Modeling Prosodic Phrasing With Multi-Task Learning in Tacotron-Based TTS

Rui Liu , *Member, IEEE*, Berrak Sisman , *Member, IEEE*, Feilong Bao, Guanglai Gao, and Haizhou Li , *Fellow, IEEE*

Abstract—Tacotron-based end-to-end speech synthesis has shown remarkable voice quality. However, the rendering of prosody in the synthesized speech remains to be improved, especially for long sentences, where prosodic phrasing errors can occur frequently. In this letter, we extend the Tacotron-based speech synthesis framework to explicitly model the prosodic phrase breaks. We propose a multi-task learning scheme for Tacotron training, that optimizes the system to predict both Mel spectrum and phrase breaks. To our best knowledge, this is the first implementation of multi-task learning for Tacotron based TTS with a prosodic phrasing model. Experiments show that our proposed training scheme consistently improves the voice quality for both Chinese and Mongolian systems.

Index Terms—Tacotron, multi-task learning, prosody.

I. INTRODUCTION

WITH the advent of deep learning, end-to-end text-to-speech (TTS) has shown many advantages over the conventional TTS techniques [1], [2]. Tacotron-based approaches [3]–[7] with an encoder-decoder architecture and attention mechanism have shown remarkable performance. The key idea is to integrate the conventional TTS pipeline into a unified network and learn the mapping directly from the text-waveform pair [8]–[10]. The recent progress in neural vocoder [4], [11]–[15] also contributes to the improvement of speech quality.

Rui Liu is with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China, and also with the National University of Singapore, Singapore 117583, Singapore (e-mail: liurui_imu@163.com).

Berrak Sisman is with the Information Systems Technology and Design (ISTD) Pillar, Singapore University of Technology and Design, Singapore 487372 (e-mail: berraksisman@u.nus.edu).

Feilong Bao and Guanglai Gao are with the Department of Computer Science, Inner Mongolia University, Hohhot 010021, China (e-mail: csfeilong@imu.edu.cn; csggl@imu.edu.cn).

Haizhou Li is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583 (e-mail: haizhou.li@nus.edu.sg).

Digital Object Identifier 10.1109/LSP.2020.3016564

Speech prosody includes affective prosody and linguistic prosody. Affective prosody represents the emotion of a speaker, while linguistic prosody relates to the language content. They are both crucial in speech communication. A TTS system is expected to synthesize the right prosodic pattern at the right time. However, most of the current end-to-end systems [3], [4], [6], [7] have not explicitly modeled speech prosody. Therefore, they can't control well the melodic and rhythmic aspects of the generated speech. This usually leads to monotonous speech, even when models are trained on very expressive speech datasets. In this letter, we would like to study the way to enable Tacotron-based TTS for expressive prosody generation.

Multi-task learning (MTL) is a learning paradigm that leverages information from multiple related tasks to help improve the overall performance [16]. MTL is inspired by human learning activities where people often apply the knowledge learned from many tasks for learning a new task, that is called inductive transfer. For example, if we learn to read and write together, the experience in reading can strengthen the writing and vice versa. MTL has been widely used in speech enhancement [17], and speech recognition [18]. It has also been used in speech synthesis [19], such as statistical parametric speech synthesis with GANs [20] and DNN-based speech synthesis with stacked bottleneck features [21]. In this letter, we apply multi-task learning to the Tacotron-based TTS for prosody modeling.

The study on expressive speech synthesis is focused on prosody modeling [22]–[24], where speech prosody generally refers to intonation, stress, speaking rate, and phrase breaks. Prosodic phrasing [25]–[28] plays an important role in both affective and linguistic expressions. Inadequate phrase breaks may lead to misperception in speech communication. There have been recent studies on prosody modeling for end-to-end TTS system [29], for example, to improve the prosodic phrasing by using contextual information [30], and syntactic features [31]. They are incorporated in the stage of text preprocessing, therefore, there are not optimized as part of the synthesis processing.

We propose a novel two-task learning scheme for Tacotron-based TTS model to improve the prosodic phrasing: 1) the main task learns the prediction of the speech spectrum parameters from character-level embedding representation, and 2) the secondary task learns the prediction of a word-level prosody embedding. During training, the secondary task serves as an additional supervision for Tacotron to learn the exquisite prosody structure associated with the input text. At run-time, the prosody embedding serves as a local condition that controls the prosodic phrasing during voice generation.

