Analysis of expressivity transfer in non-autoregressive end-to-end multispeaker TTS systems

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

Analysis of expressivity transfer in non-autoregressive end-to-end multispeaker TTS systems

Ajinkya Kulkarni, Vincent Colotte, Denis Jouvet

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

ajinkya.kulkarni@loria.fr, vincent.colotte@loria.fr, denis.jouvet@inria.fr

Abstract

The main objective of this work is to study the expressivity transfer in a speaker's voice for which no expressive speech data is available in non-autoregressive end-to-end TTS systems. We investigated the expressivity transfer capability of probability density estimation based on deep generative models, namely Generative Flow (Glow) and diffusion probabilistic models (DPM). The usage of deep generative models provides better log likelihood estimates and tractability of the system, subsequently providing high-quality speech synthesis with faster inference speed. Furthermore, we propose the usage of various expressivity encoders, which assist in expressivity transfer in the text-to-speech (TTS) system. More precisely, we used self-attention statistical pooling and multi-scale expressivity encoder architectures for creating a meaningful representation of expressivity.

In addition to traditional subjective metrics used for speech synthesis evaluation, we incorporated cosine-similarity to measure the strength of attributes associated with speaker and expressivity. The performance of a non-autoregressive TTS system with a multi-scale expressivity encoder showed better expressivity transfer on Glow and DPM-based decoders. Thus, illustrating the ability of multi-scale architecture to apprehend the underlying attributes of expressivity from multiple acoustic features.

Index Terms: expressivity, generative models, text-to-speech

1. Introduction

The main objective of the proposed work is to transfer the expressive attributes to synthesize expressive 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 with well-defined emotional classes.

Abundant research has been conducted for expressivity transfer in the context of the end-to-end (E2E) text-to-speech (TTS) system, mainly developed using E2E autoregressive TTS systems [1, 2, 3, 4]. Besides expressivity transfer, many approaches have been proposed in the context of style transfer and prosody transfer, where audiobooks, films, dialogues are used to control the style or prosody [5, 6, 7, 8]. 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, to avoid creating an expressive speech corpus every time a new speaker's voice is added to the TTS system. 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.

Since the introduction of Fastspeech [3], Fastspeech 2 [9], and ParaNet [10], non-autoregressive end-to-end TTS approaches have enabled faster inference speed. On the contrary to the frame-by-frame prediction of Mel spectrogram, non-autoregressive models generate Mel spectrogram parallelly, thus avoiding the error propagation through previously predicted frames. There has been limited research conducted on expressivity in non-autoregressive TTS [11, 12]. Moreover in [13], authors proposed to leverage the normalizing Flow approach conditioned by the speaker identity. They compared several encodings of speakers and used particularly a Glow-TTS approach for decoding the latent representation obtained by the text encoder.

In this paper, we propose expressivity encoder extension to non-autoregressive TTS systems based on deep generative models and compare the Generative Flow model (Glow) with another generative model based on Diffusion probabilistic model (DPM) approach [14, 15]. The usage of Glow and DPM provides better log-likelihood estimates, thus providing flexible sampling, inference speed, and high-quality synthesis [16, 17]. Glow has recently been used as a non-autoregressive model to synthesize speech waveform conditioned on acoustic features [18, 19]. The Glow architecture has shown state-of-art results in a neural vocoding task providing Mel spectrogram as additional input to affine coupling layer. After that, [15, 20] presented non-autoregressive TTS systems, where Glow architecture was used as a decoder network to generate Mel spectrogram from simple distribution.

In 2015 diffusion probabilistic models were first introduced to model complex data distributions using stochastic calculus [17]. The diffusion probabilistic models are inspired by non-equilibrium thermodynamics for systematically and slowly deconstructing the structure in a data distribution through an iterative forward diffusion process. After that, the reverse diffusion process learns to reconstruct the structure in data, creating a flexible and tractable generative model with an exact sampling scheme. Scoring-based DPM architectures have been proposed for speech synthesis, such as neural vocoders [21, 22, 23] with performance matching to state-of-art models. Grad-TTS [14], and Diff-TTS [24] were recently published for the TTS task where decoders are designed using diffusion probabilistic models. These proposed TTS architectures provided a flexible inference scheme with the trade-off in terms of the quality of synthesized speech and inference speed.

