INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Synthesis of expressive speaking styles with limited training data in a multi-speaker, prosody-controllable sequence-to-sequence architecture

Slava Shechtman[1], *Raul Fernandez*[2], *Alexander Sorin*[1], *David Haws*[2]

[1]IBM Haifa Research Lab, Haifa, Israel
[2]IBM TJ Watson Research Lab, Yorktown Heights, NY, USA

slava@il.ibm.com, fernanra@us.ibm.com, sorin@il.ibm.com, dhaws@us.ibm.com

Abstract

Although Sequence-to-Sequence (S2S) architectures have become state-of-the-art in speech synthesis, the best models benefit from access to moderate-to-large amounts of training data, posing a resource bottleneck when we are interested in generating speech in a variety of expressive styles. In this work we explore a S2S architecture variant that is capable of generating a variety of stylistic expressive variations observed in a limited amount of training data, and of transplanting that style to a neutral target speaker for whom no labeled expressive resources exist. The architecture is furthermore controllable, allowing the user to select an operating point that conveys a desired level of expressiveness. We evaluate this proposal against a classically supervised baseline via perceptual listening tests, and demonstrate that i) it is able to outperform the baseline in terms of its generalizability to neutral speakers, ii) it is strongly preferred in terms of its ability to convey expressiveness, and iii) it provides a reasonable trade-off between expressiveness and naturalness, allowing the user to tune it to the particular demands of a given application.

Index Terms: expressive speech synthesis, sequence to sequence speech synthesis

1. Introduction

Attempting to synthesize speech that reflects a variety of styles and emotions has been an active research area in Text-to-Speech (TTS), going back to pioneering efforts almost two decades ago in unit selection systems [1, 2], an interest that continued to evolve under parametric-synthesis frameworks based on hidden Markov models [3, 4] and on deep neural networks [5]. With the adoption of Sequence-to-Sequence (S2S) frameworks and of Tacotron-like architectures [6, 7], research has shifted to bringing expresiveness into the state-of-the-art quality afforded by these approaches. Initial work in this area sought to extend the Tacotron architecture with a reference encoding, with the goal of capturing expressive style in an unsupervised manner, and of disentangling it from text and speaker identity [8]. Although this approach was able to successfully transfer styles across speakers, it failed to disentangle style from text and worked best when the text of the reference audio closely matched the text to be synthesized. This shortcoming was addressed by the follow-up work of [9], which introduced a Global Style Tokens (GST) layer that can be seen as an unsupervised quantization of the latent space formed by the reference encoding. GST achieves better separation between the text and style, but the learned representation is not controllable, is poorly interpretable in terms of expressive styles (as noted in [10, 11]), and does not address the very important task of cross-speaker transplantation. Subsequent work in [10] and [12] proposed two semi-supervised techniques for mapping the style tokens of [9] to desired labeled expressive styles. However, these works deal with single-speaker datasets containing acted emotional speech and do not address cross-speaker style transplantation either. A variant of GST conditioned on the pitch contour and phonetic alignments was presented in [13], demonstrating cross-speaker style transplantation. However, this approach intensifies the parallel text requirement of [8], limiting general purpose applications of TTS. An unsupervised style modeling technique based on a variational autoencoder applied to a reference audio was proposed in [14], demonstrating incremental improvements over a neutral baseline in a subjective expressiveness evaluation.

To address the ongoing need for a successful framework providing both expressive control and cross-speaker transplantation, [11] introduced a semi-supervised approach which adopted the reference encoding of [8], augmenting it with a post-training quantization of the encoding latent space via principal component analysis. They demonstrated that this approach is capable of cross-speaker style transplantation at the cost of a certain degradation of speech quality and naturalness. Here we follow the setup and goals of that work, but propose a much simpler approach with a controllable trade-off between expressiveness and speech naturalness. Our proposal is in general lines a combination of supervised style embedding with the Hierarchical Prosody Controls proposed in [15] that already proved efficient for the task of word-emphasis control. Specifically, the approach learns syle embeddings in a fully supervised manner, and unlike the previously cited works, it does not use any latent style representations, which tend to suffer from all inherent drawbacks of unsupervised learning. Our approach specifically targets and successfully solves the important case of limited data resources, when we wish to transplant style from a source speaker, for which style labels are available during training, to a target speaker lacking such data. Furthermore, the system is tunable, allowing control of expression strength for each expressive style independently. We introduce the details of this architecture in Sec. 2, and various subjective evaluations validating the approach in Sec. 3 before offering some concluding remarks and future directions in Sec. 4.