The main contributions of this letter include: 1) a novel Tacotron-based TTS architecture that explicitly models prosodic phrasing; and 2) a multi-task learning scheme, that optimizes the
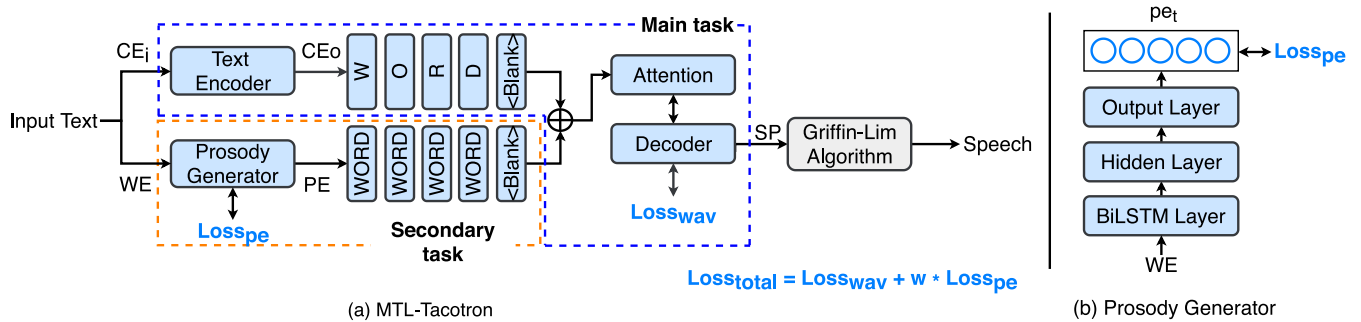
Fig. 1. Block diagrams of the proposed (a) MTL-Tacotron and (b) prosody generator. MTL-Tacotron employs a prosody generator to explicitly model prosodic phrasing. The prosody generator produces prosody embedding (*PE*) from input text, that forms a joint embedding vector with character embedding. Griffin-Lim algorithm is not involved in training. SP denotes mel-spectrum speech features. pe_t is a 5-dimension prosody embedding.

model for high quality speech spectrum, and adequate prosodic phrasing at the same time. The proposed system achieves remarkable voice quality for both Chinese Mandarin and Mongolian. To our best knowledge, this is the first multi-task Tacotron implementation that includes an explicit prosodic model.

This letter is organized as follows. Section II recaps the Tacotron TTS framework. We propose the multi-task Tacotron in Section III and report the experiments in Section IV. Section V concludes the discussion.

II. TACOTRON-BASED TTS

Tacotron [3] is a sequence-to-sequence speech synthesizer that consists of an encoder, and a decoder with attention mechanism. The encoder takes a sequence of text characters as input, where each character is encoded as a one-hot vector and embedded into a continuous vector, that is called input character embedding. The encoder is trained as part of Tacotron to take the input character embedding and generates the output character embedding. The decoder is an autoregressive recurrent neural network that converts the character embeddings into a sequence of Mel spectrum feature vectors with an attention mechanism.

Just like most of other TTS systems, Tacotron [3] is trained to predict the Mel spectrum features from input sequence of characters. Prosody, if taken into consideration, is modeled from the statistics of the training data [3], [4], [6], [8]. We note that the character sequences themselves are not the most suitable for describing prosody. They do not generalize well because prosody is manifested over a speech segment beyond characters and phonemes. There have been attempts [8], [32] to use word embedding as input to improve the expressiveness of Tacotron-based TTS model, that shows word embedding is prosody-informing.

Another idea [33]–[37] is to extract the latent prosody embeddings to characterize prosody. Some [33]–[35] learn speech variations without explicit annotations for prosody or style. The learned prosody embeddings are usually not fully controllable and interpretable. Others [36], [37] just take the prosody embeddings as an auxiliary input to the TTS model.

In this letter, we propose a novel prosody embedding, that directly interprets the phrase breaks from the input text. We also propose a novel multi-task learning framework that optimizes the system to generate Mel spectrum, at the same time, accurately predict phrase breaks, which will be the focus of Section III.

