INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Principal Style Components: Expressive Style Control and Cross-Speaker Transfer in Neural TTS

Alexander Sorin, Slava Shechtman, Ron Hoory

IBM Research - Haifa
sorin@il.ibm.com, slava@il.ibm.com, hoory@il.ibm.com

Abstract

We propose a novel semi-supervised technique that enables expressive style control and cross-speaker transfer in neural text to speech (TTS), when available training data contains a limited amount of labeled expressive speech from a single speaker. The technique is based on unsupervised learning of a style-related latent space, generated by a previously proposed reference audio encoding technique, and transforming it by means of Principal Component Analysis to another low-dimensional space. The latter space represents style information in a purified form, disentangled from text and speaker-related information. Encodings for expressive styles that are present in the training data are easily constructed in this space. Furthermore, this technique provides control over the speech rate, pitch level, and articulation type that can be used for TTS voice transformation.

We present the results of subjective crowd evaluations confirming that the synthesized speech convincingly conveys the desired expressive styles and preserves a high level of quality.

Index Terms: expressive text to speech, speaking style transfer, neural TTS

1. Introduction

Expressive speech synthesis research has a long history, starting from the era of unit selection TTS, through parametric statistical systems such as Hidden Markov Models, and continuing with modern neural TTS frameworks. We address here only a few recent publications that are most relevant to our work.

It is important to specify at the beginning an industry-oriented perspective and goals for the work reported in this publication. Typically, a TTS vendor has several professionally recorded legacy voice datasets, containing neutral speech only. Endowing the legacy TTS voices with the ability to speak in certain expressive styles is of great value, e.g., for spoken conversation in the customer care domain. The styles repertoire may change on demand from time to time. An affordable but technically challenging way to achieve the above goal would be to record a limited amount of speech from a new voice talent, speaking in the desired styles, and apply cross-speaker style transfer to the legacy voices. Accordingly, the main research goal of our work is the cross-speaker expressive style transfer, based on a limited amount of labeled single-speaker data.

Skerry-Ryan et al. [1] learned a latent style embedding space by augmenting the Tacotron network [2] with a trainable audio encoding module referred to as the *Reference Encoder*. Their models were trained on a huge unlabeled expressive audiobook dataset. At synthesis time, a reference audio sample is fed to the model and the system copies the prosody of the reference by conditioning the inference process on the reference encoding. Cross-speaker prosody transfer is demonstrated in [1]. However, this approach has a very serious limitation: the synthesized text must be very close to that of the reference (known as a *parallel text* requirement). Otherwise, the style is not replicated.

To overcome the parallel text limitation, Wang et al. [3] added a lossy decoder module to the system proposed in [1], in the form of a bank of trainable vectors referred to as Global Style Tokens (GST). Their model [3] was trained on the same single-speaker audiobook dataset containing a vast amount of expressive speech. The cross-speaker style transfer was not explored and demonstrated in [3]. One more limitation of the GST approach is that the tokens are not directly interpretable in terms of expressive styles.

Studies by Wu et al. [4] and Kwon et al. [5] focused on the GST approach applied to acted emotional single-speaker datasets. These studies sought to overcome the abovementioned limitation of the GST approach's proposing techniques for establishing a mapping between the token combinations and the desired emotional styles.

Valle et al. [6] extended the GST approach by conditioning pitch contour and phone-to-audio alignment that must be fed to the system in parallel to the reference audio. It improves the accuracy of the prosody transfer from the reference sample (possibly from an unseen speaker) to the synthesized audio. However, this approach intensifies the parallel text requirement (removed by the core GST), which is clearly a barrier for general purpose TTS applications.

A method for one-shot cross-speaker style transfer was proposed by Aggarwal et al. [7]. This method is based on a reference audio encoding using a variational auto-encoder [8]. The system was trained on a multi-speaker neutral dataset. In general, systems trained on partially expressive and labeled data are expected to perform better in an expressive synthesis task. Indeed, the subjective expressiveness evaluation results reported in [7] show only incremental improvement compared to a neutral TTS baseline.

In our semi-supervised approach proposed below, we adopt the reference audio encoding [1], and combine it with an alternative to GST [3] in the form of Principal Component Analysis applied to the latent style space learned by the reference encoding. We encourage readers to visit our demo webpage[1] and listen to speech samples synthesized with our experimental systems.

