Improving Sequence-to-sequence Tibetan Speech Synthesis with Prosodic Information

WEIZHAO ZHANG, College of Physics and Electronic Engineering, Northwest Normal University, China
HONGWU YANG, School of Educational Technology, Northwest Normal University, China

There are about 6,000 languages worldwide, most of which are low-resource languages. Although the current speech synthesis (or text-to-speech, TTS) for major languages (e.g., Mandarin, English, French) has achieved good results, the voice quality of TTS for low-resource languages (e.g., Tibetan) still needs to be further improved. Because prosody plays a significant role in natural speech, the article proposes two sequence-to-sequence (seq2seq) Tibetan TTS models with prosodic information fusion to improve the voice quality of synthesized Tibetan speech. We first constructed a large-scale Tibetan corpus for seq2seq TTS. Then we designed a prosody generator to extract prosodic information from the Tibetan sentences. Finally, we trained two seq2seq Tibetan TTS models by fusing prosodic information, including feature-level and model-level prosodic information fusion. The experimental results showed that the proposed two seq2seq Tibetan TTS models, which fuse prosodic information, could effectively improve the voice quality of synthesized speech. Furthermore, the model-level prosodic information fusion only needs 60% ~ 70% of the training data to synthesize a voice similar to the baseline seq2seq Tibetan TTS. Therefore, the proposed prosodic information fusion methods can improve the voice quality of synthesized speech for low-resource languages.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Modeling and simulation;**

Additional Key Words and Phrases: Sequence-to-sequence speech synthesis, Tibetan speech synthesis, prosodic information fusion, low-resource language

ACM Reference format:
Weizhao Zhang and Hongwu Yang. 2023. Improving Sequence-to-sequence Tibetan Speech Synthesis with Prosodic Information. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22, 9, Article 225 (September 2023), 13 pages.
https://doi.org/10.1145/3616012

1 INTRODUCTION

Speech synthesis (or text-to-speech synthesis, TTS), a process of converting arbitrary texts into speech, can allow the computer to "speak". Advanced TTS has been widely used in various

The research is supported by the research fund from the National Natural Science Foundation of China (Grant No. 62067008, No. 11664036, No. 62267008). Additionally, part of this work is also supported by the Science and Technology Program of Gansu Province (Grant No. 20JR10RA095, No. 21JR7RA117) and the University Innovation Foundation of Gansu Province (Grant No. 2022B-091, No. 2023B-239).
Authors' addresses: W. Zhang, College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China; email: zhangweizhao@nwnu.edu.cn; H. Yang (corresponding author), School of Educational Technology, Northwest Normal University, Lanzhou, China; email: yanghw@nwnu.edu.cn.
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2375-4699/2023/09-ART225 $15.00
https://doi.org/10.1145/3616012

application scenarios such as chatbots, audiobooks, or intelligent speech interaction with the rapid development of deep learning technologies. However, there is still a challenge before computers can generate natural speech with high naturalness and expressiveness like humans.

TTS technology has roughly gone through three phases over time, including unit selection-based concatenative speech synthesis [1], **hidden Markov models** (**HMMs**)-based **statistical parametric speech synthesis** (**SPSS**) [2], and deep learning-based speech synthesis [3–5]. Unit selection-based concatenative speech synthesis generates the most natural-sounding speech by concatenating small units of prerecorded waveforms with a unit selection algorithm, but its flexibility is limited. HMM-based SPSS has been the most popular TTS technology since the 1990s. Generally speaking, the SPSS pipeline is usually composed of three modules: a complex text frontend to extract various linguistic features from raw text, a parameter prediction module that learns the transformation between linguistic features and acoustic features such as fundamental frequency (F0), spectral parameters and duration, and a complex signal-processing-based vocoder to reconstruct waveform from the predicted acoustic features. The main advantage of HMM-based SPSS over concatenative speech synthesis is its flexibility in changing speech features, speaking styles, and emotions. This flexibility is mainly attributed to applying many techniques for controlling variation in speech, such as adaptation [6], interpolation [7], and eigen-voice [8]. However, the synthesized speech of HMM-based SPSS is muffled compared with natural speech [9].