2. Architecture

The model architecture adopted in this work (Fig. 1) mostly follows the prosody-controllable S2S model originally proposed for unsupervised/weakly-supervised word-emphasis realization [15]. It is based on a Tacotron2 S2S acoustic model [7], augmented with Hierarchical Prosodic Controls [15]. The S2S acoustic model generates a sequence of acoustic feature vectors (composed of mel-cepstral and periodicity components [16, 17]), where each vector corresponds to a constant-length speech frame, that are then fed to an independently trained, LPCNET-based neural vocoder [16] to generate high-quality samples in

http://dx.doi.org/10.21437/Interspeech.2021-1446

real time [17]. The inputs to the system are a set of symbolic sequences extracted from the input text by a rules-based TTS Front End module (adopted from a unit selection system [2]) or derived from existing speaking-style labels. All input sequences are aligned (by repetition) to contain the same number of symbols and are *one-hot* coded. The input sequences comprise:

- (A) phone identity (including silence phone) together with its lexical stress (primary, secondary or no stress)
- (B) phrase type (4-way: intermediate, declarative, interrogative, exclamation)
- (C) speaking style (3-way: neutral, good news, apology)

All the symbolic sequences are augmented with a special symbol for word boundary, inserted between the words with no silence between them. The *one-hot* coded input sequences are converted to a set of linear embeddings, concatenated together, and fed into Tacotron2 Encoder module (D), consisting of convolutional and bidirectional Long Short-Term Memory (Bi-LSTM) layers [7].

Unlike the speaking style fed to the encoder, a global utterance-level speaker embedding (E), broadcast over the sequence, is concatenated to the encoder output. We found such a setup helped disentangle between speaker identity and generic (speaker-agnostic) speaking style, especially in limited speaking style labeling settings (e.g. when style variability is present just in a single speaker within a multi-speaker training set).

A set of Hierarchical Prosodic Controls, extended from the one introduced in [15] (and further elaborated in Sec. 2.1), is designed to enable both the sentence- and the word-level modifications needed to realize the prosodic patterns of various styles. They are designed to be speaker-agnostic to ease cross-speaker style transplantation. During training (G) these prosodic controls are evaluated from the target waveforms, while at run time a separate predictive module (F) provides default predictions for the hierarchical prosodic trajectories.

The proposed hierarchical controls (Sec. 2.1) are properly normalized and designed to be perceptually interpretable so that they can be modified by the user at inference time with a small set of constant offsets to fine-tune the desired style (see user-exposed control offsets (H) in Fig. 1). Note that the feed-forward operation in block I is placed after the (optional) user request in H. This design is made to preserve the interpretability of the quantities the user gets to manipulate (which would not be the case if the order was reversed, and the independent prosodic targets were blended via a non-linear feed-forward operation).

The Decoder is an autoregressive network that largely follows the standard Tacotron2 architecture [7], but with modifications on the attention, choice of targets, and training losses. These are described in detail in [17] and briefly summarized as follows. The attention is an augmented two-stage attention where the hybrid content- and location-based attention mechanism of Tacotron2 [7] is followed by a structure-preserving mechanism encouraging monotonicity and unimodality in the alignment matrix [17]. The model is trained in a multi-task fashion to predict the end-of-sequence indicator and 80-dim mel cepstral features [7] in tandem with the parameters needed as inputs for an independently trained LPCNET neural vocoder [16]. For 22kHz signals, these features (which we denote as "LPC features") correspond to 256 waveform samples and consist of a 22-dim vector with 20 mel-cepstral coefficients, log f0 and f0 correlation. The predicted LPC features are processed with two post-nets (one to refine the mel-cepstrum, and one to refine the pitch parameters); no post-net is applied on the mel task.