[1] ibm.biz/BdqMAF

http://dx.doi.org/10.21437/Interspeech.2020-1854

2. Development Setup

We experimented with three high-quality proprietary single-speaker US English datasets recorded at a 22 kHz sampling rate. One of the datasets contains relatively small portions of expressive speech. Hereafter, we refer to this dataset and to the speaker voice as *Source F* (female). During the audio recording sessions for this dataset, the voice talent was instructed to say certain sentences very expressively in one of two styles relevant to customer service TTS applications. These styles are referred to as *Apology* and *Good News*. The texts of the corresponding sentences are coherent with the respective styles. The apologetic samples are expected to convey deep regret and empathy. The Good News samples are expected to sound positive and highly excited.

Two other datasets contain neutral speech only. We refer to those datasets and the corresponding speaker voices as *Target F* (female) and *Target M* (male).

The Source F dataset contains approximately 11,000 utterances including approximately 900 Apology utterances and 900 Good News utterances. The Target F and Target M datasets contain approximately 17,000 and 11,000 neutral utterances, respectively.

Our goal is to develop a multi-speaker system capable of synthesizing expressive speech in the Apology and Good News styles in all three voices, with a focus on the two target voices.

This setup exemplifies the real-life situation where the target datasets play the role of the legacy non-expressive datasets (which they in fact are) and the source dataset plays the role of a new dataset recorded on demand and containing a limited amount of labeled expressive speech data. We believe that a much smaller amount of source neutral speech is enough to accomplish cross-speaker style transfer with our proposed method. However, in our experimental setup, we used all the Source F neutral data to enable high-quality speech synthesis for this speaker voice using a combined three-speaker model.

3. Proposed Approach

Our core TTS system [9] employs the Tacotron-2 architecture [10], modified for faster convergence and better inference stability [11], cascaded with the LPCNet neural vocoder [12]. The system works in real time on a CPU, generating speech with close-to-natural perceptual quality, as shown in [9].

We focus on modifying the first network of the cascade while relying on the generalization capabilities of the legacy LPCNet per-speaker models. Hence, hereafter by the term model, we refer to the model of the first network.

To accomplish the cross-speaker style transfer, we trained common models on a combined three-speaker dataset with speaker identity embedding. We also trained single-speaker models using the Source F dataset to explore and test different style modelling approaches.

Since our data is labeled (by per-utterance Apology, Good News, and Neutral tags), the first thing we tried was supervised style learning by augmenting the network with a style embedding layer. Based on informal listening evaluations, we observed that this method yielded limited success on the Source F dataset but failed in the cross-speaker style transfer. We attributed this observation to the large diversity of homogeneously tagged samples in terms of the expressiveness level.

Next, we turned to the GST-based unsupervised style modeling [3] using the combined three-speaker dataset. We tried different GST layer configurations varying the number of tokens and attention heads. We observed that certain tokens captured certain prosodic patterns. However, these patterns cannot be referred to as certain meaningful speaking styles. We also failed to generate speech in the target expressive styles by feeding the system with expressive reference samples. This led us to conclude that the tokens trained on our data failed to capture the target expressive styles.

It seems that networks with a GST layer have too many degrees of freedom, and a vast amount of expressive training data is required for GST to capture distinct speaking styles. In addition, the dimensionality of the GST output is still too high for manual engineering of perceptually meaningful token combinations. For example, a GST layer with 6 tokens and 4 heads, yields 24 dimensions.

The above observations and speculations led us to look for an alternative to the GST representation of the reference encoding space, a representation that would be more tractable and compact. Our choice fell on the well-established Principal Component Analysis (PCA) technique, which is successfully used in numerous applications for dimensionality reduction, revealing an underlying structure of data and denoising.

3.1. Architecture and Training

We augmented our baseline network with the Reference Encoder proposed in [1] but without the final fully-connected layer followed by an activation function. Thus, the 128-dimensional GRU output forms the output of the Reference Encoder. Also, the model is augmented with a speaker embedding layer, except for when a single-speaker model is trained from the Source F dataset. The speaker embedding and reference encoding outputs are broadcast-concatenated to the phone encoder outputs.

The rest of the architecture, loss-function, and training procedure follow the description given by Shechtman et al. [9]. Note that the training is done in an unsupervised manner and the expression tags are ignored.

3.2. Analysis

Once the model is trained, we select a subset of the training audio data. The subset selection is the only step where the expression tags are used. The subset contains all the Apology samples, all the Good News samples, and a matching number (approximately 900) of randomly selected neutral samples. The neutral samples are evenly distributed between the three speakers, except for when the single speaker Source F model was trained.