III. TACOTRON WITH MULTI-TASK LEARNING

We propose multi-task learning [38] for Tacotron as illustrated in Fig. 1(a), that is referred to as *MTL-Tacotron*. The idea is to dedicate a prosody modeling task to model the prosodic phrasing, that is trainable from data. In the multi-task learning, not only do we optimize the output speech quality, but we also ensure that Tacotron is optimized to produce adequate phrase breaks. We study a two-task learning strategy, 1) the *main task* generates the Mel-spectrums from the input character sequence; and 2) the *secondary task* predicts an appropriate prosodic phrasing.

A. Main Task: Spectral Modeling

The main task has a network architecture identical to the traditional Tacotron [3] as shown in Fig. 1(a). It contains a text encoder and a decoder with attention mechanism. We first convert the word sequence in raw input text to input character embeddings, denoted as CE_i, which are encoded into output character embeddings, denoted as CE_o, from which the decoder generates Mel spectral features.

The main task is optimized using a Mel spectral loss function, $Loss_{wav} = \sum_{t=1}^{T'} L_2(y_t, y_t')$, where y_t and y_t' are the target and the predicted Mel-spectrum respectively, T' is the total number of Mel spectrum features in the utterance, and L_2 denotes a L_2 norm function.

B. Secondary Task: Prosody Modeling

The secondary task optimizes a prosody generator to predict the phrase break pattern for each word in the input text, as shown in Fig. 1(b). We define prosody embedding (*PE*) as a vector of five elements, namely *break, non-break, blank, punctuation* and *stop token*, that represents five phrase break patterns. *stop token* denotes the end of an utterance, while *punctuation* refers to any punctuation symbols other than *stop token*.

The word sequence in the input text is first represented by a sequence of word embeddings. We devise a Bidirectional Long Short-Term Memory (BLSTM) as the prosody generator, that takes the word embeddings $WE = \{we_1, \ldots, we_t, \ldots, we_T\}$ as input and generates the prosody embedding $PE = \{pe_1, \ldots, pe_t, \ldots, pe_T\}$ as output, where pe_t is a 5-dimension embedding vector. Specifically, the forward and backward LSTM reads the word embedding sequence WE from both directions. We add a hidden layer on top of the LSTM to detect higher-level feature combinations, and a softmax layer to produce the probability distribution of phrase break patterns

pe_t for each of the T words. An element in the embedding vector $pe_t = [p_t[1], \ldots, p_t[k], \ldots, p_t[5]]$, $t \in [1, T]$, represents the probability of the phrase break label k.

The secondary task minimizes the differences between the predicted prosody embedding and the ground truth one-hot vector using the cross-entropy loss $Loss_{pe} = -\sum_{t=1}^{T} \log p_t[k]$, where k represents the target phrase break pattern.

C. Multi-Task Learning

We now have two parallel feature representations of input text, as shown in Fig. 1(a). CE_o is the character representation in the main task and the prosody embedding (PE) in the secondary task. Prosody embedding serves as an auxiliary input to Tacotron that informs the phrase break information. We concatenate CE_o and PE to form a joint embedding vector as the input for the attention mechanism. In this way, we expect that Tacotron optimizes the voice quality, by also making sure that the phrase break is correct. As CE_o and PE have different time resolutions, we upsample the PE to align with CE_o as shown in Fig. 1(a).

The total loss function is given as $Loss_{total} = Loss_{wav} + w * Loss_{pe}$, with w as a weight. With the total loss, we expect that the prosody generator learns from both the phrase break annotations and the actual speech utterances to associate acoustic-prosodic patterns with the input text, thus improving the Tacotron spectral generation at run-time inference. The two-task learning strategy is also referred to as the joint training strategy.

IV. EXPERIMENTS

A. Databases

Speech Data: We use the TsingHua-Corpus of Speech Synthesis (TH-CoSS) [39] for Chinese. We use a subset of TH-CoSS, denoted as *03FR00*, that contains approximately 9 hours of speech data and 5.6 k utterances with 103 k words. The speech signals are sampled at 16 kHz and encoded at 16-bit. The Mongolian speech data as in [40] contains about 17 hours of speech in total. The speech signals are sampled at 22.05 kHz and encoded at 16-bit. For both Chinese and Mongolian, we divide the corpus into training and test sets in a ratio of 4 to 1 in all experiments.

