Expressive TTS Training with Frame and Style Reconstruction Loss

Rui Liu, Student Member, IEEE, Berrak Sisman, Member, IEEE, Guanglai Gao, Haizhou Li, Fellow, IEEE

Abstract—We propose a novel training strategy for Tacotron-based text-to-speech (TTS) system to improve the expressiveness of speech. One of the key challenges in prosody modeling is the lack of reference that makes explicit modeling difficult. The proposed technique doesn’t require prosody annotations from training data. It doesn’t attempt to model prosody explicitly either, but rather encodes the association between input text and its prosody styles using a Tacotron-based TTS framework. Our proposed idea marks a departure from the style token paradigm where prosody is explicitly modeled by a bank of prosody embeddings. The proposed training strategy adopts a combination of two objective functions: 1) frame level reconstruction loss, that is calculated between the synthesized and target spectral features; 2) utterance level style reconstruction loss, that is calculated between the deep style features of synthesized and target speech. The proposed style reconstruction loss is formulated as a perceptual loss to ensure that utterance level speech style is taken into consideration during training. Experiments show that the proposed training strategy achieves remarkable performance and outperforms a state-of-the-art baseline in both naturalness and expressiveness. To our best knowledge, this is the first study to incorporate utterance level perceptual quality as a loss function into Tacotron training for improved expressiveness.

Index Terms—Expressive speech synthesis, Tacotron, frame and style reconstruction loss, emotion recognition

I. INTRODUCTION

WITh the advent of deep learning, neural TTS has shown many advantages over the conventional TTS techniques [1]–[3]. For example, encoder-decoder architecture with attention mechanism, such as Tacotron [4]–[7], has consistently achieved high voice quality. The key idea is to integrate the conventional TTS pipeline [8], [9] into an unified framework that learns sequence-to-sequence mapping from text to a sequence of acoustic features [7], [10]–[13]. Furthermore, together with a neural vocoder [5], [14]–[19], neural TTS generates natural-sounding and human-like speech which achieves state-of-the-art performance. Despite the progress, the expressiveness of the synthesized speech remains to be improved.

Speech conveys information not only through phonetic content, but also through its prosody. Speech prosody can affect syntactic and semantic interpretation of an utterance [20], that is called linguistic prosody. Speech prosody is also used to display one’s emotional state, that is referred to as affective prosody. Both linguistic prosody and affective prosody are manifested over a segment of speech beyond short-time speech frame. Linguistically, speech prosody in general refers to stress, intonation, and rhythm in spoken words, phrases, and sentences. As speech prosody is the result of the interplay of multiple speech properties, it is not easy to define speech prosody by a simple labeling scheme [21]–[25]. Even if a labeling scheme is possible [26], [27], a set of discrete labels may not be sufficient to describe the entire continuum of speech prosody.

Besides naturalness, one of the factors that differentiates human speech from today’s synthesized speech is their expressiveness. Prosody is one of the defining features of expressiveness that makes speech lively. Several recent studies successfully improve the expressiveness of Tacotron TTS framework [28]–[32]. The idea is to learn latent prosody embedding, i.e. style token, from training data [28]. At runtime, the style token can be used to predict the speech style from text [29], or to transfer the speech style from a reference utterance to target [30]. It is observed that such speech styling is effective and consistently improves speech quality. Sun et al. [31], [32] further study a hierarchical, fine-grained and interpretable latent variable model for prosody rendering. The studies show that precise control of the prosody style leads to improvement of prosody expressiveness in the Tacotron TTS framework. However, several issues have hindered the effectiveness of above prosody modeling techniques.

First, the latent embedding space of prosody is learnt in an unsupervised manner, where the style is defined as anything but speaker identity and phonetic content in speech. We note that many different styles co-exist in speech. Some are speaker dependent, such as accent and idiolect, others are speaker independent such as prosodic phrasing, lexical stress and prosodic stress. There is no guarantee that such latent embedding space of style represents only the intended prosody. Second, while the techniques don’t require the prosody annotations on training data, they require a reference
speech or a manual selection of style token [28] in order to explicitly control the style of output speech during run-time inference. While it is possible to automate the style token selection [29], a correct prediction of style token is subject to the design of the style token dictionary, and the run-time style token prediction algorithm. Third, the style token dictionary in Tacotron is trained from a collection of speech utterances to represent a large range of acoustic expressiveness for a speaker or an audiobook [28]. It is not intended to provide differential prosodic details at phrase or utterance level. It is desirable for Tacotron system to learn to automate the prosody styling in response to input text at run-time, that will be the focus of this paper.

To address the above issues, we believe that Tacotron training should minimize frame level reconstruction loss [4], [5] and utterance level perceptual loss at the same time. Perceptual loss is first proposed for image stylization and synthesis [33]–[36], where feature activation patterns, or deep features, derived from pre-trained auxiliary networks are used to optimize perceptual quality of output image. Several computational models have been proposed to approximate human perception of audio quality, such as Perceptual Evaluation of Audio Quality (PEAQ) [37], Perceptual Evaluation of Speech Quality (PESQ) [38], and Perceptual Evaluation of Audio methods for Source Separation (PEASS) [39]. However, such models are not differentiable, hence cannot be directly employed during TTS training. We believe that utterance level perceptual loss based on deep features that reflects global speech style would be useful to improve overall speech quality.