Since 2006, deep learning has become a new research hotspot in machine learning. In SPSS, deep learning-based acoustic models can be roughly divided into two categories: one is to use **Restricted Boltzmann Machines** (**RBMs**) [10] or **Deep Belief Networks** (**DBNs**) [11] to improve the probability density function of HMM state, and the other is to use **deep neural networks** (**DNNs**) [3], **Mixture Density Networks** (**MDNs**) or **Recurrent Neural Networks** (**RNNs**) [12] to model the relationship between linguistic features and acoustic parameters. Compared with statistical models commonly used in SPSS (e.g., decision trees, HMMs), deep learning technologies better represent the mapping between the context linguistic features and acoustic parameters. Therefore, deep learning technologies help overcome the loss of detailed characteristics in synthesized speech.

Like other **sequence-to-sequence** (**seq2seq**) learning tasks, TTS essentially maps character sequences of a sentence to acoustic parameters sequences of an utterance. The successful application of the attention-based seq2seq models in machine translation [13, 14], speech recognition [15, 16] make the seq2seq-based models [17, 18] have also been applied to speech synthesis. Subsequently, the seq2seq TTS represented by Tacotron can be directly trained on the pairs of ⟨ text, speech ⟩, and automatically learn the alignment and mapping of the characters to the spectrogram frames, which are then converted to waveforms by Griffin–Lim algorithm [19]. Distinct from Tacotron, Tacotron 2 [20] refines model structure and cascade with a modified WaveNet vocoder [21] to improve the voice quality of synthesized speech. Although Tacotron 2 has good performance, it still faces two challenges. One challenge is that the seq2seq model requires a large amount of training data, which is tough to use for the low-resource languages TTS. On the other hand, prosody plays a significant role in improving the naturalness and fluency of synthesized speech. Tacotron 2 uses a concise module structure to convert phonemes or characters to the spectrogram but can't learn the text's implicit prosodic information well. The limited training data will lead to insufficient prosodic information extracted from ⟨ text, speech ⟩. If we can directly input additional prosodic information into the seq2seq TTS of low-resource languages, the voice quality of synthesized speech will be improved.

There are about 6,000 languages in the world, most of which are low-resource languages. Although the current TTS systems of major languages (e.g., Mandarin, English, French) have achieved good results, the voice quality of TTS models in low-resource languages still needs

further improvement. Tibetan is one kind of low-resource minority language in China. Compared with major languages, Tibetan speech synthesis research works started late and have a weak foundation. In our previous work, we have proposed cross-lingual methods [22, 23] to improve the Tibetan TTS, but the voice quality of synthesized Tibetan speech is still unsatisfactory. The article proposes two seq2seq Tibetan TTS methods by fusing prosodic information to improve the voice quality of synthesized Tibetan speech. We first constructed a large-scale Tibetan corpus for seq2seq TTS. Then we designed a prosody generator to extract prosodic information from the Tibetan sentences. Finally, we trained two seq2seq Tibetan TTS models by fusing prosodic information, including feature-level and model-level prosodic information fusion. The experimental results showed that the proposed two seq2seq Tibetan TTS models, which fuse prosodic information, could effectively improve the voice quality of synthesized speech. The model-level prosodic information fusion only needs 60% ∼ 70% of the training data to synthesize a similar voice to the baseline seq2seq Tibetan TTS. Therefore, the proposed methods can be used for speech synthesis of low-resource languages by modeling prosodic information fusion. Our main contributions are as follows:

(1) We construct a large Tibetan corpus for seq2seq speech synthesis and build the baseline Tibetan seq2seq TTS based on Tacotron 2.
(2) To further improve the prosody robustness of seq2seq speech synthesis, we design a prosody generator to extract prosodic information from the text. Then, the extracted prosodic information is integrated with the baseline seq2seq Tibetan TTS to realize modeling prosodic information fusion in seq2seq TTS. The prosodic information fusion methods include feature-level and model-level prosodic information fusion.
(3) We investigate Tibetan speech synthesis under low-resource conditions. The experiments showed that the model-level prosodic information fusion only needs 60% ∼ 70% of the training data to synthesize a similar voice to the baseline. Thus, it is valuable to construct a TTS for low-resource languages with prosodic information fusion.

The rest of the article is organized as follows. We first introduce the related work in Section 2. Then we illustrate the Tibetan corpus in Section 3. In Section 4, we build baseline Tibetan TTS based on Tacotron 2. Section 5 proposes two modeling prosodic information fusion methods in seq2seq speech synthesis, including feature-level and model-level prosodic information fusion. The experimental setup and experimental results are presented in Section 6. Finally, a brief conclusion and future work are provided in Section 7.

