Multi-stage attention for fine-grained expressivity transfer in multispeaker text-to-speech system

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Multi-stage attention for fine-grained expressivity transfer in multispeaker text-to-speech system

1st Ajinkya Kulkarni
Université de Lorraine, CNRS, Inria, Loria
Nancy, France
ajinkya.kulkarni@loria.fr

2nd Vincent Colotte
Université de Lorraine, CNRS, Inria, Loria
Nancy, France
vincent.colotte@loria.fr

3rd Denis Jouvet
Université de Lorraine, CNRS, Inria, Loria
Nancy, France
denis.jouvet@loria.fr

***Abstract*—The main goal of this work is to provide fine-grained transfer of expressivity in various speaker's voices for which no expressive speech data is available. Our approach conditions a multispeaker Tacotron 2 system with latent embeddings extracted from phoneme sequence, speaker identity, and reference expressive Mel spectrogram. The proposed system utilizes attention modules for discovering local and global expressivity attributes. Additionally, location-sensitive attention is applied in the decoder to learn the alignment between phoneme sequence-Mel spectrogram pair.**

In addition to conventional objective metrics for speech synthesis, we used cosine similarity and character error rate (CER) measures for the evaluation of transfer of expressivity and intelligibility. The obtained results demonstrate the presented cosine similarity metric for speaker and expressivity is consistent with the subjective evaluation. Thus, the usage of multiple evaluation measures provides a way to estimate the strength of emotions and the speaker's voice for transferred expressivity in the target speaker's voice. The obtained results show that presented fine-grained TTS systems performed better than the Tacotron 2 based baseline systems.

***Index Terms*—expressivity, transfer learning, text-to-speech**

I. Introduction

The main objective of the proposed work is to transfer the expressive attributes to synthesize speech without explicit recordings of expressive speech for the target speaker's voice. Throughout this paper, we consider only the emotional attributes of expressivity in speech. Abundant research has been conducted for expressivity transfer in the context of the end-to-end (E2E) text-to-speech (TTS) system [1]. Besides expressivity transfer, many approaches have been proposed in the context of style transfer and prosody transfer, where audio-books, films, dialogues are used to control the style or prosody [5]–[7], [19]. The labeling of styles in audiobooks is an arduous task due to a large number of possible variations in a single emotion or style. This creates difficulty in the development of expressive TTS with predefined emotions. The expressivity transfer plays a vital role in creating expressive TTS for a new speaker, where one has to build the expressive speech corpus every time a new speaker's voice is augmented to multispeaker TTS. The creation of an expressive speech corpus is an expensive process in terms of the time required as well as the cost involved in the recording.

The current E2E TTS systems heavily rely on a sequence to sequence learning framework [2]–[4]. The sequence of phonemes is mapped to the Mel spectrogram, where alignment between phoneme-Mel spectrogram is learned through a location-sensitive attention mechanism [27]. Many techniques have been proposed to use a coarse-grained fixed-length latent representation of expressivity to transfer the desired emotion to the new speaker's voice [15], [16], [18], [19], [21]. The primary goal of the coarse-grained technique is to provide a time-independent fixed dimensional embedding to represent expressivity. For transfer of expressivity, reference encoder is implemented with various deep learning architectures such as Global style token (GST) [16], Variational Autoencoder (VAE) [20], Gaussian mixture VAE (GMVAE) [22]. The reference encoder creates a fixed-length latent variable, representing expressivity or prosody. Thereafter, the extracted latent variable is concatenated with text embedding and passed through a decoder network to synthesize speech in desired expression or prosody, or style. Even though these approaches have shown promising results, they still lack in terms of fine-grained control over expressivity transfer which is conditioned on the sequence of phonemes.

In the proposed work, we use three attention mechanisms at multiple stages (MSA) of multispeaker expressive TTS. The location attention generates expressive weights using outputs of text encoder and reference encoder. This expressive attention weight uncovers the desired emotional strength to be synthesized which is dependent on the phoneme sequence. Then, the self-attention layer featuring salient expressivity information from the output of the reference encoder. Thereafter, the output of the self-attention layer is passed through the GMVAE layer to create a global representation of emotion. Lastly, location-sensitive attention for creating alignment between phoneme sequence and predicted Mel spectrogram is used in the decoder module.