Phrase Break Labels: We use the text transcript of the speech data as the training data of prosody generator. The prosodic phrases of Chinese text, *break* and *non-break*, are manually labelled. The Mongolian phrase breaks are marked by examining the text and listening to the speech samples. The *blank*, *punctuation* and *stop token* labels are naturally present in the text.

Word Embedding: We generate the word embedding WE via table look-up. For Chinese, we use the Tencent AI Lab embedding database for Chinese Words and Phrases [41]. For Mongolian, the pre-trained 200-dimension word embedding reported in [42] is used.

B. Contrastive Systems

We build three contrastive systems to validate the two ideas in the proposed *MTL-Tacotron*, namely multi-task learning, and prosody embedding in a comparative study. In all systems, we use Griffin-Lim algorithm [43] for waveform generation for rapid turn-around.

1) Traditional *Tacotron* TTS system as in [3], that doesn't explicitly model prosodic phrasing.
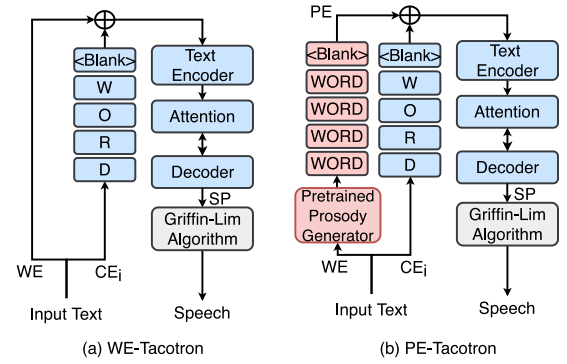

Fig. 2. Block diagrams of the baseline frameworks (a) WE-Tacotron, and (b) PE-Tacotron. Griffin-Lim algorithm is not involved in training. SP denotes mel-spectrum speech features.

2) Tacotron augmented with word embedding as in [8], denoted as *WE-Tacotron* and illustrated in Fig. 2(a). The word embedding informs *Tacotron* the word identity and its boundaries, that is shown effective [8].

3) Tacotron augmented with prosody embedding without multi-task joint training, denoted as *PE-Tacotron* and illustrated in Fig. 2(b). The prosody embeddings are derived from word embedding to encode the prosodic phrasing.

Besides the multi-task learning, *MTL-Tacotron* is also different from both *WE-Tacotron* and *PE-Tacotron* in the way that the text encoder is incorporated in order to facilitate the joint training. *PE-Tacotron* and *WE-Tacotron* share similar architecture with *Tacotron* baseline except that *PE-Tacotron* is augmented by prosody embedding, while *WE-Tacotron* is augmented by word embedding. Unlike *MTL-Tacotron*, they incorporate the embeddings that are trained independently of *Tacotron*. They are the contrastive models for *MTL-Tacotron* to show the effect of multi-task learning.

We note that prosody embedding is derived from word embedding. *MTL-Tacotron* and *PE-Tacotron* are trained to predict the phrase breaks explicitly from word embeddings, while *WE-Tacotron* use the word embeddings directly. Therefore, *WE-Tacotron* serves as the contrastive model for *MTL-Tacotron* and *PE-Tacotron* to show the advantage of the proposed prosody embedding.

C. Experimental Setup

The Chinese text is encoded in *Pinyin* string with tones, and Mongolian text in Latin transliteration. For both languages, we generate 80-channel Mel-spectrum as output. CE_i (or CE_o) and PE size are set to 256 and 5 respectively. The number of output frames is controlled by a hyperparameter reduction factor (r), which is set to 5, and the weight w is set to 0.5. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 10^{-3} exponentially decaying to 10^{-5} starting with 50 k steps. We also apply L_2 regularization with weight 10^{-6}. All models are trained with a batch size of 32. The final models are trained with 200 k steps for all systems.

The prosody generator in the *MTL-Tacotron* is jointly trained with other Tacotron modules. The size of the LSTM layer is set to 200 in both directions for all experiments, the size of hidden layer is set to 50. The prosody generator in *PE-Tacotron* has exactly the same configuration as that in *MTL-Tacotron*, except that it is pre-trained on the training data.