2 RELATED WORK

Seq2seq speech synthesis is a state-of-the-art speech synthesis method. To make the synthesized speech more expressive, the researchers attempt to extract the information representing the speaker and prosody as an additional input to the seq2seq structure, making the acoustic model can automatically learn the speaker representation and implicit prosodic information in English seq2seq TTS [24–27]. These methods extract a latent representation for the entire speech, and a global representation is needed to capture the whole variation space of speech signals of any length. In contrast, the model proposed in [28] uses a fine-grained structure to encode the prosody associated with each phoneme. Furthermore, many researchers have developed prosody modeling of speech synthesis based on **Variational Auto-encoder** (**VAE**) and its improved methods [29] to model fine-grained prosody attributes (e.g., phoneme-level, word-level).

Chinese character is different from English and other Roman character-based languages. Because there is no distinct separator between adjacent words, an occasional word segmentation error leads to semantic confusion and prosodic errors in speech synthesis. In [30], the

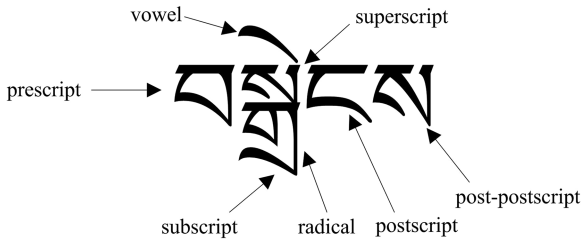


Fig. 1. The typical longest Tibetan character.

experiments showed that enhanced input text by integrating prosodic information (e.g., prosodic word boundaries, prosodic phrase boundaries, HTS-based context-dependent information) significantly improves the naturalness of the synthesized speech in seq2seq TTS. In [31–33], the text embedding, extracted by the pre-training model **Bidirectional Encoder Representations from Transformers** (**BERT**), is an additional input to add to the seq2seq TTS based on Tacotron 2. Because these text embedding features contain linguistic and semantic-related information, this information can help the speech synthesis system to generate more natural speech.

Tibetan belongs to the Sino-Tibetan language family, so Tibetan has many similarities with Chinese in linguistics and phonetics. Tibetan has three dialects, including the Lhasa dialect, the Kang dialect, and the Amdo dialect. Because the Lhasa dialect is the standard pronunciation of Tibetan, the article focuses on realizing Lhasa dialect speech synthesis. Although the above three Tibetan dialects have different pronunciations, they use the same Tibetan characters. A Tibetan character with horizontal and vertical spelling structure differs from English with fully linear spelling. The typical Tibetan character consists of seven parts, as shown in Figure 1. The spelling order of Tibetan characters is prescript, superscript, radical, subscript, vowel, postscript, and post-postscript. A Tibetan character has at least one radical. In seq2seq TTS, we should first convert the input Tibetan character sequences into phoneme sequences. In our previous work, we have realized the Tibetan grapheme-to-phoneme conversion module to convert Tibetan characters into initial and final sequences [22].

The voice quality of synthesized Tibetan speech still needs improvement compared to major languages. In [34, 35], the unit selection-based concatenative Tibetan speech synthesis and HMM-based Tibetan SPSS have been proposed, respectively. In [36], although the authors have realized seq2seq Tibetan speech synthesis, the voice quality of synthesized speech still needs to be improved due to the limitation of the corpus. In our previous works, we also have realized an HMM-based and deep learning-based Mandarin-Tibetan cross-lingual speech synthesis [22, 23]. In [23], we developed a Tibetan text analyzer for generating context-dependent labels from Tibetan sentences. The text analyzer consists of text normalization, word segmentation, prosody prediction, and grapheme-to-phoneme conversion. At the same time, we investigated the low-resource language speech synthesis method by borrowing major languages corpus in [23].

3 TIBETAN CORPUS FOR SPEECH SYNTHESIS

The Tibetan corpus for speech synthesis, called Tibetan_Lasa_1, is a standard Lhasa dialect corpus of female speakers. The utterances are recorded in a professional recording studio. We use high-fidelity microphones and professional audio workstations to record speech corpus and conduct legality checks on all recordings to avoid errors in the recording process.