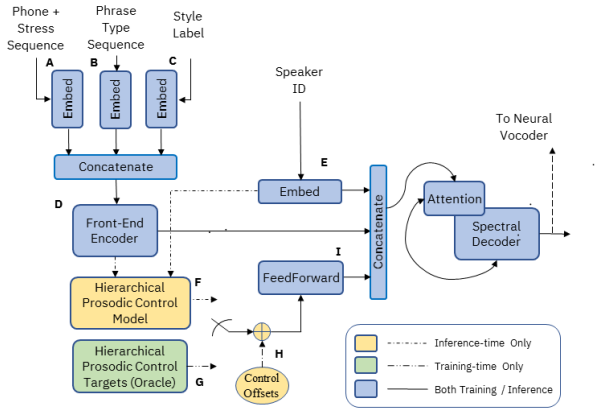

Figure 1: *Multi-speaker S2S synthesis architecture with expressive support.*

Unlike [15, 17], the autoregressive feedback mechanism in the decoder is kept unmodified from the original Tacotron2 model.

Let y_t^M and y_t^L represent the target sequences for the mel and LPC tasks respectively, $\tilde{y}_t^M$ and $\tilde{y}_t^L$ their final predictions, and $\hat{y}_t^L$ the "intermediate" LPC-feature prediction (before the post-net). Then the combined acoustic loss function is used to train the system:

$$\mathcal{L} = MSE(\tilde{y}_t^M, y_t^M) + 0.8MSE(\hat{y}_t^L, y_t^L) + 0.4MSE(\tilde{y}_t^L, y_t^L) + 0.4MSE(\Delta\tilde{y}_t^L, \Delta y_t^L), \quad (1)$$

where the Δ operator applies the first difference in time to a sequence, and $MSE(,)$ is the mean-squared error. The above combined acoustic loss is added to the end of-sequence indicator cross-entropy loss [7] to yield the total training loss. For the sake of space, we omit some detail in this exposition, and refer the reader to [18, 17] for additional background and formulae. The proposed architecture is explored in comparison to a baseline architecture referred to as *Classic Supervision*, in which Hierarchical Prosodic Controls (i.e., blocks F,G,H, and I in Fig. 1) are missing.

2.1. Hierarchical Prosodic-Control Model

In this work we extend a set of four perceptually-interpretable prosodic measurements introduced in [15], evaluated over sentence- and word-intervals, with four more components to better suit speaking-style modeling. We make use of the following statistics:

- S_{dur}: The log of the average per-phone durations, along a sentence (and excluding any silence).
- $S_{\Delta f_0}$: The f_0 dynamics (i.e., the difference between the 95- and 5-percentiles of log-f_0), along a sentence.
- S_{f_0}: The median log-f_0 along a sentence.
- $S_{\angle f_0}$: The log-f_0 linear regression slope along a sentence (excluding any silence).
- W_{dur}: The log of the average per-phone durations (as above), along each word.
- $W_{\Delta f_0}$: The f_0 dynamics (as above), along each word.
- W_{f_0}: The median log-f_0 (as above), along each word.
- $W_{\angle f_0}$: The log-f_0 linear regression slope (as above), along each word.

- V_{f_0}: The median log-f_0 along a single speaker data set. We use it for normalization of S_{f_0} (see below)

Note that the average per-phone durations in the above definitions are estimated as the duration (in seconds) of the relevant spans (word or sentence) divided by the number of phone symbols contained therein, and that therefore no fine-level phonetic alignment is required in the computation (only coarse word-level alignments and either phonetic transcriptions or a dictionary). The above sentence- and word-level properties are propagated down to the temporal granularity of the phonetic encoder outputs (i.e., phones) to form piecewise functions that are constant within a (sentence or word) unit. From this we define the following eight-component prosodic-control target vector:

$$P = Norm_\sigma\{[S_{dur}, S_{\Delta f_0}, S_{f_0} - V_{f_0}, S_{\angle f_0}, W_{dur} - S_{dur}, W_{\Delta f_0} - S_{\Delta f_0}, W_{f_0} - S_{f_0}, W_{\angle f_0} - S_{\angle f_0}]\}, \quad (2)$$

where $Norm_\sigma\{\}$ is the linear map $[-3\sigma^2, 3\sigma^2] \rightarrow [-1, 1]$, and σ^2 is the global (multi-speaker corpus-wide) variance for each of the statistics in P. Note that all measurements in P are gender-agnostic. For that purpose, a gender-specific median pitch S_{f_0} measurement was normalized by subtraction of the speaker statistics (V_{f_0}) to make it speaker- and gender-agnostic.