We are motivated to study a novel training strategy for TTS systems, that learns to associate prosody styles with input text implicitly. We would like to avoid the use of prosody annotations. We don’t attempt to model prosody explicitly, but rather learn the association between prosody styles and input text using existing neural TTS system, such as Tacotron. As the training strategy is only involved during training, it doesn’t change the run-time inference process for neural TTS system. At run-time, we don’t require any reference signal nor manual selection of prosody style.

The main contributions of this paper include: 1) we propose a novel training strategy for Tacotron TTS that effectively models both spectrum and prosody generation; 2) we propose to supervise the training of Tacotron with a fully differentiable perceptual loss, which is derived from a pre-trained auxiliary network, in addition to frame reconstruction loss; 3) we successfully implement a system that doesn’t require any reference speech nor manual selection of prosody style at run-time; and 4) we successfully validate the proposed perceptual loss by showing consistent speech quality improvement. To our best knowledge, this is the first study to incorporate perceptual loss into Tacotron training for improved expressiveness.

This paper is organized as follows: In Section II, we present the research background and related work to motivate our study. In Section III, we propose a novel training strategy for TTS system with frame and style reconstruction loss. In Section IV, we report the subjective and objective evaluations. Section V concludes the discussion.

II. BACKGROUND AND RELATED WORK

This work is built on several previous studies on neural TTS, prosody modeling, perceptual loss, and speech emotion recognition. Here we briefly summarize the related previous work to set the stage for our study, and to place our novel contributions in a proper context.

A. Tacotron2-based TTS

In this paper, we adopt Tacotron2-based [5] TTS model as a reference baseline, which is also referred to as Tacotron baseline for brevity. For rapid turn-around, we use Griffin-Lim waveform reconstruction instead of WaveNet vocoder in this study. We note that the selection of waveform generation technique will not affect our judgment and performance comparison.

The overall architecture of the reference baseline includes encoder, attention-based decoder and Griffin-Lim algorithm as illustrated in Fig. 1. The encoder consists of two components, a convolutional neural network (CNN) module [41], [42] that has 3 convolutional layers, and a bidirectional LSTM (BLSTM) [43] layer. The decoder consists of four components: a 2-layer pre-net, 2 LSTM layers, a linear projection layer and a 5-convolution-layer post-net. The decoder is a standard autoregressive recurrent neural network that generates mel-spectrogram features and stop tokens frame by frame.

Just like other TTS systems, Tacotron [4], [5] TTS system predicts mel-spectrum features from input sequence of characters by minimizing a frame level reconstruction loss. Such frame level objective function focuses on the distance between spectral features. It does not seek to optimize the perceptual quality at utterance level. To improve the suprasegmental expressiveness, there have been studies [29], [32], [44] on latent prosody representations, that make possible prosody styling in Tacotron TTS framework. However, most of the studies rely on the style tokens mechanism to explicitly model the prosody. Simply speaking, they build a Tacotron TTS system that synthesizes speech, and learns the global style tokens (GST) at the same time. At run-time inference, they apply the style tokens to control the expressive effect [28], [30], that is referred to as the GST-Tacotron paradigm.
In this paper, we advocate a new way of addressing the expressiveness issue by integrating a perceptual quality motivated objective function into the training process, in addition to the frame level reconstruction loss function. We no longer require any dedicated prosody control mechanism during run-time inference, such as style tokens in Tacotron system.

### B. Prosody Modeling in TTS

Prosody conveys linguistic, para-linguistic and various types of non-linguistic information, such as speaker identity, intention, attitude and mood [45], [46]. It is inherently supra-segmental [1], [47] due to the fact that prosody patterns cannot be derived solely from short-time segments [48]. Prosody is hierarchical in nature [48]–[51] and affected by long-term dependencies at different levels such as word, phrase and utterance level [52]. Studies on hierarchical modeling of F0 in speech synthesis [1], [53], [54] suggest that utterance-level prosody modeling is more effective. Similar studies, such as continuous wavelet transform, can be found in many speech synthesis related applications [52], [55]–[58]. In this paper, we will study a novel technique to observe utterance-level prosody quality during Tacotron training to achieve expressive synthesis.

The early studies of modeling speaking styles are carried out on Hidden Markov Models (HMM) [9], [59], where we can synthesize speech with an intermediate speaking style between two speakers through model interpolation [60]. To improve the HMM-based TTS model, there have been studies to incorporate unsupervised expression cluster information during training [61]. Deep learning opens up many possibilities for expressive speech synthesis, where speaker, gender, and age codes can be used as control vectors to change TTS output in different ways [62]. The style tokens, or prosody embeddings, represent one type of such control vectors, that is derived from a representation learning network. The success of prosody embedding motivates us to further develop the idea.

Tacotron TTS framework has achieved remarkable performance in terms of spectral feature generation. With large training corpus, it may be able to generate natural prosody and expression by remembering the training data using a large number of network parameters. However, its training process doesn’t aim to optimize the system for expressive prosody rendering. As a result, Tacotron TTS system tends to generate speech outputs that represent model average, rather than the intended prosody.