In this paper, we propose to compare different architectures for transferring the expressivity for analyzing the effect of several encodings of expressivity, leveraging two deep generative models (Glow and DPM) as decoder networks and training on small-size corpora. Section 2 presents the proposed architectures: global architectures and several encoders). Section 3 and 4 describe data preparation and experiment setup. Finally, Section 5 presents and discusses the results.

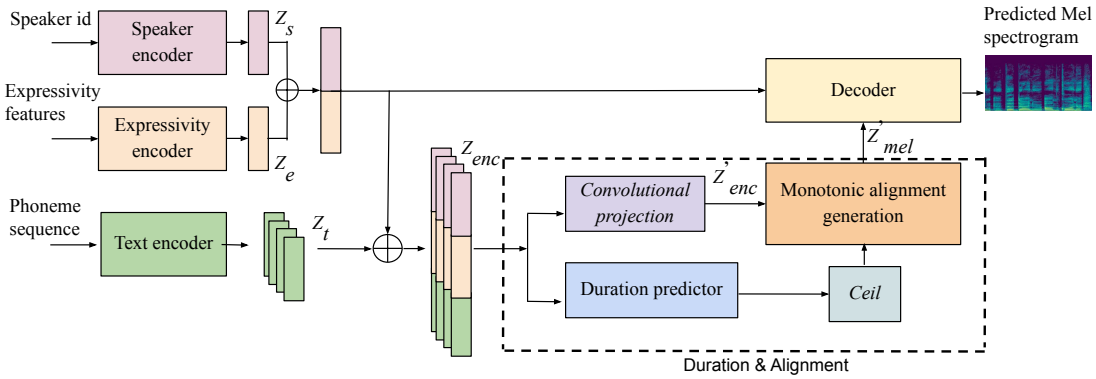

Figure 1: *Non-autoregressive expressive TTS architecture by explicitly conditioning duration predictor, alignment and decoder on expressivity and speaker embeddings*

2. Proposed Architectures

The presented non-autoregressive TTS systems use a framework similar to Glow-based [15] and DPM-based [14] systems, as a text encoder, a duration predictor, a monotonic alignment, and a decoder. Our architecture of non-autoregressive TTS systems is illustrated in Fig. 1. We propose to add a speaker encoder and an expressivity encoder to our proposed TTS systems to enable a multispeaker expressive TTS configuration. The embeddings acquired from the expressivity encoder Z_e, and the speaker encoder Z_s are lengthwise concatenated to the text embedding Z_t. By doing so, the single speaker TTS system is adapted to the multispeaker expressive setting. For the decoder step, we used two variants of networks, namely Generative Flow (Glow) and Diffusion probabilistic models (DPM). During the training of the decoder network, change in distributions from input to output is learned from the provided training data of phoneme sequence and Mel spectrogram.

In the proposed work, the Glow decoder is conditioned with Z_e and Z_s in the affine coupling layer as additional channel inputs. For the DPM decoder, Z_e and Z_s are given to the decoder as additional channels along with the output from the duration predictor and alignment step. For the duration predictor and alignment step (dashed block in in Fig. 1), we used the same implementation as stated in Glow-based [15] and DPM-base [14] systems. We use a monotonic alignment search based on the Viterbi method [25] to find alignment between phoneme sequence and Mel spectrogram. This step is also conditioned on speaker embedding and expressivity embedding. For the alignment, Z'_{enc} which contains statistics from Z_{enc} (μ_{enc} and σ_{enc} on the text length) is aligned with Z'_{mel} (μ_{mel} and σ_{mel} on the Mel spectrogram frame length) which represents the prior distribution for the input of decoder.

The use of monotonic alignment assures monotonicity and surjectivity, ensuring that spoken text is synthesized in the correct order without skipping any input phoneme sequence. In the inference phase, the alignment is predicted by the duration predictor, and the latent variables are sampled from the prior distribution Z'_{mel} (obtained from Z'_{enc}). (*Ceil* block in Fig. 1 is to convert predicted duration value in integer). For the Glow-based TTS system, we used log-likelihood loss and duration loss as described in [15]. For the DPM-based TTS system, we used diffusion loss, log-likelihood loss on Z'_{enc} and duration loss as detailed in [14].

2.1. Speaker encoder

For enabling a multispeaker setting, we created a speaker embedding by providing a one-hot representation to the speaker encoder network. We assigned each speaker identity a unique integer value, which is converted to one-hot encoding. We used nn.Embedding() from PyTorch library to implement speaker encoder network. We derived a speaker embedding Z_s of size 80. Unlike Casanova *et al.* work [13] for which the speaker embedding is pre-trained independently, our speaker encoder is trained with the whole system.