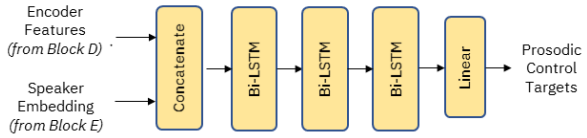

Figure 2: *Architecture of the prosodic controls predictor.*

The prosodic-control predictor (Fig. 2) consists of stacked Bi-LSTMs (3x128), terminated with a linear layer. The predictor is trained separately (after the training of the main model has ended and all its weights are frozen) with MSE loss and ADAM [19] optimizer. At inference time, the predictions of the prosodic-control subnet are rectified to be piecewise constant as the oracle values that the S2S system was trained with. To that end, a mean pooling function is applied to the prediction to be constant between the (known) sentence and word boundaries.

The predicted Prosody Controls (Fig. 1, F) can be adjusted at run time by a set of constant additive offsets (Fig. 1, H), that were tuned in advance by synthesizing and iteratively listening to a small independent development set (20 sentences per style). Note that the global offsets can be applied to both sentence-level and word-level components We applied the fine-tuning to derive distinctive moderate and high levels of expression for each of the non-neutral speaking styles available in the dataset, namely *good news (GDN)* and *apologetic (APO)* styles. During fine tuning, we noticed that applying offsets both on the sentence level and the word level can be useful to control the desired speaking style strength since the prosodic definitions above have correlates in terms of expanding/contracting pitch range, sharpening pitch accents, modifying the speaking rate, etc. The offsets (Δp_1,...,Δp_8) we used for the strong and the moderate level of expression of the speaking styles (per style and per speaker) are provided in Table 1.

3. Evaluation

The training material for the systems comprised four proprietary corpora from three professional native speakers of US English, broken down as follows: a 9.5-hour corpus from a female speaker ($F1$); 2 hrs of expressive recordings from the same female speaker, 1 hr for each one of *GDN* and *APO* styles ($F1_{exp}$) and two more distinct corpora from a female and a male speakers ($F2$ and $M1$) comprising 17.3 and 10.8 hours respectively. The corpus $F1_{exp}$ was collected by indicating to the speaker the expressive style she was to convey and instructing her to be particularly expressive. Notice that the labeled expressive data is available for only one speaker, and that the size of this corpus is considerably smaller than that of the base corpora. The training data for all the evaluated systems comprised all the four corpora, where 5-fold replication was used for $F1_{exp}$ to compensate for the lower prior.

Table 1: *Run-time offsets for speakers F and M.*

	APO	APO$^+$	GDN	GDN$^+$
Δp_1	0.0	0.1	0.0	-0.1
Δp_2	-0.4	-0.5	0.15	0.3
Δp_3	0.0	0.15	0.0	0.1
Δp_4	0.0	0.0	0.0	0.0
Δp_5	0.0	0.0	0.0	-0.05
Δp_6	0.0	0.35	-0.15	0.15
Δp_7	0.0	0.0	0.0	0.0
Δp_8	-0.2	-0.3	-0.2	-0.8

(a) *Run time offsets for F*

	APO	APO$^+$	GDN	GDN$^+$
Δp_1	0.05	0.15	0.0	-0.1
Δp_2	-0.15	-0.5	0.0	0.5
Δp_3	0.1	0.35	0.0	0.05
Δp_4	0.0	0.0	0.0	-0.2
Δp_5	0.0	0.1	0.0	0.0
Δp_6	0.0	0.6	-0.1	0.1
Δp_7	0.0	0.0	0.0	0.0
Δp_8	-0.2	-0.25	-0.35	-0.5

(b) *Run time offsets for M*

In a set of subjective evaluations presented below we would like to assess how well the proposed prosody-controllable system conveys the desired speaking styles, to their various extents, while preserving a decent quality and naturalness in its output. To that end, we consider the following systems:

- **BaseExp**: A baseline Tacotron2-like system with no Prosodic Controls, implementing *Classic Supervision.*
- **PCExp**: The proposed system with Hierarchical Prosodic Controls and with Prosody Control Offsets suited to a *moderate* level of expressiveness (See APO and GDN in Table 1), fine tuned on 20 sentences.
- **PCExp$^+$**: The proposed system with Hierarchical Prosodic Controls and with Prosody Control Offsets suited to a *high* level of expressiveness (See APO$^+$ and GDN$^+$ in Table 1), fine tuned on 20 sentences.
- **PCNeu**: The proposed system (with Hierarchical Prosodic Controls) in a *neutral* speaking style. We use

this system as a quality reference for MOS test only.