The idea of global style tokens [28], [29] represents a success in controlling prosody style of Tacotron output. Style tokens learn to represent high level styles, such as speaker style, pitch range, and speaking rate across a collection of utterances or a speech database. We argue that they neither necessarily represent the useful styles to describe the continuum of prosodic expressions [63], nor provide the dynamic and differential prosodic details with the right level of granularity at utterance level. Sun et al. [31], [32] study a way to include a hierarchical, fine-grained prosody representation, that represents the recent attempts to address the problems in GST-Tacotron paradigm.

We would like to address three issues in the existing prosody modeling in Tacotron framework, 1) lack of prosodic supervision during training; 2) limitation of explicit prosody modeling, such as style tokens, in describing the continuum of prosodic expressions; 3) lack of dynamic and differential prosody at utterance level.

### C. Perceptual Loss for Style Reconstruction

It is noted that frame-level reconstruction loss, denoted as frame reconstruction loss in short, is not always consistent with human perception because it doesn’t take into account human sensitivities to temporal and spectral information, such as prosody and temporal structure of the utterance. For example, if one repeatedly asks the same question two times, despite the perceptual similarity of two utterances, they would be very different as measured by frame-level losses.

Perceptual loss refers to the training loss derived from a pre-trained auxiliary network [34]. The auxiliary network is usually trained on a different task that provides perceptual quality evaluation of an input at a higher level than a speech frame. The intermediate feature representations, generated by the auxiliary network in form of hidden layer activations, are usually referred to as deep features. They are used as the high level abstraction to measure the training loss between reconstructed signals and reference signals. Such training loss is also called deep feature loss [64], [65].

In speech enhancement, perceptual loss has been used successfully in end-to-end speech denoising pipeline, with an auxiliary network pre-trained on audio classification task [66]. Kataria et al. [64] propose to use perceptual loss which optimizes the enhancement network with an auxiliary network pre-trained on speaker recognition task. In voice conversion, Lo et al. [67] propose deep learning-based assessment models to predict human ratings of converted speech. Lee [68] propose a perceptually meaningful criterion where human auditory system was taken into consideration in measuring the distances between the converted speech and the reference.

In speech synthesis, Oord et al. propose to train a WaveNet-like classifier with perceptual loss for phone recognition [69]. As the classifier extracts high-level features that are relevant for phone recognition, this loss term supervises the training of WaveNet to look after temporal dynamics, and penalize bad pronunciations. Cai et al. [70] study to use a pre-trained speaker embedding network to provide feedback constraint, that serves as the perceptual loss for the training of a multi-speaker TTS system.

In the context of prosody modeling, the perceptual loss in the above studies can be generally described as style reconstruction loss [34]. Following the same principle, we would like to propose a novel auxiliary network, that is pre-trained on a speech emotion recognition (SER) task, to extract high level prosody representations. By comparing prosody representations in a continuous space, we measure perceptual loss between two utterances. While perceptual loss is not new in speech reconstruction, the idea of using a pre-trained emotion recognition network for perceptual loss is a novel attempt in speech synthesis.
D. Deep Features for Perceptual Loss

Now the question is which deep features could be suitable for measuring perceptual loss. We benefit from the prior work in prosody modeling. Prosody embedding in Tacotron is a type of feature learning, that learns the representation for prediction or classification tasks. With deep learning algorithms, automatic feature learning can be achieved in either supervised, such as multilayer perceptron [71], or unsupervised manner, such as variational autoencoder [72].

Deep features are usually more generalizable, and easier to manage than hand-crafted or manually designed features [73]. There have been studies on representation learning for prosody patterns, such as speech emotion [74], and speech styles [28].

Affective prosody refers to the expression of emotion in speech [75], [76]. It is prominently exhibited in emotional speech database. Therefore, the studies in speech emotion recognition provide valuable insights into prosodic modeling. Emotion are usually characterized by discrete categories, such as happy, angry, and sad, and continuous attributes, such as activation, valence and dominance [77], [78]. Recent studies show that latent representations of deep neural networks also characterize well emotion in a continuous space [71].

There have been studies to leverage emotion speech modeling for expressive TTS [30], [61], [79]–[81]. Eyben et al. [61] incorporate unsupervised expression cluster information into a HMM-based TTS system. Skerry-Ryan et al. [30] study learning prosody representation from animated and emotive storytelling speech corpus. Wu et al. [79] propose a semi-supervised training of Tacotron TTS framework for emotional speech synthesis, where style tokens are defined to represent emotion categories. Gao et al. [80] propose to use an emotion recognizer to extract the style embedding for speech style transfer. Um et al. [81] study a technique to apply style embedding to Tacotron system to generate emotional speech, and to control the intensity of emotion expressiveness.

All the studies point to the fact that emotion-related deep features serve as excellent descriptors of speech prosody and speech styles. In this paper, instead of using the style tokens to control the TTS outputs, we would like to study how to use deep style features to measure perceptual loss for training of neural TTS system in general. While the idea is proposed for neural TTS, we use Tacotron TTS system as an example to carry out the study.