TABLE I
COMPARISON OF PHRASE BREAK PREDICTION IN TERMS OF PRECISION (P),
RECALL (R) AND F-SCORE (F) FOR TWO SYSTEMS THAT EMPLOY PROSODY
EMBEDDING, AND MEAN OPINION SCORE (MOS) IN LISTENING TESTS FOR
ALL SYSTEMS

System	Language	P	R	F	MOS
Tacotron	Chinese	NA	NA	NA	3.71 ± 0.04
	Mongolian	NA	NA	NA	3.60 ± 0.02
WE-Tacotron	Chinese	NA	NA	NA	3.79 ± 0.03
	Mongolian	NA	NA	NA	3.72 ± 0.02
PE-Tacotron	Chinese	90.01	90.88	90.68	3.86 ± 0.02
	Mongolian	88.83	89.42	89.11	3.79 ± 0.04
MTL-Tacotron	Chinese	90.77	91.54	**91.39**	**3.91** ± 0.01
	Mongolian	90.01	90.79	**90.33**	**3.83** ± 0.03

D. Phrase Break Prediction

We report the phrase break prediction performance of the prosody generator in *MTL-Tacotron* and *PE-Tacotron* where prosody embedding is used. As the text in the datasets has already been annotated with prosody labels, it serves as the ground truth for reporting the performance. At run-time inference, the phrase break pattern of a word we_t is predicted as $\hat{k} = \arg\max_k p_t[k]$. We report the performance in terms of Precision (P), Recall (R) and F-score (F) which is defined as the harmonic mean of the P and R. F values range from 0 to 1, with a higher value indicating better performance.

As shown in Table I, *MTL-Tacotron* clearly outperforms *PE-Tacotron* in phrase break prediction. By comparing *MTL-Tacotron* and *PE-Tacotron*, we confirm the advantage of joint training over pre-trained prosody embedding. We expect that *MTL-Tacotron* will reflect the improved phrase break prediction into actual prosodic rendering in speech.

E. Subjective Listening Test

We conduct listening experiments for all systems.[1] 20 Chinese and 15 Mongolian speakers participated in the listening tests. Each subject listens to 80 converted utterances of his/her native language.

1) Voice Quality: We first evaluate the voice quality with mean opinion score (MOS) among these four systems. The listeners rate the quality on a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. In Table I, we observe that *PE-Tacotron* and *MTL-Tacotron* consistently outperform traditional *Tacotron* that doesn't explicitly model prosodic phrasing. The results validate the idea of prosody embedding. Moreover, *MTL-Tacotron* outperforms *WE-Tacotron* and *PE-Tacotron* consistently that confirms the advantage of the proposed joint training.

2) Prosodic Embedding vs. Word Embedding: To confirm the advantage of prosody embedding over word embedding [8], we further conduct ABX preference tests between pairs of systems. The subjects are asked to choose their preferred utterances in terms of the rhythm and prosody break between a pair of synthesized utterances. The results in Table II suggest that *MTL-Tacotron* system with prosodic phrasing significantly outperforms others in both Chinese and Mongolian experiments.

We also observe that both *MTL-Tacotron* and *PE-Tacotron* outperform *WE-Tacotron* system. As *MTL-Tacotron* and *PE-Tacotron* model the phrase breaks explicitly, the results suggest