In particular, we are interested in examining two test-case scenarios. In the first case, the target synthesis voice matches a speaker for whom we have labeled expressive data (i.e., the matched condition). In the second we assume that the target synthesis voice lacks any such labeled resources for training, and, therefore, any use the system makes of supervised information is done indirectly by transferring knowledge from one speaker to another (we refer to this as the transplant condition).

Table 2: *Results of preference test between baseline (BaseExp) and expressive ((PCExp) systems for the less expressive tuning, broken down by style (APO, GDN) and speaker (*F1*,* M1*)*

Speaker	Style	System			pVal
		BaseExp	NoPref	PCExp	
F1	APO	36.53	13.75	49.72	<0.01
	GDN	45.59	11.76	42.65	0.21
M1	APO	29.84	11.61	58.55	<0.01
	GDN	30.43	9.00	60.57	<0.01

To evaluate the systems defined above, while addressing the matched and transplant cases respectively, we conducted two sets of independent listening tests where the target speakers were F1 (whose training data contains the expressive dataset) and M1 (whose training data does not). We selected a male voice, as the cross-gender style transfer is a more challenging use case. No natural recordings were included in MOS tests since no common set of expressive utterances existed for both voices, and we wanted to run parallel tests. Instead, we opted for an evaluation set of 40 unseen sentences (20 for each expression), where the proposed system in a neutral style serves as a quality reference in MOS test (we selected this reference, as it was already demonstrated in [18] that a prosody-controlled neutral system provides a better quality than a similar system without the prosody controls)[1].

We present the subjective evaluation results: MOS for quality and naturalness and, for expressiveness, an ABX preference tests with a reference anchor. In the latter case, subjects were shown pairs of parallel audio samples sharing the same text, and a fixed stylistic reference audio sample illustrating the target style. Listeners were instructed to select which of the two samples (or neither) better matched the style of the anchor sample; they were made aware that the text of the reference differed from the two evaluated samples, and to ignore textual differences. The raw ratings were subject to an outlier-removal procedure. In the ABX tests, each stimulus (20 per speaker, per style) received 33 independent ratings on average after outlier removal (Tables 2 and 3). In MOS tests, each stimulus (40 per speaker) received 22 independent ratings after outlier removal (Table 4). For the speech quality MOS evaluation, p-values are estimated based on the Student's T cumulative distribution function. For the ABX evaluation, p-values are based on the binomial cumulative distribution function with equal outcome probabilities after dividing the number of 'no preference' votes equally between the two outcomes.

[1] Audio samples are available at http://ibm.biz/IS2021_S2S.

Table 3: *Results of preference test between baseline (BaseExp) and expressive (PCExp$^+$) systems for the more expressive tuning, broken down by style (APO, GDN) and speaker (*F1*,* M1*)*

Speaker	Style	System			pVal
		BaseExp	NoPref	PCExp$^+$	
F1	APO	26.41	10.94	62.66	<0.01
	GDN	29.72	16.39	53.89	<0.01
M1	APO	22.60	9.20	68.20	<0.01
	GDN	32.86	9.00	58.14	< 0.01

Table 4: *Mean opinion scores (and standard deviation) for overall naturalness and quality for baseline (BaseExp), less (PCExp) and more expressive (PCExp$^+$), and neutral (PCNeu) systems for matched- (*F1*) and transplant-speaker (*M1*) conditions.*

Speaker	Systems			
	BaseExp	PCExp	PCExp$^+$	PCNeu
F1	3.93 (.86)	3.84 (.85)	3.82 (.86)	4.00 (.86)
M1	3.62 (.96)	3.79 (.85)	3.68 (.89)	3.91 (.85)

4. Discussion and Conclusions

For the proposed system with moderate level of expressiveness (Table 2) we see a clear difference between the matched (F1) and transplanted (M1) conditions when realizing stylized speech. In the matched condition, classical supervision provides the same level of expressiveness for one of the styles, although it lacks tunability of expressiveness. However, we observe the classical supervision baseline fails to generalize to a speaker who lacks labeled style training data whereas the expressive framework is able to successfully achieve this transplantation of expressive style (with the added feature of making the degree of expressiveness controllable by tuning). This difference between the approaches is consistent with findings in our previous works [15]. Specifically, given supervised data for all speakers, classical supervised methods usually suffice. However, when style annotated data is limited - as is often the case - our expressive system can work well in transplant conditions.