III. TACOTRON WITH FRAME AND STYLE RECONSTRUCTION LOSS

We propose a novel training strategy for Tacotron with both frame and style reconstruction loss. As the style reconstruction loss is formulated as a perceptual loss (PL) [34], the proposed frame and style training strategy is called Tacotron-PL in short. It seeks to optimize both frame-level spectral loss, that is frame reconstruction loss; as well as utterance-level style loss, that is style reconstruction loss, at the same time.

The overall framework is illustrated in Fig. 2, that has three stages: 1) training of style descriptor, 2) the proposed frame and style training for Tacotron-PL model, and 3) run-time inference. In Stage I, we train an auxiliary network to serve as the style descriptor for input speech utterances. In Stage II, the proposed frame and style training strategy is implemented to associate input text with acoustic features, as well as prosody style of natural speech, that is assisted by the style descriptor obtained from Stage I. In Stage III, the Tacotron-PL system takes input text and generates expressive speech in the same way as a standard Tacotron does. Unlike other Tacotron variants [28], Tacotron-PL doesn’t require any add-on module or process for run-time inference.

As discussed in Section II-A, traditional Tacotron architecture contains a text encoder and an attention-based decoder. We first encode input character embedding into hidden state, from which the decoder generates mel-spectrum features. During training, we adopt a frame-level mel-spectrum loss as in [5], which is a $L_2$ loss between the synthesized mel-spectrum $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_T\}$ and target mel-spectrum $Y = \{y_1, \ldots, y_T\}$. We have $Loss_{frame}$ as follows,

$$Loss_{frame}(Y, \hat{Y}) = \sum_{t=1}^{T} L_2(y_t, \hat{y}_t)$$ (1)

which is designed to minimize frame level distortion. As it doesn’t guarantee utterance level similarity concerning speech expressions, such as speech prosody and speech styles. We will study a new loss function $Loss_{style}$ next, that measures the utterance-level style reconstruction loss.

A. Stage I: Training of Style Descriptor

One of the great difficulties of prosody modelling is the lack of reference problem. In linguistics, we usually describe prosody styles qualitatively. However, precise annotation of
speech prosody is not straightforward. One of the ways to describe a prosody style is to show by example. The idea of style token [28] shows a way to compare two prosody styles quantitatively using deep features.

Manual prosodic annotations of recorded speech [26] provide quantifiable prosodic labels that allow us to associate speech styles with actual acoustic features. Prosody labelling schemes often attempt to describe prosodic phenomena, such as the supra-segmental features of intonation, stress, rhythm and speech rate, in discrete categories. Categorical labels of speech emotion [82] also seek to achieve a similar goal. The prosody labelling schemes serve as a type of style descriptor. With deep neural network, one is able to learn the feature representation of the data at different level of abstraction in a continuous space [83]. As speech styles naturally spread over a continuum rather than forced-fitting into a finite set of categorical labels, we believe that deep neural network learned from animated and emotive speech serves as a more suitable style descriptor.

We propose to use a speech emotion recognizer (SER) [75], [76] as a style descriptor $F(\cdot)$, which extracts deep style features $\Psi$ from an utterance $Y$, or $\Psi = F(Y)$. We use neuronal activations of hidden units in a deep neural network as the deep style features to represent high level prosodic abstraction at utterance level. In practice, we first train a SER network with highly animated and emotive speech with supervised learning. We then derive deep style features from a small intermediate layer. As the intermediate layer is small relative to the size of the other layers, it creates a constriction in the network that forces the information pertinent to emotion classification into a low dimensional prosody representation [84]. If the network classifies the emotion well, the so derived deep features are believed to describe well prosody style of speech.

We follow the SER implementation in [85] as illustrated in Fig. 3, that forms part of Fig. 2. The SER network includes 1) a three-dimensional (3-D) CNN layer; 2) a BLSTM layer [86]; 3) an attention layer; and 4) a fully connected (FC) layer. The 3-D CNN [85] first extracts a latent representation from mel-spectrum, its delta and delta-delta values from input utterance, converting the input utterance of variable length into a fixed size latent representation, denoted as deep features sequence $\Psi_{\text{low}}$, that reflects the semantics of emotion. The BLSTM summarizes the temporal information of $\Psi_{\text{low}}$ into another latent representation $\Psi_{\text{middle}}$. Finally, the attention layer assigns weights to $\Psi_{\text{middle}}$ and generates $\Psi_{\text{high}}$ for emotion prediction.

The question is which of the latent representations, $\Psi_{\text{low}}$, $\Psi_{\text{middle}}$, and $\Psi_{\text{high}}$, is suitable to be the deep style features. To validate the descriptiveness of deep style features, we perform an analysis on LJ-Speech corpus [87]. Specifically, we randomly select five utterances from each of the six style groups from the database, each group having a distinctive speech style, namely, 1) Short question; 2) Long question; 3) Short answer; 4) Short statement; 5) Long statement and 6) Digit string. The complete list of utterances can be found at Table III in Appendix A.

We visualize the $\Psi_{\text{low}}$, $\Psi_{\text{middle}}$ and $\Psi_{\text{high}}$ of utterances using the t-SNE algorithm in a two dimensional plane [88], as shown in Fig. 4. It is observed that $\Psi_{\text{low}}$, $\Psi_{\text{middle}}$ and $\Psi_{\text{high}}$ of utterances form clear style groups in terms of feature distributions, that is encouraging. We will further compare the performance of different deep style features through TTS experiments in Section IV.