The selected audio samples are fed to the Reference Encoder. Its outputs are stored forming a set $\boldsymbol{R}=\{\boldsymbol{r}^{(i)}, i=1,\ldots,2700\}$ of 128-dimenisonal reference encodings, which we refer to as the *analysis set in R space*. The analysis set is centered, i.e., the mean reference encoding $\bar{\boldsymbol{r}}$ is calculated and subtracted from each element of the set. Then PCA is applied to this set, producing an orthonormal basis of 128 *principal components* $\{\boldsymbol{p}_i, i=0,\ldots,127\}$, which are the eigenvectors of the covariance matrix of the analysis set, scaled to unit norm and sorted in the descending order of their corresponding eigenvalues.

Each reference encoding vector $\boldsymbol{r}^{(i)}$ is transformed from R space to the principal component space P_N (where N is the

number of principal components, *0<N<128*) by subtracting the mean and projecting onto the individual principal components:

$$\boldsymbol{\pi}^{(i)} = [\boldsymbol{p}_0 \dots \boldsymbol{p}_{N-1}] \cdot (\boldsymbol{r}^{(i)} - \bar{\boldsymbol{r}}) \quad (1)$$

In (1), the matrix $[\boldsymbol{p}_0 \dots \boldsymbol{p}_{N-1}]$ is formed by stacking the principal-component column vectors. The coordinates $\pi_j^{(i)}$ of the transformed reference encoding are the projection coefficients to the corresponding principal components:

$$\pi_j^{(i)} = \langle \boldsymbol{p}_j, (\boldsymbol{r}^{(i)} - \bar{\boldsymbol{r}}) \rangle \quad (2)$$

In this way, we get the analysis set representation $\Pi=\{\boldsymbol{\pi}^{(i)}, i=1,\dots,2700\}$ in P_N space.

In Figure 1, corresponding to a model trained on the combined three-speaker dataset, each transformed reference encoding $\boldsymbol{\pi}^{(i)}$ is depicted by a point in P_3 space formed by the projection coefficients $(\pi_0^{(i)}, \pi_1^{(i)}, \pi_2^{(i)})$ corresponding to the first three principal components. The points associated with the Neutral, Apology, and Good News samples are colored in green, blue, and red correspondingly.

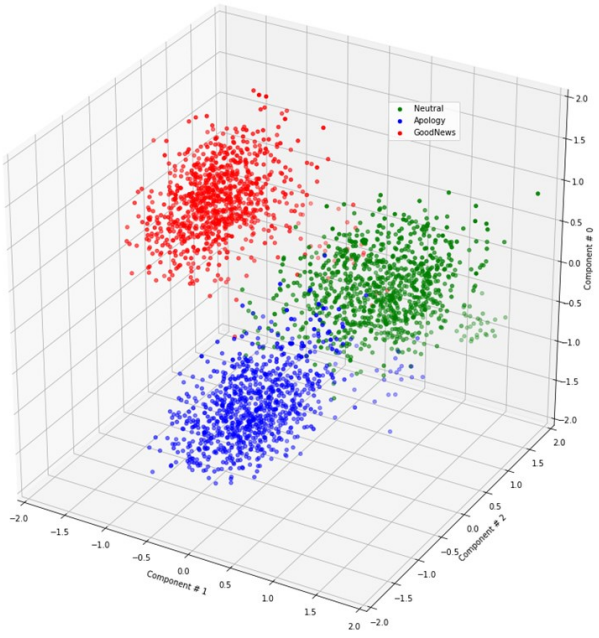

Figure 1: *Reference encodings transformed to the space of the first three principal components. Green points – neutral samples, blue points – Apology samples, and red points – Good News samples.*

We observed that the expressive styles are very well separated even in the plane of the first two components, with the third component somewhat contributing to the separation. We also found that higher components are not indicative of the expressive styles. The latter is evident from examining histograms of the projection coefficients $\pi_j^{(i)}$ onto individual principal components, as introduced in the next section. It is also evident from Figure 1 that the Neutral cluster is not split into separate clouds, even though it contains points originating from three speakers. Thus, the above analysis indicates that the Reference Encoder managed to extract the relevant expressive style information. The analysis also reveals that the style information is fully represented by the first three principal components in a speaker invariant form. Note that the projection coefficients $\{\pi_j^{(i)}, j = 3,\dots,127\}$ corresponding to higher principal components, on the average contribute more than 80% to the norm of the $\boldsymbol{\pi}^{(i)}$ vector. Presumably, the higher principal components account for most of the text structure related information and other information irrelevant to the expressive style.