The MOS of Table 4 shows a trade-off between expressiveness and the overall quality of a system. For the expressive system this translates into a graceful degradation (of about 0.12-0.16 MOS) and represents a sensible operating point for an application seeking to balance high quality with expressiveness. Although there is a degradation in MOS for the more expressive system, this may be acceptable depending on the task. We have shown, in any case, that obtaining diverse operating points trading quality for more convincing expressions are possible through the flexibility of the tunable framework. Future work will include improving the quality-expressivity trade-off. Though our focus has been on prosody, speakers modulate more than intonation and tempo to convey different affects (e.g., timbre and voice quality). Although our approach can reproduce spectral characteristics that are consistent with f_0 (e.g., lower f_0 correlating with low spectral energy and creaky voice), further work will look at augmenting the set of acoustic descriptors to address the complexity inherent in the realization of styles.

5. References

[1] E. Eide, "Preservation, identification, and use of emotion in a text-to-speech system," in *Proc. 2000 IEEE Workshop on Speech Synthesis*, September 2002, pp. 127–130.

[2] J. Pitrelli, R. Bakis, E. Eide, R. Fernandez, W. Hamza, and M. Picheny, "The IBM Expressive Text-to-Speech Synthesis System for American English." *IEEE Trans. Audio, Speech and Lang. Processing*, vol. 14, no. 4, pp. 1099–1108, July 2006.

[3] J. Yamagishi, K. Onishi, T. Masuko, , and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *EICE Trans. Inf. Syst.*, vol. 88-D, pp. 502–509, 2005.

[4] J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo, J. Ferreirosa, J. Yamagishi, and J. Montero, "Emotion transplantation through adaptation in HMM-based speech synthesis," *Computer Speech and Language*, vol. 34, no. 1, pp. 292–307, November 2015.

[5] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in *Proc. APSIPA*, Malaysia, July 2017.

[6] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. Weiss, N. Jaitly, Z. Yang, Y. Ying Xiao, Z. Chen, B. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135

[7] J. Shen, R. 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," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 4779–4783.

[8] R. Skerry-Ryan, E. Battenberg, X. Y., 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," *CoRR*, vol. abs/1803.09047, 2018. [Online]. Available: http://arxiv.org/abs/1803.09047

[9] 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," *CoRR*, vol. abs/1803.09017, 2018. [Online]. Available: http://arxiv.org/abs/1803.09017

[10] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *Proc. APSIPA*, Lanzhou, China, 2019, pp. 623–627.

[11] A. Sorin, S. Shechtman, and R. Hoory, "Principal style components: Expressive style control and cross-speaker transfer in neural TTS," in *Proc. Interspeech*, Shanghai, China, 2020, pp. 3411–3415.

[12] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1383–1387, 2019.

[13] R. Valle, J. Li, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 6189–6193.

[14] 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,," in *Proc. ICASSP*, Barcelona, Spain, 2020, pp. 6179–6183.

[15] S. Shechtman, R. Fernandez, and D. Haws, "Supervised and unsupervised approaches for controlling narrow lexical focus in sequence-to-sequence speech synthesis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, January 2021, pp. 431–437.

[16] J. M. Valin and J. Skoglund, "LPCNET: Improving neural speech synthesis through linear prediction," in *ICASSP*, Brighton, England, 2019, pp. 5891–5895.

[17] S. Shechtman, R. Rabinovitz, A. Sorin, Z. Kons, and R. Hoory, "Controllable sequence-to-sequence neural TTS with LPCNET backend for real-time speech synthesis on CPU," *CoRR*, 2020. [Online]. Available: http://arxiv.org/abs/2002.10708

[18] S. Shechtman and A. Sorin, "Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities," in *Proc. SSW10*, Vienna, Austria, 2019, pp. 275–280.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, San Diego, May 2015. [Online]. Available: https://arxiv.org/abs/1412.6980