B. Stage II: Tacotron-PL Training

During the training of Tacotron-PL, the SER-based style descriptor $F(\cdot)$ is used to extract the deep style features $\Psi$. We define a style reconstruction loss that compares the prosody style between the reference speech $Y$ and the generated speech $\hat{Y}$.

$$\text{Loss}_{\text{style}}(Y, \hat{Y}) = L_2(\Psi, \hat{\Psi})$$

(2)

where $\Psi = F(Y)$ and $\hat{\Psi} = F(\hat{Y})$. As illustrated in Fig. 3, the proposed training strategy involves two loss functions: 1) $\text{Loss}_{\text{frame}}$ that minimizes the loss between synthesized and original mel-spectrum at frame level; and 2) $\text{Loss}_{\text{style}}$ that minimizes the style differences between the synthesized and reference speeches at utterance level.

$$\text{Loss}_{\text{total}}(Y, \hat{Y}) = \text{Loss}_{\text{frame}}(Y, \hat{Y}) + \text{Loss}_{\text{style}}(Y, \hat{Y})$$

(3)
Fig. 4: t-SNE plot of the distributions of deep style features \( \Psi_{\text{low}} \), \( \Psi_{\text{middle}} \) and \( \Psi_{\text{high}} \) for six groups of utterances in LJ-Speech corpus. The list of utterances can be found at Table III in Appendix A.

where \( \text{Loss}_{\text{frame}} \) is also the loss function of a traditional Tacotron [5] system.

Style reconstruction loss can be seen as a perceptual quality feedback at utterance level to supervise the training of prosody style. All parameters in the TTS model are updated with the gradients of the total loss through back-propagation. We expect that mel-spectrum generation will learn from local and global viewpoint through the frame and style reconstruction loss.

C. Stage III: Run-time Inference

The inference stage follows exactly the same Tacotron workflow, that only involves the TTS Model in Fig. 3. The difference between Tacotron-PL and other global style tokens variation of Tacotron is that Tacotron-PL encodes prosody styling inside the standard Tacotron architecture. It doesn’t require any add-on module.

At run-time, the Tacotron architecture takes text as input and generate expressive mel-spectrum features as output, that is followed by Griffin-Lim algorithm [40] in this paper to generates audio signals.

IV. EXPERIMENTS

We train a SER as the style descriptor on IEMOCAP dataset [82], which consists of five sessions, each of which is displayed by a pair of speakers (female and male) in scripted and improvised scenarios. The dataset contains a total of 10,039 utterances, with an average duration of 4.5 seconds at a sampling rate of 16 kHz. We only use a subset of improvised data with four emotional categories, namely, happy, angry, sad, and neutral, which are recorded in hypothetical scenarios designed to elicit specific types of emotions.

With the style descriptor, we further train a Tacotron system on LJ-Speech database [87], which consists of 13,100 short clips with a total of nearly 24 hours of speech from one single speaker reading 7 non-fiction books. The speech samples are available from the demo link 1.

A. Comparative Study

We develop five Tacotron-based TTS systems for a comparative study, that include the Tacotron baseline, and four variants of Tacotron with the proposed training strategy, Tacotron-PL. To study the effect of different style descriptors, we compare the use of four deep style features, which includes three single features and a combination of them, in \( \text{Loss}_{\text{style}} \), as illustrated in Fig. 5, and summarized as follows:

- Tacotron: Tacotron [5] trained with \( \text{Loss}_{\text{frame}} \) as in Eq. (1).
- Tacotron-PL(L): Tacotron-PL which uses \( \Psi_{\text{low}} \) in \( \text{Loss}_{\text{style}} \).
- Tacotron-PL(M): Tacotron-PL which uses \( \Psi_{\text{middle}} \) in \( \text{Loss}_{\text{style}} \).
- Tacotron-PL(H): Tacotron-PL which uses \( \Psi_{\text{high}} \) in \( \text{Loss}_{\text{style}} \).
- Tacotron-PL(LMH): Tacotron-PL which uses \( \{ \Psi_{\text{low}}, \Psi_{\text{middle}}, \Psi_{\text{high}} \} \) in \( \text{Loss}_{\text{style}} \).

B. Experimental Setup

For SER training, we split the speech signals into segments of 3 seconds as in [85]. Then 40-channel mel-spectrum features are extracted with a frame size of 50ms and 12.5ms frame shift. The first convolution layer has 128 feature maps, while the remaining convolution layers have 256 feature map. The filter size for all convolution layers is 5\times5, with 5 along the time axis, and 3 along the frequency axis, and the pooling size for the max pooling layer is 2\times2. We add a linear layer with 200 output units after 3-D CNN for dimension reduction.