3.3. Synthesis

The above analysis suggests a way for constructing a text- and speaker-invariant style encoding $\boldsymbol{e}^{(s)}$ for style $s \in \{Apology, Good\ News, Neutral\}$ by picking a point $\boldsymbol{\alpha}^{(s)} = (\alpha_0^{(s)}, \alpha_1^{(s)}, \alpha_2^{(s)})$ in P_3 space and transforming it back to the original reference encoding space R:

$$\boldsymbol{e}^{(s)} = \sum_{j=0}^{2} \alpha_j^{(s)} \cdot \boldsymbol{p}_j + \bar{\boldsymbol{r}} \quad (3)$$

We refer to the values $(\alpha_0^{(s)}, \alpha_1^{(s)}, \alpha_2^{(s)})$ as *style controls*.

Histograms of the projection coefficients $\{\pi_j^{(i)}, j = 0,1,2\}$ over the set Π within the Apology, Good News, and Neutral clusters provide a convenient means for the style controls selection. Such histograms are shown in Figure 2.

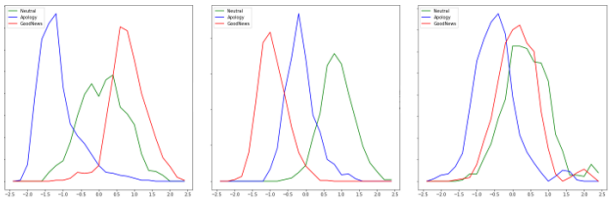

Figure 2: *Histograms of the projection coefficient onto principal components #0 (left plot), #1 (middle plot), and #2 (right plot). Histograms corresponding to the Apology, Good News, and Neutral clusters are shown in blue, red, and green, respectively.*

For example, to construct an encoding for style *s=Apology* using Equation (3), the control values $(\alpha_0^{(s)}, \alpha_1^{(s)}, \alpha_2^{(s)})$ can be selected as the points of maxima of the blue curves presented on the left, middle, and right plot of Figure 2, respectively. We observed that to intensify the style perception, the control values should be extrapolated, where possible. In our example with the Apology style, this means that the first and third controls move to the left from the point of maximum position.

The style encodings prepared offline as described above are used at synthesis time to condition the inference process.

We observed that speech synthesized with the conditioning on expressive style encodings, especially extrapolated Good News encodings, sometimes features unnaturally sounding pitch rising towards the end of the sentences. We devised a simple and effective technique that removes this negative effect. Regularly, during the inference, the same style encoding is concatenated to each phone encoding in the sentence. We changed this behavior and blended the expressive style encoding with a neutral style encoding towards the end of each sentence. The blending starts in the vicinity of the sentence break, e.g., at the eighth phone before the end. Its strength evolves linearly in time, ending with a purely neutral encoding at the last phone. The effect of this technique is demonstrated on our demo webpage.

Constructing encodings for the three-speaker system in a similar way from individual higher principal components, we observed that the components with indexes 3, 4, and 5 (i.e., coefficients α_3, α_4, α_5) control global prosodic features such as speech rate, pitch level, and articulation (legato – staccato), respectively. The effect of these components is similar for all three speakers. These prosodic features are not necessarily related to the expressive styles, but their control can be used for TTS voice transformation. The samples demonstrating the effect of the three higher components can be found on our demo webpage.

We call the proposed method presented in this section the Principal Style Components (PSC) method.

4. Evaluation

We performed a crowd subjective evaluation on the Amazon Mechanical Turk (AMT) platform comparing the expressiveness of speech synthesized with the PSC method in the Apology and Good News styles against baseline systems. We adopted an ABX test scheme, like the one used in [3]. We collected 42 text passages (each containing 1 or 2 sentences) on the Internet. Of the passages, 21 convey apologetic messages and the remaining 21 contain good news announcements. Although the style control enables expressive rendering of an arbitrary text message, this choice of texts improves the conditions for the baseline system to produce speech with the appropriate expression. The two text sets were synthesized using the PSC method with conditioning on fixed Apology and Good News style encodings, respectively. This was done for all three speaker voices. The same text sets were also synthesized for all three voices by the baseline systems.

The PSC samples in Source F voice were synthesized at an initial stage of the study from a single speaker model. The baseline Source F samples were generated from the same PSC system conditioned on a Neutral-style encoding. All the other PSC samples were generated from a three-speaker model, and the rest of the baseline samples were synthesized from corresponding pre-existing models [9].

In this way, we produced two parallel competing sets of samples per speaking style and per speaker voice. Two highly expressive held-out natural speech samples were selected from the Source F dataset as references: one sample featuring the Apology style and the other featuring the Good News style. The reference samples can be found on our demo webpage. Each pair of the synthesized samples was rated by each one of the 40 subjects. In each step of the evaluations, subjects were presented with a pair of competing synthesized samples (in random order) and the reference featuring the corresponding style. Subjects were asked to select the sample that was closer to the reference in terms of overall speaking style.