This paper implements a two-staged training approach, firstly the same pair of reference Mel spectrogram and target Mel spectrogram is provided for training of proposed architecture. When model parameters of proposed architecture start converging, cluster sampling is employed to provide reference Mel spectrogram belonging to the same expressive label, which is selected randomly. This process avoids the source speaker leakage, where synthesized speech has a voice quality of reference Mel spectrogram contrary to the target speaker's voice.

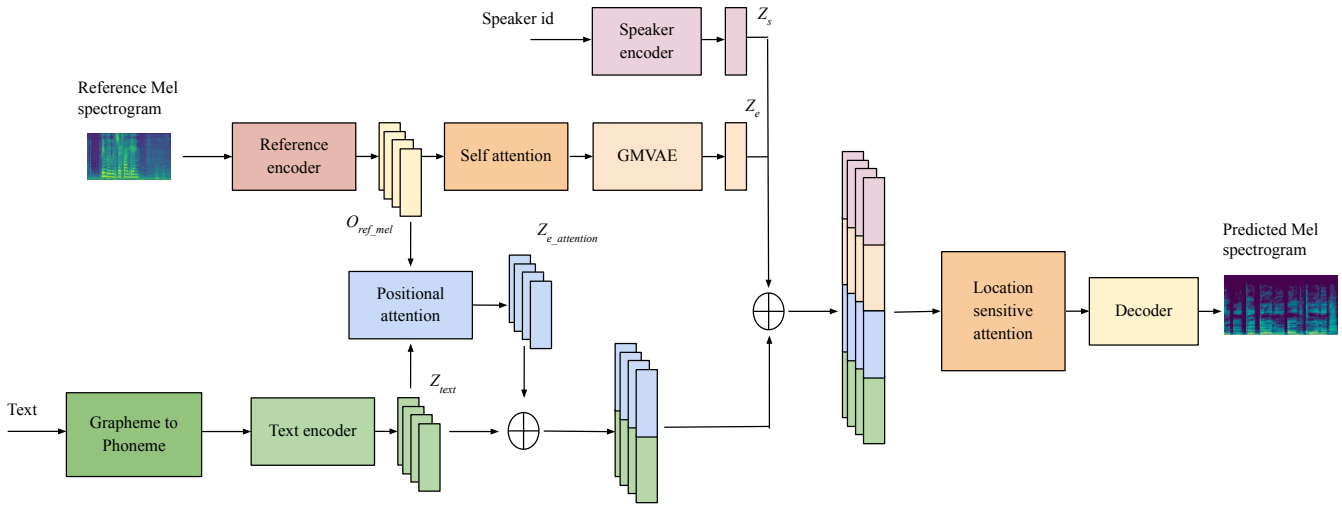

Fig. 1: Framework for multi-stage attention based fine-grained expressivity transfer in Multi-stage-attention (MSA) II TTS system

II. Related Work

A fine-grained VAE framework is proposed to extract latent variables at each token of phoneme embedding [8]. This approach uses sequential prior in a discrete latent space implemented with the help of vector quantization, where each token of phoneme embedding is aligned to the target Mel spectrogram. This work is further extended by creating multilevel alignment for phoneme, word, and utterance [9]. Instead of creating multilevel alignment, we propose to use variable-length latent representation from the reference encoder to create expressive attention weights. These attention weights emphasize emotional prosody extracted from reference Mel spectrogram and trained without any explicit cost function such as vector quantization loss.

Similar to our approach, fine-grained control over prosody is achieved by dot-product attention from phoneme embedding and output of reference encoder in single speaker TTS setting [10]. In addition to emotional attention weights, we extracted global latent expressive information using the GMVAE layer. The GMVAE creates hierarchical disentanglement of a latent variable over the expressive latent attribute in an unsupervised setting. The global latent variable assists in influencing overall expressivity in synthesized speech.

III. Baseline Models

The baseline TTS system use system architecture stated in Tacotron 2 [27], with the addition of an expressivity encoder. The expressivity encoder is implemented using Global style token (GST) [16] and Variational autoencoder (VAE) [18]. First, a sequence to sequence acoustic model predicts mel-spectrograms from a sequence of phoneme-level linguistic inputs, along with expressivity embedding and speaker embedding as explained in [11]. Then a Waveglow neural vocoder converts the mel-spectrograms into a high fidelity audio waveform [12].