In this way, the 3-D CNN extracts a fixed size of latent representation with 150 \times 200 dimension from the input utterance, that we use as the deep style features \( \Psi_{\text{low}} = F_{\text{low}}(\cdot) \), which represents a temporal sequence of 150 segment, each having an embedding of 200 elements. As each direction

1 Speech Samples: https://ttslr.github.io/Expressive-TTS-Training-with-Frame-and-Style-Reconstruction-Loss/
of BLSTM layer contains 128 cells, in two directions, we obtain 256 output activations for each input segment, that are further mapped to 200 output units via a linear layer. BLSTM summarizes the temporal information of $\Psi_{\text{low}}$ into another fixed size latent representation $\Psi_{\text{middle}} = F_{\text{middle}}(\cdot)$ of $150 \times 200$ dimension. The attention layer assigns the weights to $\Psi_{\text{middle}}$ and generate a new latent representation $\Psi_{\text{high}} = F_{\text{high}}(\cdot)$. All latent representation $\Psi_{\text{low}}$, $\Psi_{\text{middle}}$, $\Psi_{\text{high}}$ have the same dimension.

The fully connected layer contains 64 output units. Batch normalization [89] is applied to the fully connected layer to accelerate training and improve the generalization performance. The parameters of the SER model were optimized by minimizing the cross-entropy objective function, with a minibatch of 40 samples, using the Adam optimizer with Nestorov momentum. The initial learning rate is set to $10^{-4}$ and the momentum is set to 0.9.

The SER-based style descriptor is used to extract deep style features for the computing of $\text{Loss}_{\text{style}}$. For TTS training, the encoder takes 256-dimensions character sequence as input and the decoder generates the 40-channel mel-spectrum. The training utterances from LJ-Speech database are of variable length. Mel-spectrum features are also extracted with a frame size of 50ms and 12.5ms frame shift. They are normalized to zero-mean and unit-variance to serve as the reference target.

D. Objective Evaluation

We conduct objective evaluation experiments to compare the systems in a comparative study. The results are summarized in Table I.

1) Performance Evaluation Metrics: Mel-cepstral distortion (MCD) [90] is used to measure the spectral distance between the synthesized and reference mel-spectrum features. MCD is calculated as:

$$
\text{MCD} = \frac{10\sqrt{2}}{\ln 10} \frac{1}{N} \sqrt{\sum_{k=1}^{N} (y_{t,k} - \hat{y}_{t,k})^2}
$$

where $N$ represents the dimension of the mel-spectrum, $y_{t,k}$ denotes the $k^{th}$ mel-spectrum component in $t^{th}$ frame for the reference target mel-spectrum, and $\hat{y}_{t,k}$ for the synthesized mel-spectrum. Lower MCD value indicates smaller distortion.

We use Root Mean Squared Error (RMSE) as the evaluation metrics for F0 modeling, that is calculated as:

$$
\text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (F0_t - \hat{F0}_t)^2}
$$

where $F0_t$ and $\hat{F0}_t$ denote the reference and synthesized F0 at $t^{th}$ frame. We note that lower RMSE value suggests that the two F0 contours are more similar.

Moreover, we propose to use frame disturbance, denoted as FD, to calculate the deviation in the dynamic time warping (DTW) alignment path [91]–[93]. FD is calculated as:

$$
\text{FD} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (a_{t,x} - a_{t,y})^2}
$$

where $a_{t,x}$ and $a_{t,y}$ denote the x-coordinate and the y-coordinate of the $t^{th}$ frame in the DTW alignment path. As FD represents the duration deviation of the synthesized speech from the target, it is a proxy to show the duration distortion. A larger value indicates poor duration modeling performance and a smaller value indicates otherwise.

2) Spectral Modeling: We observe that all implementations of Tacotron-PL model consistently provide lower MCD values than Tacotron baseline, with Tacotron-PL(L) representing the lowest MCD, as can be seen in Table I. We also visualize the spectrograms of same speech content synthesized by five different models, together with that of the reference natural speech in Fig. 7. A visual inspection of the spectrograms suggests that Tacotron-PL models consistently provide finer spectral details than Tacotron baseline. All results confirm the observations in Fig. 6, that Tacotron-PL training provides a better spectral model.
TABLE I: The MCD, RMSE and FD results of different systems.

| System       | MCD  [dB] | RMSE [Hz] | FD  [frame] |
|--------------|----------|-----------|-------------|
| Tacotron     | 7.01     | 1.33      | 15.59       |
| Tacotron-PL(L)| 6.37     | 0.94      | 13.96       |
| Tacotron-PL(M)| 6.70     | 1.21      | 14.20       |
| Tacotron-PL(H)| 6.88     | 1.42      | 15.41       |
| Tacotron-PL(LMH)| 6.62    | 1.16      | 14.15       |

3) F0 Modeling: Fundamental frequency, or F0, is an essential prosodic feature of speech [29], [32]. As there is no guarantee that synthesized speech and reference speech have the same length, we apply DTW [94] to align speech pairs and calculated RMSE between the F0 contour of them. The results are reported in Table I. It is observed that Tacotron-PL models consistently generate F0 contours which are closer to reference speech than Tacotron baseline.

We note that both F0 and prosody style contribute to RMSE measurement. To show the effect of various deep style features on the F0 contours, we also plot the F0 contours of the utterances in Fig. 7. A visual inspection suggests that the Tacotron-PL models benefit from the perceptual loss training, and produce F0 contour with a better fit to that of the reference speech, with Tacotron-PL(L) producing the best fit (see Fig. 7(c)).