There are 10,000 recordings in total from a 25-year-old female Lhasa dialect broadcaster. All sentences are mainly declarative utterances from news, Tibetan websites, and Tibetan books. All recordings are saved in the Microsoft WAV format (mono-channel, 16-bit depth, sampled at 16 kHz).

Table 1. The Distribution of Utterance Length in the Tibetan_Lasa_1

	mean	max	min
Utterance length(character)	16	30	2
Utterance length(second)	2.55	5.26	0.68

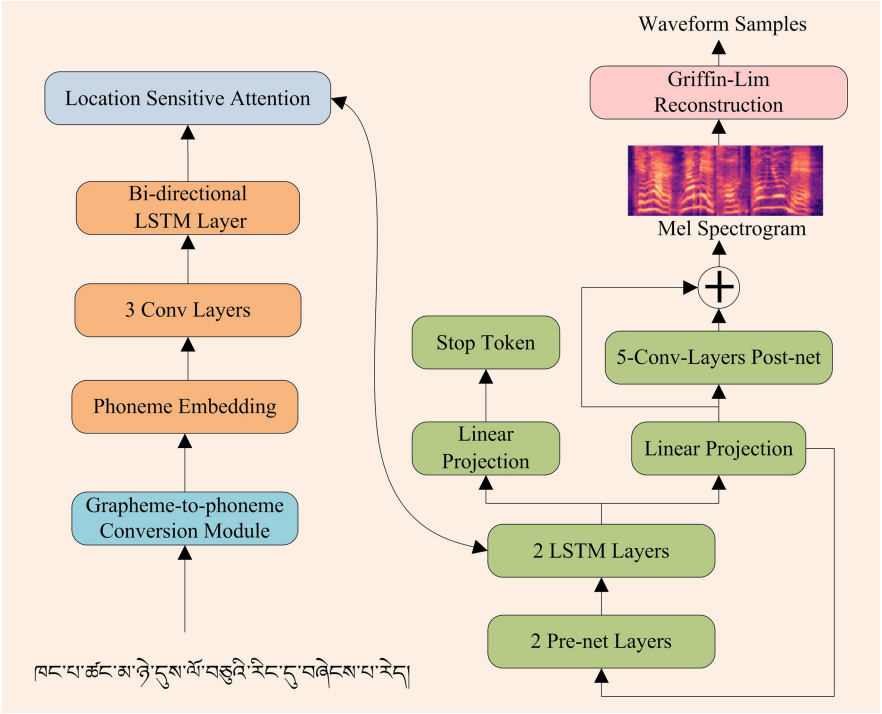


Fig. 2. Baseline of Seq2seq Tibetan TTS model.

For example, one of the recordings is named TIL_F001_00001.wav, where TIL stands for Tibetan Lhasa dialect, F001 stands for the first female speaker, and 00001 stands for the file number of the speaker. All recordings removed silence segment are in total of 7.08 hours. Table 1 shows the distribution of some essential attributes across the entire corpus.

4 BASELINE OF SEQ2SEQ TIBETAN SPEECH SYNTHESIS

The Tacoton2-based baseline of seq2seq Tibetan TTS is shown in Figure 2. The network consists of a grapheme-to-phoneme conversion module, an encoder, and a decoder with location-sensitive attention. Since the article investigates prosodic information fusion on the voice quality improvement of seq2seq Tibetan speech synthesis, we use the Griffin-lim algorithm instead of a neural vocoder to reconstruct speech waveforms to speed up the model's training.

The grapheme-to-phoneme conversion module (color marked with light blue) converts Tibetan characters into initial and final sequences. The architecture of the grapheme-to-phoneme conversion module is the same as our previous work in [22, 23].

The encoder (color marked with orange) consists of a phoneme embedding layer, 3-convolution-layers (3 Conv Layers), and a Bi-directional long short-term memory layer (Bi-directional LSTM

Layer). The encoder converts the Tibetan initial and final sequences into hidden states of the encoder.

The decoder (color-coded with green) is an autoregressive recurrent neural network. It consists of 2 layers pre-net (2 Layers Pre-net), 2 LSTM layers, a linear projection layer (Linear Projection), and 5-convolution-layers post-net (5-Conv-Layers Post-net). The decoder with location-sensitive attention predicts mel spectrogram from the encoded output hidden feature representation sequences. In parallel to mel spectrogram frames prediction, another linear projection predicts whether the output sequence would complete.