IV. Proposed Architecture

For developing fine-grained expressive end-to-end TTS, we modified the multispeaker Tacotron 2 system described in [11]. The proposed approach takes input as text, reference Mel spectrogram, and speaker identity. The text is mapped into a sequence of phoneme identity using explicit grapheme to phoneme converter. We used the same text encoder as used by Tacotron 2 to create text embeddings, z_{text}. The fixed dimensional speaker embedding, z_s, is extracted from the speaker encoder for provided input speaker identities.

We extracted the expressivity information using a reference Mel spectrogram. The reference Mel spectrogram is passed to the reference encoder, which generates a segmental representation of expressivity, denoted as O_{refMel}. The reference encoder comprises six layers of stacked 2D convolutional layers with batch normalization. The gated recurrent units are used for recurrent pooling to compress variable-length O_{refMel} to a fixed dimensional vector. Then, it is passed through the self-attention layer to highlight salient expressivity features. We employed a hierarchical generative model based on multivariate Gaussian mixture variational autoencoder [15] to disentangle the global representation of expressivity denoted as z_e. The GMVAE layer models expressivity as latent attributes using a mixture of Gaussian distribution. This allows the discovery of hidden expressive attributes and makes it easier to disentangle latent space.

For fine-grained expressivity transfer, we extract local information by obtaining attention weights as a correlation between text embedding and segmental representation, O_{refMel}. We integrated both representations using location-sensitive attention, which aligns the z_{text} and O_{refMel}. This attention output provides insight into expressivity strength for each sequence of text embedding. The output of attention is denoted as $z_{e,attention}$, which has the same length, L_{text} as of text

embedding.

In case of MSA I, speaker embedding, global expressive embedding, and expressive attention outputs are concatenated together to obtain encoder output vector from all encoders as $L_{text} \times (z_s, z_e, z_{e,attention})$. After that, we experimented with another system termed MSA II. The proposed system MSA II is shown in Fig 1. The main extension in MSA II is that speaker embedding, global expressive embedding, expressive attention outputs, and text embedding are concatenated together to obtain encoder output vector from all encoders as $L_{text} \times (z_s, z_e, z_{e,attention}, z_{text})$, as described in Fig 1.

We applied a third attention mechanism, location-sensitive attention, to align the target Mel spectrogram and encoder output. The decoder uses the encoder output to generate the Mel spectrogram frame by frame. The output from the previous frame is first passed through the pre-net. The pre-net is composed of fully connected layers with the ReLU activation function. The predicted Mel spectrogram from the pre-net and recurrent network is passed through the post-net. We used the same implementation of post-net as the Tacotron 2 system.

V. 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) [30], Synpaflex speech corpus (approx. 7hrs) [31], and Caroline expressive speech corpus [32]. 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. 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.

VI. Experimental setup

For training the baseline TTS systems, we used the same model hyperparameters as explained in [16], [18], [27] for implementing the Tacotron 2 system and expressivity encoders based on GST, VAE. We have used a 256 dimensional latent variable of expressivity for both GST and VAE. We used 8 heads for the implementation of the self-attention layer. The latent representations, z_e, z_s, z_{text}, $z_{eattention}$, and O_{refMel} are set to 256 dimensions. For training the TTS systems, we incorporated two losses for training the TTS systems, which are mean squared error loss on predicted Mel spectrogram and gate loss on the decoder's location-sensitive attention layer.

The variational inference-based frameworks (VAE, GM-VAE, and MSA) often suffer from Kublack Leibler (KL) annealing problem. In KL annealing, the divergence term abruptly drops to a value close to zero. Therefore, we multiplied the KL divergence term with an additional weight of 0.001 and gradually increased over the training epoch with 0.0001. This technique is also used for training baseline TTS systems with an expressivity encoder implemented for GST and VAE. We have incorporated Waveglow [33] based neural vocoder for synthesizing speech waveform and trained it on 4 French speech synthesis corpora.

VII. Results

A. Objective evaluation

We used Mel Cepstrum Distortion (MCD), F0 Root Mean Squared Error (F0 RMSE), and Band aperiodicity distortion (BAP) for an objective evaluation between reference speech samples and synthesized speech samples. The objective evaluation results are presented in Table I and explained in section VII.

