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 paper, 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.

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 paper, 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 paper, 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 paper include: 1) a novel Tacotron-based TTS architecture that explicitly models prosodic phrasing; and 2) a multi-task learning scheme, that optimizes the 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 paper is organized as follows. Section II recap 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 paper, 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 $CE_t$, 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 $WE = \{we_1, ..., we_t, ..., we_T\}$ as input and generates the prosody embedding $PE = \{pe_1, ..., pe_t, ..., pe_T\}$ 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 $pe_t = [p_t[1], ..., p_t[k], ..., p_t[5]]$, $t \in [1, T]$, 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 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 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 and PE have different time resolutions, we upsample the PE to align with CE, as shown in Fig. 1(a).

The total loss function is given as \( \text{Loss}_{\text{total}} = \text{Loss}_{\text{wav}} + w \ast \text{Loss}_{\text{pe}} \), 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 5.6k utterances with 103k 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 for 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 \( W \) 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.

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 (or CE), 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 50k 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 200k 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.

D. Phrase Break Prediction

We report the phrase break prediction performance of the prosody generator in MTL-Tacotron and PE-Tacotron where
Table I

| System      | Language | P  | R  | F   | MOS     |
|-------------|----------|----|----|-----|---------|
| Tacotron    | Chinese  | NA | NA | NA  | 3.71 ± 0.04 |
|             | Mongolian| NA | NA | NA  | 3.69 ± 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 | 90.08 | 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 | 3.85 ± 0.03  |

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 \( w_{ei} \) is predicted as \( \hat{k} = \arg \max_k p_k[k] \). 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 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. Conclusions

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.

\(^1\) Speech samples in the listening tests: https://ttslr.github.io/SPL2020

Fig. 3. The preference percentage (%) with 95% confidence interval between Tacotron and MTL-Tacotron for various text length.