5 THE PROPOSED SEQ2SEQ TIBETAN SPEECH SYNTHESIS WITH PROSODIC INFORMATION FUSION

5.1 Prosody Generator

In the Tacotron2-based seq2seq TTS, the acoustic model needs a large number of ⟨ text, speech ⟩ pairs to explore the syntactic and semantic information from the input sentences, which is difficult for low-resource languages. To adequately explore the syntactic and semantic information, we design a prosody generator to extract the prosodic information from the sentence. The prosody generator includes a text analyzer, a feature vector extraction module, a question set, and a hidden feature extraction module.

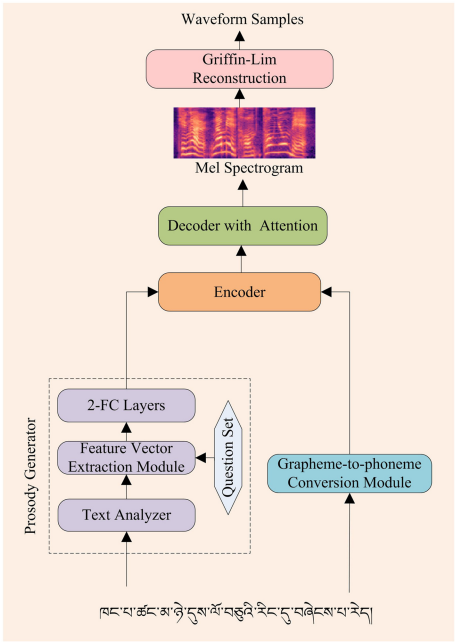
All initials and finals of Tibetan are used as the speech synthesis unit. The text analyzer generates HTS-based context labels, which are designed by taking into account the following context-dependent features [22].

- unit level: the {pre-preceding, preceding, current, succeeding, suc-succeeding} unit identity, the position of the current unit in the current syllable.
- syllable level: the {initial, final, tone type, number of units} of the {preceding, current, succeeding} syllable, the position of the current syllable in the current {word, prosodic word, phrase}.
- word level: the {POS, number of syllables} of the {preceding, current, succeeding} word, the position of the current word in the current {prosodic word, phrase}.
- prosodic word level: the number of {syllables, words} in the {preceding, current, succeeding} prosodic word, the position of the current prosodic word in the current phrase.
- phrase level: the intonation type of the current phrase, the number of the {syllables, words, prosodic words} in the {preceding, current, succeeding} phrase.
- utterance level: whether the utterance has question intonation or not, the number of {syllables, words, prosodic words, phrases} in this utterance.

The feature vector extraction module converts the context-dependent labels of the current synthesis unit into a 764-dimensional numerical vector according to the question set. In the seq2seq Tibetan speech synthesis model with feature-level prosodic information fusion, the hidden feature extraction module consists of 2 fully-connected-layers (2-FC layers). However, unlike feature-level prosodic information fusion, the hidden feature extraction module consists of a Bi-directional LSTM layer instead of 2-FC layers to better extract prosodic information in model-level prosodic information fusion.

5.2 The seq2seq Tibetan Speech Synthesis Model with Feature-level Prosodic Information Fusion

The seq2seq Tibetan TTS model with feature-level prosodic information fusion is shown in Figure 3(a). The prosodic feature vectors generated by the prosody generator and corresponding



(a) The seq2seq Tibetan TTS model with feature-level prosodic information fusion

(b) The seq2seq Tibetan TTS model with model-level prosodic information fusion

Fig. 3. The seq2seq Tibetan TTS model by fusing prosodic information.

phoneme embedding are concatenated to enhance the prosodic information of the original Tacotron 2.