2.2. Expressivity encoder

We implemented three techniques to create expressivity embedding. In the first technique, we assigned an integer value to each expressivity class and derived expressivity embedding similar to speaker embedding. This identity-based expressivity encoder is termed **one-hot embedding**.

The second technique involves providing reference Mel spectrogram to expressivity encoder, with a model based on convolutional recurrent neural network (CRNN) and self-attentive statistical pooling (SASP). We termed this reference Mel spectrogram based network as **CRNN SASP**. Using self-attentive statistical pooling allows assigning different attention weights to different frames and generating weighted means and weighted standard deviations. Thus, self-attention allows deriving long-term variations in expressivity more efficiently. The CRNN is constructed using a single 2-D convolutional layer with single-channel input-output, kernel size of 3x3, along with a single BLSTM layer of 256 hidden units. After the SASP step, we used 1024 and 512 as hidden units to implement feedforward layers (with Selu activation function), which are mapped to 80-dimensional expressivity embedding Z_e as an output of the last feedforward layer.

For the third technique, we first extract four sets of emotional acoustic features for describing: articulation, prosody, phonation, and WORLD [26] vocoder features. This kind of sets are traditionally used in emotion recognition systems [27, 28] as an input to a DNN [29]. Features sets are described in Section 3. For each feature set, we use a CRNN SASP network similar to the previously described CRNN SASP network. The outputs of 4 CRNN SASP networks are concatenated together and given to two feedforward layers (same as above) to obtain the expressivity embedding, Z_e. We termed this expres-

sivity encoder as **multiscale CRNN SASP** because the encoder is designed to encompass a variety of traditional emotional features computed at different frame lengths. The usage of explicitly providing various emotional features assists the expressivity encoder in learning better variations across the frames and underlining the principle factors in defining expressivity classes.

2.3. Text encoder

For each text sentence, we used the Soja grapheme to phoneme converter developed internally in our team to obtain a sequence of phonemes that is given as input to the text encoder. Text encoder is composed of transformer network with similar implementation as used in Transformer TTS [4] with the exclusion of positional encoding.

3. Data preparation

We have used 2 French female speech synthesis corpora for implementing an end-to-end multispeaker expressive TTS system: SIWIS, neutral speech corpus (approx. 5hrs) [32] and Caroline expressive speech corpus (in house speech synthesis corpus) [33]. 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. We have used a sampling rate of 22050 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.

For the multiscale CRNN SASP expressive encoder, we used DisVoice[1] python library to compute features representing articulation, prosody, and phonation. 58 *Articulation features* are computed using frames of 40 ms duration every 20 ms in onset transitions. These features are mainly computed using Bark band energies, and Mel frequency cepstrum coefficient with derivatives. *Prosody features* are 13: Lagrange polynomial to model F0 contour, Lagrange polynomial to model energy contour, and duration of the voiced segment. *Phonation features* include 7 features computed for a frame length of 40 ms (first and second F0 derivatives, jitter, shimmer, amplitude and pitch perturbation quotients, logarithmic energy). For WORLD [26] vocoder features, we used 187 acoustic features, referring to 180 Mel Generalized Cepstrum coefficients, 3 Logarithmic F0 values, 3 Band of Aperiodicity values, and 1 voiced-unvoiced flag extracted over a frame length of 5 ms (computed with pyWorld library).

4. Experimental setup

For transfer of expressivity, we considered two cases: parallel transfer and non-parallel transfer. In the case of parallel transfer, the reference speech utterance given to the expressivity encoder has the same textual content as the input phoneme sequence given to the text encoder (but a different speaker). On the other hand, the non-parallel transfer involves a mismatch of textual content between the input given to the expressivity encoder and the input given to the text encoder (and also a different speaker). For non-parallel transfer, we selected reference acoustic features extracted from long speech utterances of desired expressivity. It is worth noting that the training was done for both cases in a parallel way.