4) Duration Modeling: Frame disturbance is a proxy to the duration difference [93] between synthesized speech and reference natural speech. We report frame disturbance of five systems in Table I. As shown in Table I, Tacotron-PL models obtain significantly lower FD value than Tacotron baseline, with Tacotron-PL(L) giving the lowest FD. From Figs. 6 and 7, we can also observe that Tacotron-PL(L) example clearly provides a better duration prediction than other models. We can conclude that perceptual loss training with style reconstruction loss helps Tacotron to achieve more accurate rendering of prosodic patterns.

5) Deep Style Features: We compare four different deep style features by evaluating the performance of their use in Tacotron-PL models, namely Tacotron-PL(L), Tacotron-PL(M), Tacotron-PL(H) and Tacotron-PL(LMH).

In supervised feature learning, the features that are near the input layer are related to the low level features, while those that are near the output are related to the supervision target, that are the categorical labels of the emotion. While we expect the style descriptors to capture utterance level prosody style, we don’t expect style reconstruction loss function to directly relate to emotion categories. Hence, lower level deep features, $\Psi_{low}$, as illustrated in Fig. 5, would be more appropriate than higher level deep features, such as $\Psi_{middle}$ and $\Psi_{high}$.

We observe that $\Psi_{low}$ is more descriptive than other deep style features for perceptual loss evaluation, as reported in spectral modeling (MCD), F0 modeling (RMSE), duration modeling (FD) for Tacotron-PL experiment in Table I. The observations confirm our intuition and the analysis in Fig. 4.

E. Subjective Evaluation

We conduct listening experiments to evaluate several aspects of the synthesized speech, and the choice of deep style features for $\text{Loss}_{style}$.

1) Voice Quality: Each audio is listened by 15 subjects, each of which listens to 90 synthesized speech samples. We first evaluate the voice quality in terms of mean opinion score (MOS) among Tacotron, Tacotron-PL(L), Tacotron-PL(M), Tacotron-PL(H), and Tacotron-PL(LMH). As shown in Fig. 8, Tacotron-PL models consistently outperforms Tacotron baseline, while Tacotron-PL(L) achieves the best result.

2) Expressiveness: In the objective evaluations and MOS listening tests, Tacotron-PL(L) and Tacotron-PL(LMH) consistently offer better results. We next focus on comparing Tacotron-PL(L) and Tacotron-PL(LMH) with Tacotron baseline. We first conduct the AB preference test to assess speech expressiveness of the systems. Each audio is listened by 15 subjects, each of which listens to 90 synthesized speech samples. Fig. 9 reports the speech expressiveness evaluation results. We note that Tacotron-PL(L) outperforms both Tacotron baseline and Tacotron-PL(LMH) in the preference test. The results suggest that $\Psi_{low}$ is more effective than other deep style features to inform the speech style.
that Tacotron-PL(LMH) outperforms both Tacotron baseline and Tacotron-PL(LMH) in the preference test. The results confirm that $\Psi_{low}$ is more effective to inform the speech style.

4) Deep Style Features: We finally conduct Best Worst Scaling (BWS) listening experiments to compare the four different Tacotron-PL systems with different deep style features. The subjects are invited to evaluate multiple samples derived from the different models, and choose the best and the worst sample. We perform this experiments for 18 different utterances, and each subject listens to 72 speech samples in total. Each audio is listened by 15 subjects.

Table II summarizes the results. We can see that Tacotron-PL(L) is selected for 83% of time as the best model and only 2% of time as the worst model, that shows $\Psi_{low}$ is the most effective deep style features.

V. CONCLUSION

We have studied a novel training strategy for Tacotron-based TTS system that include frame and style reconstruction loss. We implement a SER model as the style descriptor to extract deep style features to evaluate the style reconstruction loss. We have conducted a series of experiments and demonstrated that the proposed Tacotron-PL training strategy outperforms the start-of-the-art Tacotron baseline without the need of any add-on mechanism at run-time. While we conduct the experiments only on Tacotron, the proposed idea is applicable to other end-to-end neural TTS systems, that will be the future work in our plan.

APPENDIX A

TABLE II: Best Worst Scaling (BWS) listening experiments that compare four deep style features in four Tacotron-PL models.

| System       | Best (%) | Worst (%) |
|--------------|----------|-----------|
| Tacotron-PL(L)| 83       | 2         |
| Tacotron-PL(M)| 7        | 28        |
| Tacotron-PL(H)| 0        | 50        |
| Tacotron-PL(LMH)| 10     | 20        |

TABLE III: The scripts of utterances in six distinctive style groups from LJ-Speech database, the deep style features of which are visualized in Fig. 4.