Specifically, we define the prosodic feature vectors generated by the prosody generator as $\{\boldsymbol{p}_j\}_{j=1}^{N}$, and the corresponding phoneme embedding sequences as $\{\boldsymbol{e}_j\}_{j=1}^{N}$. The phoneme embedding vector sequences are concatenated with the prosodic feature vectors before feeding into the encoder. The goal of the seq2seq TTS model is to convert $\{\boldsymbol{e}_j, \boldsymbol{p}_j\}_{j=1}^{N}$ to the output acoustic feature vector sequences $\{\boldsymbol{y}_j\}_{j=1}^{T}$ with length $T(N \neq T)$. The output of the encoder is shown in Equation (1), and $\boldsymbol{h}_j$ is the hidden state of the encoder.

$$\{\boldsymbol{h}_j\}_{j=1}^{N} = \text{Encoder}\left(\{\boldsymbol{e}_j, \boldsymbol{p}_j\}_{j=1}^{N}\right). \tag{1}$$

The output hidden state sequences of decoder can be expressed as $\{\boldsymbol{d}_i\}_{i=1}^{T}$, as shown in Equation (2).

$$\boldsymbol{d}_i = \text{Decoder}\left(\boldsymbol{d}_{i-1}, \boldsymbol{c}_i, \boldsymbol{s}_i\right). \tag{2}$$

Where, the output state sequences of attention RNN can be expressed as $\{\boldsymbol{s}_j\}_{j=1}^{T}$, as shown in Equation (3). The context vector $\boldsymbol{c}_i$ is calculated by Equation (4).

$$\boldsymbol{s}_i = \text{RNN}_{\text{att}}\left(\boldsymbol{s}_{i-1}, \boldsymbol{c}_{i-1}, \boldsymbol{y}_{i-1}\right), \tag{3}$$

$$\boldsymbol{c}_i = \sum \alpha_{i,j} \boldsymbol{h}_j. \tag{4}$$

The alignment $\alpha_{i,j}$ can be calculated by Equations (5) and (6).

$$\alpha_{i,j} = \exp(e_{i,j}) / \sum_{j=1}^{L} \exp(e_{i,j}), \tag{5}$$

$$e_{i,j} = w^{\top} \tanh(Ws_{i-1} + Vh_j + Uf_{i,j} + b), \tag{6}$$

where w and b are vectors, W, V, U are matrices. $e_{i,j}$ indicates the scoring mechanism that extend the content-based attention mechanism to be location sensitive by making it take into account the alignment produced at the previous step. $f_{i,j} \in \mathbb{R}^k$ can be calculated by convolving every position j of the previous alignment α_i with a matrix $F \in \mathbb{R}^{k \times r}$:

$$f_i = F * \alpha_{i-1}. \tag{7}$$

5.3 The seq2seq Tibetan Speech Synthesis Model with Model-level Prosodic Information Fusion

The seq2seq Tibetan TTS model with model-level prosodic information fusion is shown in Figure 3(b). Unlike the feature-level prosodic information fusion, we use a Bi-directional LSTM layer instead of 2-FC layers to extract prosodic information better. The hidden states of the encoder are concatenated with the prosodic feature vectors $\{\boldsymbol{p}_j\}_{j=1}^{N}$ before feeding into the next module. We define this prosodic information fusion as model-level prosodic information fusion. The context vector $\boldsymbol{c}_i$ is calculated as Equation (8). Other parameters are consistent with the seq2seq Tibetan TTS model with feature-level prosodic information fusion.

$$\boldsymbol{c}_i = \sum \alpha_{i,j}\{\boldsymbol{h}_j, \boldsymbol{p}_j\}. \tag{8}$$

6 EXPERIMENTS

6.1 Experimental Data

In the experiment, we selected training data from the Tibetan_Lasa_1. First, two sets of 150 utterances were randomly selected from Tibetan_Lasa_1 to serve as a test set and development set, respectively. Then the remaining 9,700 utterances were used to form six training sets {A, B, C, D, E, F} containing {9700, 9000, 8000, 7000, 6000, 5000} utterances. We trained the Tibetan seq2seq TTS models on the above six subsets.

6.2 Experimental Setup

We build the following three TTS models to verify the effectiveness of the proposed two prosodic information fusion methods in low-resource language TTS.

- seq2seq-Baseline: seq2seq-baseline, trained on subset A, is the baseline seq2seq TTS model, and its architecture is shown in Figure 2.
- seq2seq-CP: We call the seq2seq TTS model with feature-level prosodic information fusion as seq2seq-CP, which is trained on training sets {A, B, C, D, E, F}, respectively.
- seq2seq-MCP: We call the seq2seq TTS model with model-level prosodic information fusion as seq2seq-MCP, which is trained on training sets {A, B, C, D, E, F}, respectively.