Furthermore, we selected only three distinctive expressivity classes, anger, sadness, and joy, out of six expressivity classes available for Caroline data. After analyzing Ekman's emotion model [34, 35], these three expressivity classes are selected, because they show significant distinguishing characteristics of expressivity. As the aim is to transfer expressive attributes from Caroline data to the SIWIS speaker's voice, we decided not to use Caroline neutral speech data in the training process. Each speech corpus is split into train, validation, and test sets in 80:10:10 ratio, respectively. In this work, we compare six different non-autoregressive TTS models for expressivity transfer based on variations in decoder architecture and expressivity encoder. All the non-autoregressive TTS models were trained for 1500 epochs. As vocoder, we used a pretrained HiFi-GAN model trained on the LJS dataset released on the official implementation website [2].

5. Results and Discussion

In Table 1, we present the performance of E2E TTS systems for the classical task of TTS (i.e. same speaker for encoders and same linguistic content). E2E-GST N-Pair, E2E-VAE N-Pair [30] and multi-stage-attention [31] illustrate the performance of autoregressive (AR) E2E systems trained with Caroline expressive corpora in the multispeaker setting.

We evaluated E2E TTS systems using Mean Opinion Score (MOS) [36] based listening test. 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. A total of 20 French listeners participated in this MOS test and results are displayed in Table 1 with an associated 95% confidence interval. From Table 1, the DPM-multiscale-CRNN-SASP system showed better performance compared to other E2E TTS systems, which also includes the AR E2E TTS systems.

In addition to Mel Cepstrum Distortion, we present cosine similarity scores to measure the strength of speaker attributes and expressivity in synthesized speech. We trained emotion recognition and speaker recognition systems based on the CRNN network (trained using French speech corpora). We computed speaker similarity score by estimating cosine distance between the pre-computed mean of speaker embeddings and embedding extracted from synthesized speech. Similar to the speaker similarity score, we estimated the expressive similarity score with the help of an emotion recognition system for extracting respective embeddings. For TTS task, similarity measures show that the speaker identity and expressivity are preserved for all systems. These values can be seen as a reference (upper limit) when evaluating the systems in the expressivity transfer task.

Table 2 describes the performance of expressivity transfer in a parallel and non-parallel transfer setting with subjective evaluation metrics and objective evaluation metrics. 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 14 French listeners performed both listening tests mentioned above, where each listener scored 18 speech utterances

[1] https://github.com/jcvasquezc/DisVoice

[2] https://github.com/jik876/hifi-gan

Table 1: *Evaluation metrics computed to measure the performance of non-autoregressive end-to-end TTS system in a classical task of TTS. MCD stands for Mel Cepstrum Distortion (between Mel-Generalized Cepstral features computed from the Mel spectrogram).*

Models	AR/ N-AR	MOS	MCD	Speaker similarity	Expressivity similarity
E2E-GST N-Pair [30]	AR	3.72 ± 0.4	4.33	0.91	0.90
E2E-VAE N-Pair [30]	AR	3.47 ± 0.3	4.21	0.92	0.91
Multi-Stage-Attention [31]	AR	3.85 ± 0.2	4.50	0.83	0.94
Glow-one-hot-embedding	N-AR	3.63 ± 0.1	4.57	0.79	**0.81**
Glow-CRNN-SASP	N-AR	3.72 ± 0.2	4.56	**0.81**	0.79
Glow-multiscale-CRNN-SASP	N-AR	$\mathbf{3.86 \pm 0.1}$	**4.53**	0.80	0.76
DPM-one-hot-embedding	N-AR	3.68 ± 0.1	4.61	0.77	0.82
DPM–CRNN-SASP	N-AR	3.85 ± 0.2	4.54	**0.79**	**0.83**
DPM-multiscale-CRNN-SASP	N-AR	$\mathbf{3.94 \pm 0.2}$	**4.51**	0.71	0.78

Table 2: *Evaluation metrics computed to measure the performance of expressivity transfer in parallel and non-parallel setting: several autoregressive (AR) are displayed as comparison point and best results are in bold for Glow- and for DPM-based non-autoregressive (N-AR) systems regarding expressivity encoding architectures.*

