive speech corpus was recorded with several emotions irrespective of emotional information derived from the textual content.

The available training expressive speech data is limited to 1hr per emotion. This poses a challenge in training complex deep neural network frameworks for which large training data is usually expected. Due limited availability of training data for expressive speech corpus, we opt for parametric speech synthesis framework. The current state-of-art TTS frameworks focus mainly on transferring prosody or speaking style. These frameworks use precomputed means of emotions to perform the interpolation. In our approach, multiclass N-pair loss reduces the distance between latent variables belonging to the same emotion class. This creates a tightly bounded representation of expressivity. For instance, we can notice that trajectory of lf0 in Fig 5.a is certainly modified after transferring anger emotion as shown in Fig 5.b. The contour of lf0 for RCVAE acoustic model trained with N-pair loss (blue) has higher local differences (between high and low lf0 values) than the RCVAE acoustic model trained without N-pair loss (orange). This contributes to explain the better expressivity score obtained in contrast to the acoustic model without N-pair loss. We activated the N-pair loss objective after few training epochs to have a warm start for the network. Also, this ensures that reconstruction loss is converging in the right direction first. This avoids the overfitting of the clustering of latent variables. The variational metric learning provides benefits of variational inference along with a robust representation of emotion.

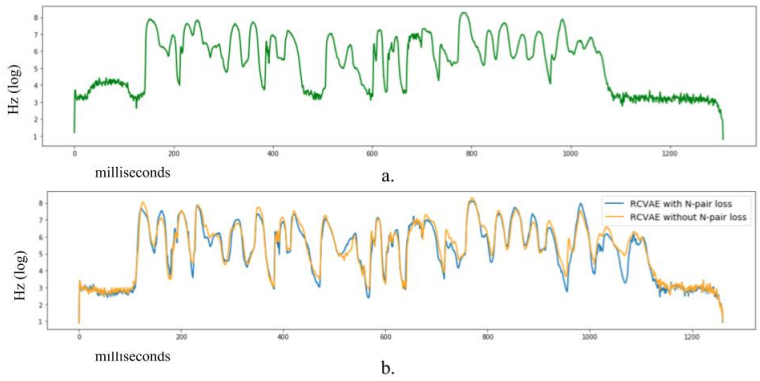

Fig. 5. lf0 trajectories for same utterance generated for a. neutral emotion for Lisa speaker, b. anger emotion for Lisa speaker synthesized using RCVAE without N-pair loss (blue) and using RCVAE with N-pair loss (orange).

6 Conclusion

We presented variational autoencoder architecture trained with multi-class N-pair loss for transferring expressivity in a multispeaker text-to-speech synthesis for French. Multi-class N-pair loss function is used for the disentanglement of

information in the latent space which is a deep metric learning objective. The deep variational metric learning enforce the better clustering of emotions in latent space representation.

In the presented work, speaker embeddings allow inheriting knowledge from the speaker recognition task in the TTS system. We trained speaker encoder network on speakers from our French speech synthesis corpora. The speaker representation learned in such a way ease the convergence of multispeaker TTS system. The perception tests conducted show that the proposed approach retains the target speaker voice while transferring the expressivity. This is the first approach that uses deep metric learning in a variational inference to improve the performance of latent space representation in transferring the expressivity. In the future, we would like to adopt a similar RCVAE based deep variational metric learning in an end-to-end TTS system.