Due to the unavailability of reference emotional speech samples for Lisa, Siwis, and Synpaflex speech corpora, we propose to conduct an objective evaluation of expressivity transfer using cosine similarity score and recognition performance. We develop emotion recognition and speaker recognition systems trained on French speech synthesis corpora.

We implemented a convolutional recurrent neural network model for recognition tasks trained using the cross-entropy loss function. The pre-computed mean of each label of recognition systems is compared with embedding extracted from synthesized speech. The higher cosine similarity scores for speaker and expressivity indicates better expressivity transfer without retaining speaker quality from the reference mel spectrogram used. In addition to cosine similarity, we also measure intelligibility by computing character error rate (CER) with acoustic model of automatic speech recognition system [13] trained on French language.

B. Subjective evaluation

At first, we evaluated the multispeaker expressive TTS systems using Mean Opinion Score (MOS) [34] based listening test. In this work, we used the absolute category ranking scale. Each listener had to assign a score for synthesized speech utterance on a scale between 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 20 French listeners participated in this MOS test and results are displayed in Table I 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.

TABLE I: Subjective and objective evaluation of E2E TTS systems on text-to-speech task, where reference CER is 13.24%

Model	MOS	MCD	F0 RMSE	BAP	CER
GST	3.51 ± 0.3	**4.36**	18.23	0.83	19.70
VAE	3.38 ± 0.4	4.76	18.79	0.88	19.36
MSA I	3.81 ± 0.2	4.49	16.68	**0.82**	17.29
MSA II	**3.85 ± 0.2**	4.51	**16.66**	0.83	**16.59**

TABLE II: Subjective and objective evaluation of E2E TTS systems on expressivity transfer task, where reference CER is 7.22%

Model	CER	Speaker MOS	Expressive MOS	Speaker similarity	Expressive similarity
GST	31.08	2.57 ± 0.2	3.05 ± 0.2	0.62	0.53
VAE	31.15	2.71 ± 0.3	3.12 ± 0.2	**0.69**	0.55
MSA I	25.74	2.75 ± 0.3	3.40 ± 0.3	0.68	0.59
MSA II	**22.85**	**2.83 ± 0.3**	**3.58 ± 0.2**	0.68	**0.61**

For speaker MOS, 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 20 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.

VIII. Discussion

Table I details the evaluation conducted to measure the performance of E2E TTS systems on speech synthesis task (with reference CER computed on reference speech samples from datasets). Furthermore, Table II describes the evaluation of systems on expressivity transfer. From Table I, II, the objective evaluation results are consistent with the subjective evaluation metrics namely MOS, speaker MOS, and expressive MOS. Both MSA I and MSA II systems received the lowest F0 RMSE score. Consequently, MSA systems performed better than the baseline coarse-grained TTS systems with expressivity encoders (GST and VAE). The obtained results on expressivity transfer clearly indicate that MSA I performs slightly lower than the MSA II system. Thus, this demonstrates that concatenation of output of text encoder and expressivity attention weights plays a vital role in improving overall performance in synthesizing expressive speech.

Furthermore, we created a matrix for expressivity similarity (in Fig 2) and speaker similarity (in Fig 3) on speech synthesis results of the MSA II system to measure the similarity/dissimilarity with other classes. From Fig 2 matrix of expressivity similarity scores demonstrates the closeness of expressivity class anger, joy and surprise, while other expressivity classes are distant. Thus, it allows us to observe the capability of the MSA II system to synthesize expressive

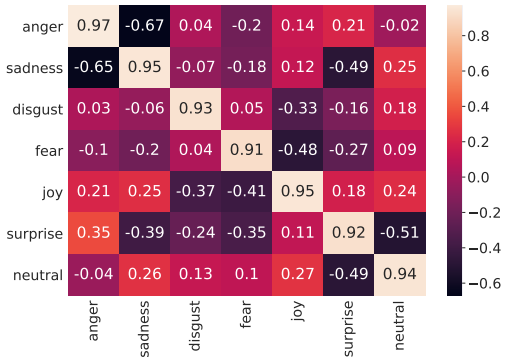

Fig. 2: Scores illustrating closeness of each expressivity class with others created using expressive similarity score computed on MSA II TTS system for speech synthesis datasets