| Group 1 (Digit String) | (1) What did he say to that?  
| (2) Where would be the use?  
| (3) Where is it?  
| (4) The soldiers then?  
| (5) What is my proposal? |
|------------------------|----------------------------------|
| Group 2 (Long Statement) | (1) Could you advise me as to the general view we have on the American Civil Liberties Union?  
| (2) Why not relieve Newgate by drawing more largely upon the superior accommodation which Millbank offered?  
| (3) Who ever heard of a criminal being sentenced to catch the rheumatism or the typhus fever?  
| (4) Why not move the city prison bodily into this more rural spot, with its purer air and greater breathing space?  
| (5) Great Britain in many ways has advanced further along lines of social security than the United States? |
| Group 3 (Short Answer) | (1) Answer: Yes.  
| (2) Answer: No.  
| (3) Answer: Thank you.  
| (4) Answer: No. sir.  
| (5) Answer: By not talking to him. |
| Group 4 (Short Statement) | (1) In September he began to review Spanish.  
| (2) They agree that Hosty told Revill.  
| (3) Hardly any one.  
| (4) They are photographs of the same scene.  
| (5) And other details in the picture. |
| Group 5 (Long Statement) | (1) I only know that his basic desire was to get to Cuba by any means, and that all the rest of it was window dressing for that purpose. End quote.  
| (2) He tried to start a conversation with me several times, but I would not answer. And he said that he didn’t want me to be angry at him because this upset him.  
| (3) Several of the publications furnished the Commission with the prints they had used, or described by correspondence the retouching they had done.  
| (4) From an examination of one of the photographs, the Commission determined the dates of the issues of the Militant and the Worker which Oswald was holding in his hand.  
| (5) He later wrote to another official of the Worker, seeking employment, and mentioning the praise he had received for submitting his photographic work. |
| Group 6 (Digit String) | (1) Nineteen sixty-three.  
| (2) Fourteen sixty-nine, fourteen seventy.  
| (3) March nine, nineteen thirty-seven. Part one.  
| (4) Section ten. March nine, nineteen thirty-seven. Part two  
| (5) On November eight, nineteen sixty-three. 
I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. The MIT Press, 2016.

D. Yu and M. L. Seltzer, “Improved bottleneck features using pretrained deep neural networks,” in Proc. Interspeech 2011, 2011, pp. 237–240.

M. Chen, X. He, J. Yang, and H. Zhang, “3-d convolutional recurrent neural networks with attention model for speech emotion recognition,” IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440–1444, 2018.

K. Greff, R. K. Srivastava, J. Koutn., B. R. Steunebrink, and J. Schmidhuber, “Lstm: A search space odyssey,” IEEE Transactions on Neural Networks & Learning Systems, vol. 28, no. 10, pp. 2222–2232, 2017.

K. Ito, “The lj speech dataset,” https://keithito.com/LJ-Speech-Dataset/, 2017.

L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” Journal of machine learning research, vol. 9, no. Nov, pp. 2579–2605, 2008.

S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International Conference on Machine Learning, 2015, pp. 448–456.

R. Kubichek, “Mel-cepstral distance measure for objective speech quality assessment,” in Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing, vol. 1. IEEE, 1993, pp. 125–128.

B. Sisman, G. Lee, H. Li, and K. C. Tan, “On the analysis and evaluation of prosody conversion techniques,” in 2017 International Conference on Asian Language Processing (IALP), 2017, pp. 44–47.

A. Z. Jusoh, R. Togneri, S. Nordholm, N. Sulaiman, and M. H. Khairoolanuar, “The investigation of frame disturbance (fd) in perceptual evaluation speech quality (pesq) as a perceptual metric,” ARPN Journal of Engineering and Applied Sciences, vol. 10, no. 15, pp. 6365–6369, 2015.

C. Gupta, H. Li, and Y. Wang, “Perceptual evaluation of singing quality,” in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2017, pp. 577–586.

M. Müller, “Dynamic time warping,” Information retrieval for music and motion, pp. 69–84, 2007.

Guanglai Gao received the B.S. degree from Inner Mongolia University, Hohhot, China, in 1985, and the M.S. degree from the National University of Defense Technology, Changsha, China, in 1988. He was a Visiting Researcher at University of Montreal, Canada. Currently, he is a Professor with the Department of Computer Science, Inner Mongolia University. His research interests include artificial intelligence and pattern recognition.

Haizhou Li Haizhou Li (M91-SM01-F14) received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently a Professor at the Department of Electrical and Computer Engineering, National University of Singapore (NUS). His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Prior to joining NUS, he taught in the University of Hong Kong (1988-1990) and South China University of Technology (1990-1994). He was a Visiting Professor at CRIN in France (1994-1995), Research Manager at the Apple-ISS Research Centre (1996-1998), Research Director in Lernout & Hauspie Asia Pacific (1999-2001), Vice President in InfoTalk Corp. Ltd. (2001-2003), and the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003-2016). Dr Li served as the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015-2018), a Member of the Editorial Board of Computer Speech and Language (2012-2018). He was an elected Member of IEEE Speech and Language Processing Technical Committee (2013-2015), the President of the International Speech Communication Association (2015-2017), the President of Asia Pacific Signal and Information Processing Association (2015-2016), and the President of Asian Federation of Natural Language Processing (2017-2018). He was the General Chair of ACL 2012, INTERSPEECH 2014 and ASRU 2019. Dr Li is a Fellow of the IEEE and the ISCA. He was a recipient of the National Infocomm Award 2002 and the Presidents Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and Bremen Excellence Chair Professor in 2019.