7 Acknowledgements

Experiments presented in this paper were carried out using the Grid5000 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. Yuxuan Wang and R. J. Skerry-Ryan and Daisy Stanton and Yonghui Wu and Ron J. Weiss and Navdeep Jaitly and Zongheng Yang and Ying Xiao and Zhifeng Chen and Samy Bengio and Quoc V. Le and Yannis Agiomyrgiannakis and Rob Clark and Rif A. Saurous. Tacotron: A Fully End-to-End Text-To-Speech Synthesis Model. Journal: CoRR, arxiv.org, volume: abs/1703.10135, 2017.
2. Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Omer Arik, Ajay Kannan Sharan Narang, Jonathan Raiman and John Miller. Deep Voice 3: 2000-Speaker Neural Text-to-Speech, CoRR, arxiv.org, volume abs/1710.07654, 2017.
3. Sotelo J., Mehri S., Kumar K., Santos J.F., Kastner K., Courville A., Bengio Y.: Char2Wav: End-to-End Speech Synthesis. ICLR, 2017.
4. Taigman, Y., Wolf, L., Polyak, A., Nachmani, E. VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop. ICLR, 2017.
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, 2018.
6. Akuzawa K., Yusuke, I., and Yutaka, M. Expressive Speech Synthesis via Modeling Expressions with Variational Autoencoder. Interspeech, 2018.
7. Hsu W. N., Zhang Y., R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang. Hierarchical Generative Modeling for Controllable Speech Synthesis. ICLR, 2019.
8. Wang Y., Stanton D., Zhang Y., Skerry-Ryan R. J., Battenberg E., Shor J., Xiao Y., Jia Y., Ren F., Saurous R. Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. ICML, 2018.

9. Skerry-Ryan, R. J., 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. ICML, 2018.
10. Lee, Younggun and Taesu Kim. Robust and Fine-grained Prosody Control of End-to-end Speech Synthesis. ICASSP, 2019.
11. Parker, J., Stylianou, Y., Cipolla, R. Adaptation of an Expressive Single Speaker Deep Neural Network Speech Synthesis System. ICASSP, 2018.
12. Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet. Layer adaptation for transfer of expressivity in speech synthesis. Language & Technology Conference (LTC), 2019.
13. Lin, Xudong, Duan, Yueqi, Dong, Qiyuan, Lu, Jiwen and Zhou, Jie. Deep Variational Metric Learning.The European Conference on Computer Vision, 2018.
14. Kaya, Mahmut and BİLGE, Hasan Şakir, Deep Metric Learning: A Survey. Symmetry, volume 11 ISSN, 2073-8994, 2019.
15. Sohn, Kihyuk. Improved Deep Metric Learning with Multi-class N-pair Loss Objective., NIPS 2016.
16. Snyder, David, Daniel Garcia-Romero, Gregory Sell, Daniel Povey and Sanjeev Khudanpur. X-Vectors: Robust DNN Embeddings for Speaker Recognition. ICASSP, 2018.
17. Kingma, Diederik P. and Max Welling. Auto-Encoding Variational Bayes. CoRR, arxiv.org, abs/1312.6114 2013.
18. Rezende, Danilo Jimenez, Shakir Mohamed and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. ICML, 2014.
19. Bowman S. R., L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, Generating Sentences from a Continuous Space. SIGNLL Conference on Computational Natural Language Learning, 2016.
20. Chung J.S., Nagrani A., Zisserman A. VoxCeleb2: Deep Speaker Recognition. Interspeech 2018.
21. Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukás Burget, Ondrej Glembek, Nagendra Kumar Goel, Mirko Hannemann, Petr Motlícek, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer and Karel Veselý. The Kaldi Speech Recognition Toolkit. ASRU conference, 2011.
22. Morise M., Yokomori F., Ozawa K. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. IEICE Transactions 2016.
23. Yamagishi, Junichi, Pierre-Edouard Honnet, Philip Neil Garner and Alexandros Lazaridis. The SIWIS French Speech Synthesis Database. 2017.
24. Stan, Adriana, Oliver Watts, Yoshitaka Mamiya, Mircea Giurgiu, Robert A. J. Clark, Junichi Yamagishi and Simon King. TUNDRA: a multilingual corpus of found data for TTS research created with light supervision. Interspeech, 2013.
25. Streijl, Robert, C., Winkler, S., Hands, D. S. Mean Opinion Score (MOS) Revisited: Methods and Applications, Limitations and Alternatives. Multimedia System. Volume 22.2, 2016.
26. Dahmani S., Colotte V., Girard V. and Ouni S. Conditional Variational Auto-Encoder for Text-Driven Expressive AudioVisual Speech Synthesis. Interspeech 2019.
27. Zhizheng Wu, Oliver Watts, Simon King. Merlin: An Open Source Neural Network Speech Synthesis System. ISCA Speech Synthesis Workshop (SSW9), 2016.