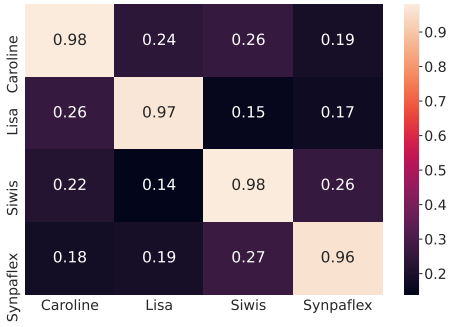

Fig. 3: Scores demonstrating speaker similarity score computed on MSA II TTS system for 4 French speech synthesis datasets

speech with well-segregated expressivity class, i.e., lesser disambiguation between expressivity classes. From Fig 3 matrix for speaker similarity scores provides information regarding MSA II's capability to learn the speaker representation distinctively from other systems. MSA II obtained the lowest CER scores and the highest expressivity similarity scores, demonstrating the importance of expressivity representation at the local and global levels.

IX. Conclusions

We presented a multi-stage attention approach towards extracting phoneme-dependent expressivity transfer to create a local and global representation of expressivity. We proposed using fine-grained TTS architectures in autoregressive TTS settings, where information from the expressivity encoder and text encoder is used to construct expressive attention weights. These attention weights take into account positional phoneme information and its correlation with prosodic information from hidden representation from the expressivity encoder.

Our experimentation with fine-grained TTS systems suggests that concatenation of output of text encoder and expressivity attention weights plays a vital role in improving overall performance in synthesizing speech, thus resulting in better intelligibility than other baseline TTS systems (GST

and VAE). Additionally, we incorporated various objective evaluation metrics such as CER, speaker similarity, and expressive similarity to assist in evaluating various attributes of the expressive TTS system vital for expressivity transfer.

X. Acknowledgements

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] Tan, X., Qin, T., Soong, F.K., Liu, T. "A Survey on Neural Speech Synthesis", ArXiv, abs/2106.15561 2021.

[2] Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q.V., Agiomyrgiannakis, Y., Clark, R.A., Saurous, R.A. "Tacotron: Towards End-to-End Speech Synthesis", INTERSPEECH, 2017

[3] Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M. Neural Speech Synthesis with Transformer Network. AAAI, 2019.

[4] Arik, S.Ö., Chrzanowski, M., Coates, A., Diamos, G.F., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., Shoeybi, M. "Deep Voice: Real-time Neural Text-to-Speech", ICML, 2017.

[5] Zaïdi, J., Seuté, H., Niekerk, B.V., Carbonneau, M. "Daft-Exprt: Robust Prosody Transfer Across Speakers for Expressive Speech Synthesis", ArXiv, abs/2108.02271, 2021.

[6] Yang, F., Yang, S., Wu, Q., Wang, Y., Xie, L. "Exploiting Deep Sentential Context for Expressive End-to-End Speech Synthesis", ArXiv, abs/2008.00613, 2020.

[7] Kim, M., Cheon, S.J., Choi, B.J., Kim, J.J., Kim, N.S. "Expressive Text-to-Speech Using Style Tag", Interspeech 2021.

[8] Sun, G., Zhang, Y., Weiss, R.J., Cao, Y., Zen, H., Rosenberg, A., Ramabhadran, B., Wu, Y. "Generating Diverse and Natural Text-to-Speech Samples Using a Quantized Fine-Grained VAE and Autoregressive Prosody Prior", ICASSP 2020.

[9] Sun, G., Zhang, Y., Weiss, R.J., Cao, Y., Zen, H., Wu, Y. "Fully-Hierarchical Fine-Grained Prosody Modeling For Interpretable Speech Synthesis", ICASSP 2020.

[10] Klimkov, V., Ronanki, S., Rohnke, J., Drugman, T. "Fine-grained robust prosody transfer for single-speaker neural text-to-speech", ArXiv, abs/1907.02479 2019.

[11] Kulkarni, A., Colotte, V., Jouvet, D. "Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis", EUSIPCO 2021.

[12] Prenger, R.J., Valle, R., Catanzaro, B. "Waveglow: A Flow-based Generative Network for Speech Synthesis", ICASSP 2019.

[13] Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J., Castonguay, P., Popova, M., Huang, J., Cohen, J.M. "NeMo: a toolkit for building AI applications using Neural Modules", ArXiv, abs/1909.09577 2019.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

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

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

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

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

[27] 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.

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

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

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

[31] 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.

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

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

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

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