		parallel setting				*non-parallel setting*	
Models	AR/ N-AR	Speaker MOS	Expressive MOS	Speaker similarity	Expressivity similarity	Speaker similarity	Expressivity similarity
E2E-GST N-Pair [30]	AR	2.65 ± 0.2	3.15 ± 0.4	0.65	0.55	-	-
E2E-VAE N-Pair [30]	AR	2.90 ± 0.3	3.33 ± 0.4	0.71	0.57	-	-
Multi-Stage-Attention [31]	AR	2.83 ± 0.3	3.58 ± 0.2	0.68	0.61	0.84	0.21
Glow-one-hot-embedding	N-AR	3.12 ± 0.1	2.81 ± 0.2	0.68	**0.35**	0.75	**0.35**
Glow-CRNN-SASP	N-AR	3.20 ± 0.1	2.73 ± 0.1	0.70	0.31	0.76	0.30
Glow-multiscale-CRNN-SASP	N-AR	$\mathbf{3.34 \pm 0.2}$	$\mathbf{2.87 \pm 0.1}$	**0.73**	0.30	**0.79**	0.29
DPM-one-hot-embedding	N-AR	3.13 ± 0.3	2.78 ± 0.2	0.68	**0.35**	0.65	**0.30**
DPM-CRNN-SASP	N-AR	3.06 ± 0.2	2.73 ± 0.1	0.62	0.32	0.72	**0.30**
DPM-multiscale-CRNN-SASP	N-AR	$\mathbf{3.15 \pm 0.1}$	$\mathbf{2.81 \pm 0.2}$	**0.69**	0.34	**0.77**	0.22

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[3]. Speaker MOS and expressive MOS scores indicate that usage of CRNN SASP based expressivity encoder with multiscale improve the overall performance in expressivity transfer. Specifically, performance using Glow decoder demonstrates slightly better expressivity transfer in comparison to DPM decoder, thus indicating normalizing Flows were able to understand latent semantic attributes such as speaker and expressivity and disentangle in latent space representation. Out of three presented approaches for expressivity representation and for both decoder networks, usage of multi-scale emotional features provided better insight in capturing emotion as expressive information. In a parallel transfer setting, the AR E2E approaches were better at transferring expressivity compared to non-autoregressive TTS systems, but AR also showed reduced speaker MOS. Hence there is a trade-off between speaker and expressivity attributes. The similarity measures show the same trend (slightly less marked for speaker).

For non-parallel settings, we noticed reduced expressive similarity scores in the AR E2E system than the non-autoregressive TTS systems. Furthermore, we observed degraded performance in a non-parallel expressivity transfer setting. Thus, usage of small size expressive speech corpora limits E2E TTS systems to apprehend expressivity irrespective of textual content.

[3]Some examples can be found at `https://github.com/ajinkyakulkarni14/NAR_TTS_samples_interspeech_2022`

6. Conclusion

We presented various multispeaker expressive E2E TTS systems developed for the expressivity transfer task. The paper details the effect of decoder and expressivity encoder on expressivity transfer in a non-autoregressive E2E TTS systems. The obtained results indicate that the multiscale architecture of the expressivity encoder assists in apprehending better expressivity representation using various acoustic features ranging from the pitch, aperiodicity, Mel spectrogram, energy, jitter, shimmer, etc. Therefore, an multiscale expressivity encoder creates the embedding containing information associated with various prosodic factors.

We proposes to condition deep generative models on expressivity and speaker embeddings to synthesize expressive speech in a multispeaker setting. The presented work provides an analysis regarding the effect of two deep generative models, Glow and DPM on disentangling expressivity and speaker attributes in expressivity transfer. We also studied two expressivity transfer settings, parallel transfer, and non-parallel transfer. The E2E systems encounter difficulty to transfer expressivity in a non-parallel setting. But the presented non-autoregressive TTS systems were able to transfer expressivity in a parallel setting with limited available expressive speech corpora.

7. 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).

8. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *proceedings of INTERSPEECH*, 2017.

[2] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta, and M. Shoeybi, "Deep voice: Real-time neural text-to-speech," in *proceedings of International Conference on Machine Learning (ICML)*, 2017.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in *proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[4] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[5] D. S. R. Mohan, Q. Hu, T. H. Teh, A. Torresquintero, C. G. R. Wallis, M. Staib, L. Foglianti, J. Gao, and S. King, "Ctrl-p: Temporal control of prosodic variation for speech synthesis," in *proceedings of INTERSPEECH*, 2021.

[6] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. S'aez-Trigueros, and T. Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," in *proceedings of INTERSPEECH*, 2020.

[7] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *proceedings of INTERSPEECH*, 2018.

[8] K. Lee, K. Park, and D. Kim, "Styler: Style modeling with rapidity and robustness via speechdecomposition for expressive and controllable neural text to speech," in *proceedings of INTERSPEECH*, 2021.