It should be clear from the PSC method description that the reference samples (like any other individual sample) were not used for the style encoding construction. The texts of the reference and test samples are completely different, and so are the speaker voices, except for the test samples synthesized in the Source F voice. The only purpose of having the reference was to give the subjects a clear idea of what the target speaking style was. We preferred this method over describing the target speaking style in words, with a risk of a misinterpretation.

The expressiveness evaluation results presented in Table 1 indicate a consistent and strong preference of PSC over the baseline. All the differences are statistically significant with p-value < 0.01.

To assess the overall quality of speech synthesized with the conditioning on the expressive style encodings, we carried out a crowd mean opinion score (MOS) test on the AMT platform. The same synthesized samples evaluated in the ABX test were grouped according to the speaker voice and system. Thus, for each speaker voice, we have two groups corresponding to the Baseline and PSC systems. As a hidden high anchor, a third group for each speaker voice was composed from natural speech recordings (featuring other texts) of that speaker. Thirty subjects rated each sample. The results are presented in Table 2. All the intergroup differences are statistically significant with p-value < 0.01. The quality scores confirm that the baseline system generates high quality speech, with a small gap of 0.12 MOS points to natural speech. With respect to the Target F and Target M voices, this is not a new finding. The performance of these models was reported in our previous work [9]. The baseline performance of the new model (Source F PSC-model conditioned on a Neutral style encoding) is at the same level.

Although the quality scores of the expressive PSC samples are high, they are somewhat below the baseline with gaps of 0.13, 0.17, and 0.11 MOS points for the three voices, correspondingly. In our opinion, the expressiveness of the PSC samples came at the cost of a slightly muffled quality. Restoration of the voice sound crispness is a subject for future work.

Table 1: *Expressive style AXB preference results*

Voice	Style	Baseline	No preference	PSC
Source F	Apology	29.8%	3.1%	67.1%
Source F	Good news	35.4%	6.2%	58.4%
Target F	Apology	28.2%	15.2%	56.6%
Target F	Good news	17.7%	4.8%	77.5%
Target M	Apology	36.7%	6.6%	56.7%
Target M	Good news	20.6%	4.9%	74.5%

Table 2: *Quality MOS results*

Voice	Natural	Baseline	PSC
Source F	4.17	4.05	3.92
Target F	4.03	3.91	3.74
Target M	4.33	4.21	4.10

5. Conclusions

The novel technique proposed in this work enables cross-speaker expressive style transfer when available training data contains a limited amount of labeled expressive speech from a single speaker. The technique is based on unsupervised learning of a style-related latent space, generated by the reference audio encoding, and transforming it using Principal Component Analysis to another low-dimensional space. The latter space represents style information in a purified form, disentangled from text and speaker-related information. Encodings for expressive styles present in the training data are easily constructed in this space. Furthermore, this technique provides control over speech rate, pitch level, and articulation type that can be used for TTS voice transformation.

6. References

[1] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron," arXiv preprint arXiv:1803.09047, 2018.

[2] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, RJ. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, and Q. V. Le, "Tacotron: A Fully End-to-End Text-to-Speech Synthesis Model," arXiv preprint arXiv:1703.10135. 2017.

[3] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis," arXiv preprint arXiv:1803.09017, 2018.

[4] P. Wu, Z. Ling, L. Liuy, Y. Jiangy, H. Wuy and L. Dai, "End-to-End Emotional Speech Synthesis Using Style Tokens and Semi-Supervised Training," arXiv preprint arXiv: 1906.10859, 2019.

[5] O. Kwon, I. Jang, C. H. 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, September 2019.

[6] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, Barcelona, May 2020.

[7] 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," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, Barcelona, May 2020.

[8] D. P. Kingma and M.Welling, "Auto-Encoding Variational Bayes," in Proceedings of the International Conference on Learning Representations (ICLR-14), 2014.

[9] S. Shechtman, C. Rabinovitz, A. Sorin, Z. Kons and R. Hoory, "Controllable Sequence-To-Sequence Neural TTS with LPCNET Backend for Real-time Speech Synthesis on CPU," *IEEE Transactions on Acoustics, Speech and Signal Processing*, arXiv preprint arXiv: arXiv:2002.10708, 2020.

[10] S, Jonathan, et al., "Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, April 2018.

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

[12] V. Jean-Marc and J. Skoglund, "LPCNet: Improving Neural Speech Synthesis Through Linear Prediction." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton UK, May 2019.