[1]Speech samples in the listening tests: https://ttslr.github.io/SPL2020

TABLE II
THE PREFERENCE PERCENTAGE (%) WITH 95% CONFIDENCE INTERVAL SIX
COMPETING PAIRS ON COMMON TEST DATA

| Competing pair	Language	Preference(%)			*p*-value
		Former	Neutral	Latter	
Tacotron vs. WE-Tacotron	Chinese	30.56	28.56	41.19	0.00105
	Mongolian	29.00	24.92	46.08	0.00014
Tacotron vs. PE-Tacotron	Chinese	29.38	27.18	43.44	0.00248
	Mongolian	27.33	24.42	48.25	0.00134
Tacotron vs. MTL-Tacotron	Chinese	27.44	26.25	46.31	0.00176
	Mongolian	25.91	21.01	53.08	0.00054
WE-Tacotron vs. MTL-Tacotron	Chinese	39.56	9.75	50.69	0.00392
	Mongolian	37.42	11.41	51.17	0.00217
WE-Tacotron vs. PE-Tacotron	Chinese	38.81	9.69	51.50	0.00282
	Mongolian	37.33	12.00	50.67	0.00153
PE-Tacotron vs. MTL-Tacotron	Chinese	40.06	12.69	47.25	0.00047
	Mongolian	41.50	9.92	48.58	0.00318
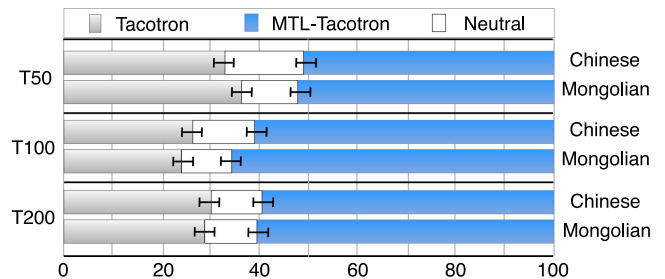
Fig. 3. The preference percentage (%) with 95% confidence interval between *Tacotron* and *MTL-Tacotron* for various text length.

that modeling phrase breaks explicitly is more effective than using word embedding as a proxy to inform the prosody [8]. As *MTL-Tacotron* consistently offers superior performance, we are convinced that multi-task learning improves the accuracy of the prosody model over *PE-Tacotron*, thereby generating more accurate prosody embeddings.

3) Effect of Text Length: We further investigate how the systems perform with regard to the length of input text. By grouping the test sentences by length, we create three subsets: 1) T50 with sentences up to 50 characters; 2) T100 with sentences of 51 to 100 characters; and 3) T200 with sentences of 101 to 200 characters. We select 80 utterances from each group for evaluation of expressiveness, and report the subjective listening test in Fig. 3. We observe that *MTL-Tacotron* consistently outperforms the *Tacotron* baseline. It is worth noting that *MTL-Tacotron* performs remarkably well for long sentences for T100 and T200, which is encouraging.

V. CONCLUSION

We have proposed a novel multi-task Tacotron model to model the prosodic phrasing in speech synthesis, where a word-level prosody generator is introduced as the secondary task. The experiments show that the proposed *MTL-Tacotron* consistently outperforms all contrastive systems. The modeling technique for prosodic phrasing can be easily extended to the modeling of other melodic and rhythmic aspects of speech, such as intonation and stress.

REFERENCES

[1] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proc. IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[2] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7962–7966.

[3] Y. Wang *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[4] J. Shen *et al.*, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 4779–4783.

[5] Y. Lee and T. Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5911–5915.

[6] R. Liu, B. Sisman, J. Li, F. Bao, G. Gao, and H. Li, "Teacher-student training for robust tacotron-based TTS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6274–6278.

[7] R. Liu, B. Sisman, F. Bao, G. Gao, and H. Li, "WaveTTS: Tacotron-based TTS with joint time-frequency domain loss," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2020, pp. 245–251.

[8] Y.-A. Chung, Y. Wang, W.-N. Hsu, Y. Zhang, and R. Skerry-Ryan, "Semi-supervised training for improving data efficiency in end-to-end speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6940–6944.

[9] M. He, Y. Deng, and L. He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural TTS," in *Proc. Interspeech*, 2019, pp. 1293–1297.

[10] H.-T. Luong, X. Wang, J. Yamagishi, and N. Nishizawa, "Training multi-speaker neural text-to-speech systems using speaker-imbalanced speech corpora," in *Proc. Interspeech*, 2019, pp. 1303–1307.

[11] T. Hayashi, A. Tamamori, K. Kobayashi, K. Takeda, and T. Toda, "An investigation of multi-speaker training for WaveNet vocoder," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 712–718.

[12] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders," in *Proc. Interspeech*, 2019, pp. 1308–1312.

[13] B. Sisman, M. Zhang, and H. Li, "A voice conversion framework with tandem feature sparse representation and speakera-adapted WaveNet vocoder," in *Proc. Interspeech*, 2018, pp. 1978–1982.