[9] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *proceedings of International Conference on Learning Representations (ICLR)*, 2020.

[10] K. Peng, W. Ping, Z. Song, and K. Zhao, "Non-autoregressive neural text-to-speech," in *proceedings of International Conference on Machine Learning (ICML)*, 2020.

[11] K. Lee, K. Park, and D. Kim, "Styler: Style modeling with rapidity and robustness via speechdecomposition for expressive and controllable neural text to speech," in *proceedings of INTERSPEECH*, 2021.

[12] R. Shah, K. Pokora, A. Ezzerg, V. Klimkov, G. Huybrechts, B. Putrycz, D. Korzekwa, and T. Merritt, "Non-autoregressive tts with explicit duration modelling for low-resource highly expressive speech," in *proceedings of International Workshop on Semantic Search Over the Web (SSW)*, 2021.

[13] E. Casanova, C. Shulby, E. Gölge, N. M. Müller, F. S. de Oliveira, A. Candido, A. S. Soares, S. M. Aluísio, and M. Ponti, "Sc-glowtts: an efficient zero-shot multi-speaker text-to-speech model," in *proceedings of INTERSPEECH*, 2021.

[14] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *proceedings of International Conference on Machine Learning (ICML)*, 2021.

[15] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[16] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[17] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *proceedings of International Conference on Machine Learning (ICML)*, 2015.

[18] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *proceedings of ICASSP*, 2019.

[19] S. Kim, S. gil Lee, J. Song, J. Kim, and S. Yoon, "Flowavenet : A generative flow for raw audio," in *proceedings of International Conference on Machine Learning (ICML)*, 2019.

[20] C. Miao, S. Liang, M. Chen, J. Ma, S. Wang, and J. Xiao, "Flow-tts: A non-autoregressive network for text to speech based on flow," in *proceedings of ICASSP*, 2020.

[21] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," in *proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[22] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, N. Dehak, and W. Chan, "Wavegrad 2: Iterative refinement for text-to-speech synthesis," in *proceedings of INTERSPEECH*, 2021.

[23] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *proceedings of International Conference on Learning Representations (ICLR)*, 2021.

[24] M. Jeong, H. Kim, S. J. Cheon, B. J. Choi, and N. S. Kim, "Diff-tts: A denoising diffusion model for text-to-speech," in *proceedings of INTERSPEECH*, 2021.

[25] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *proceedings of the IEEE*, 1989.

[26] M. Morise, F. Yolomori, and K. Ozawa, "World: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, 2016.

[27] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, J. F. Vargas-Bonilla, R. Arora, N. Dehak, P. S. Nidadavolu, H. Christensen, F. Rudzicz, M. Yancheva, H. Chinaei, A. Vann, N. Vogler, T. Bocklet, M. Cernak, J. Hannink, and E. Nöth, "Neurospeech: An open-source software for parkinson's speech analysis," in *proceedings of International Conference on Digital Signal Processing*, 2018.

[28] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, 2007.

[29] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale cnn and attention," in *proceedings of ICASSP*, 2021.

[30] A. Kulkarni, V. Colotte, and D. Jouvet, "Improving transfer of expressivity for end-to-end multispeaker text-to-speech synthesis," in *29th European Signal Processing Conference (EUSIPCO)*, Dublin / Virtual, Ireland, Aug. 2021.

[31] A. Kulkarni, V. Colotte, and D. Jouvet, "Multi-stage attention for fine-grained expressivity transfer in multispeaker text-to-speech system," in *30th European Signal Processing Conference (EUSIPCO)*, 2022.

[32] P.-E. Honnet, A. Lazaridis, P. N. Garner, and J. Yamagishi, "The siwis french speech synthesis database – design and recording of a high quality french database for speech synthesis," IDIAP, Tech. Rep., 2017. [Online]. Available: https://datashare.ed.ac.uk/handle/10283/2353

[33] S. Dahmani, V. Colotte, V. Girard, and S. Ouni, "Conditional variational auto-encoder for text-driven expressive audio visual speech synthesis," in *proceedings of INTERSPEECH*, 2019.

[34] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[35] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[36] M. D. Polkosky and J. R. Lewis, "Expanding the MOS: development and psychometric evaluation of the MOS-R and MOS-X," *International Journal of Speech Technology*, vol. 6, pp. 161–182, 2003.