[14] B. Sisman, M. Zhang, and H. Li, "Group sparse representation with WaveNet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 6, pp. 1085–1097, Jun. 2019.

[15] B. Sisman, M. Zhang, S. Sakti, H. Li, and S. Nakamura, "Adaptive wavenet vocoder for residual compensation in GAN-based voice conversion," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 282–289.

[16] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.

[17] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3274–3278.

[18] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 4835–4839.

[19] Q. Hu, Z. Wu, K. Richmond, J. Yamagishi, Y. Stylianou, and R. Maia, "Fusion of multiple parameterisations for DNN-based sinusoidal speech synthesis with multi-task learning," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 854–858.

[20] S. Yang *et al*, "Statistical parametric speech synthesis using generative adversarial networks under a multi-task learning framework," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 685–691.

[21] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4460–4464.

[22] I. Jauk, J. Lorenzo Trueba, J. Yamagishi, and A. Bonafonte Cávez, "Expressive speech synthesis using sentiment embeddings," in *Proc. Interspeech*, 2018, pp. 3062–3066.

[23] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Proc. Interspeech*, 2018, pp. 3067–3071.

[24] Y. Mass *et al.*, "Word emphasis prediction for expressive text to speech." in *Proc. Interspeech*, 2018, pp. 2868–2872.

[25] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *J. Acoust. Soc. Amer.*, vol. 91, no. 3, pp. 1707–1717, 1992.

[26] H. Kim, T. Yoon, J. Cole, and M. Hasegawa-Johnson, "Acoustic differentiation of L- and L-L% in switchboard and radio news speech," in *Proc. Speech Prosody*, 2006, p. 214.

[27] P. Taylor and A. W. Black, "Assigning phrase breaks from part-of-speech sequences," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 99–117, 1998.

[28] T. Mishra, Y.-j. Kim, and S. Bangalore, "Intonational phrase break prediction for text-to-speech synthesis using dependency relations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 4919–4923.

[29] R. Sloan, S. S. Akhtar, B. Li, R. Shrivastava, A. Gravano, and J. Hirschberg, "Prosody prediction from syntactic, lexical, and word embedding features," in *Proc. 10th ISCA Speech Synthesis Workshop*, 2019, pp. 269–274.

[30] Y. Lu, M. Dong, and Y. Chen, "Implementing prosodic phrasing in chinese end-to-end speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7050–7054.

[31] H. Guo, F. K. Soong, L. He, and L. Xie, "Exploiting syntactic features in a parsed tree to improve end-to-end TTS," in *Proc. Interspeech*, 2019, pp. 4460–4464.

[32] H. Ming, L. He, H. Guo, and F. K. Soong, "Feature reinforcement with word embedding and parsing information in neural TTS," 2019, *arXiv:1901.00707*.

[33] Y. Wang *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5180–5189.

[34] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 595–602.

[35] R. Skerry-Ryan *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 94–99.

[36] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6905–6909.

[37] G. Sun *et al.*, "Generating diverse and natural text-to-speech samples using a quantized fine-grained VAE and auto-regressive prosody prior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6699–6703.

[38] Caruana and R. A., "Multitask learning: A knowledge-based source of inductive bias," in *Proc. Int. Conf. Mach. Learn.*, 1993, pp. 41–48.

[39] L. Cai, D. Cui, and R. Cai, "TH-CoSS, a Mandarin speech corpus for TTS," *J. Chin. Inf. Process.*, vol. 21, no. 2, pp. 94–99, 2007.

[40] J. Li, H. Zhang, R. Liu, X. Zhang, and F. Bao, "End-to-end Mongolian text-to-speech system," in *Proc. 11st Int. Symp. Chin. Spoken Lang. Process.*, 2018, pp. 3062–3066.

[41] Y. Song, S. Shi, J. Li, and H. Zhang, "Directional skip-gram: Explicitly distinguishing left and right context for word embeddings," in *Proc. 16th Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2018, pp. 175–180.

[42] R. Liu, F. Bao, G. Gao, and W. Wang, "Improving Mongolian phrase break prediction by using syllable and morphological embeddings with BiLSTM model," in *Proc. Interspeech*, 2018, pp. 57–61.

[43] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 2, pp. 236–243, Apr. 1984.