Adam optimizer was used in the experiments to train the above TTS models. The learning rate stepped from 10^{-3} and halved every 10k step. All models were trained until the loss of the development set started increasing. The experiments used two NVIDIA Tesla P4 GPUs, and batch_size was set to 64. The 2-FC layers of seq2seq-CP's prosody generator consist of 2 fully connected layers with 128 hidden ReLU units, while the Bi-directional LSTM of seq2seq-MCP's prosody generator consists of 2 layers LSTM with 128 units. The other parameters of all models are similar to the original Tacotron 2.

6.3 Experimental Evaluations

To evaluate the effectiveness of feature-level prosodic information fusion and model-level prosodic information fusion in low resource language seq2seq TTS, we evaluate the voice quality of

Table 2. The MCD of Different TTS Models on the Test Set

Models	Training sets					
	A	B	C	D	E	F
seq2seq-Baseline	4.376	–	–	–	–	–
seq2seq-CP	4.227	4.318	4.537	4.762	4.906	5.763
seq2seq-MCP	4.185	4.200	4.242	4.297	4.431	4.475

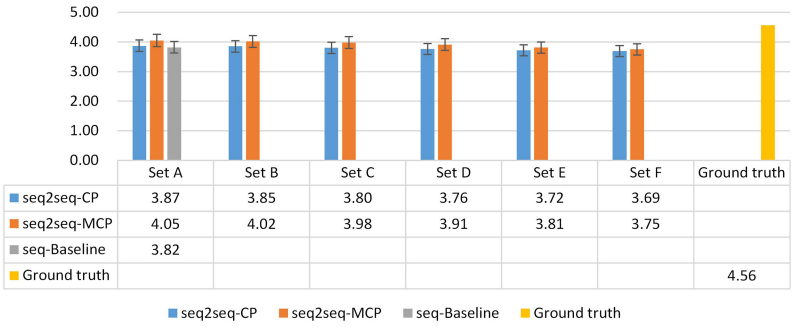


Fig. 4. The average MOS scores of synthesized speech under 95% confidence intervals.

synthesized speech of the above models under different training sets. The evaluations include objective evaluation and subjective evaluation.

6.3.1 Objective Evaluation. In the objective evaluations, we evaluated the **Mel-cepstral distortion** (**MCD**) between the original speech samples and predicted speech samples of each model. We marked the above three seq2seq TTS acoustic models on different training sets as a "model-training set". For example, the seq2seq-Baseline-A represents the seq2seq-Baseline TTS model trained on subset A and is similar to seq2seq-CP and seq2seq-MCP.

The MCD of different seq2seq TTS models on the test set is shown in Table 2. From Table 2., the findings can be analyzed that the seq2seq TTS with prosodic information fusion help to improve the voice quality of the synthesized speech. Compared with seq2seq-CP, seq2seq-MCP has higher modeling accuracy. Especially as the utterance number of the training set decreases, this advantage becomes more evident. For example, when the training data of seq2seq-MCP is 60%−70% in size of the training set A, the MCD of synthesized speech is similar to that of seq2seq-Baseline trained on A.

6.3.2 Subjective Evaluation. For subjective evaluations, we conducted a **mean opinion score** (**MOS**) test and an AB preference test to evaluate the quality of synthesized speech. We invited 30 native Tibetan listeners as subjects. In the MOS test, 20 synthesized utterances were randomly selected from the test sets of the above three seq2seq TTS acoustic models. The subjects were asked to rate the naturalness of the synthesized speech using a 5-point scale score. The average MOS scores of synthesized speech are shown in Figure 4.

We generated 20 paired synthesized utterances from the above three seq2seq TTS acoustic models in the AB preference test. Each pair of synthesized utterances was played at random. The subjects were asked to listen and judge the quality of which utterance was better (or "neutral" means that the subjects had no preference). The preference results of the synthesized speech are shown in Table 3.

From the subjective evaluations, we can find that the seq2seq-MCP-A acoustic model obtains the best subjective evaluation scores, and the subjective evaluation scores of the seq2seq-MCP-E

Table 3. The Preference Scores(%) of the Synthesized Speech with $p < 0.01$

	seq2seq-Baseline-A	seq2seq-CP-A	seq2seq-MCP-A	seq2seq-MCP-E	Neutral
1	20.2	69.7	–	–	10.2
2	19.2	–	70.7	–	10.2
3	–	20.5	68.0	–	11.2
4[a]	38.8	–	–	37.7	23.5

[a] $p = 0.06$.

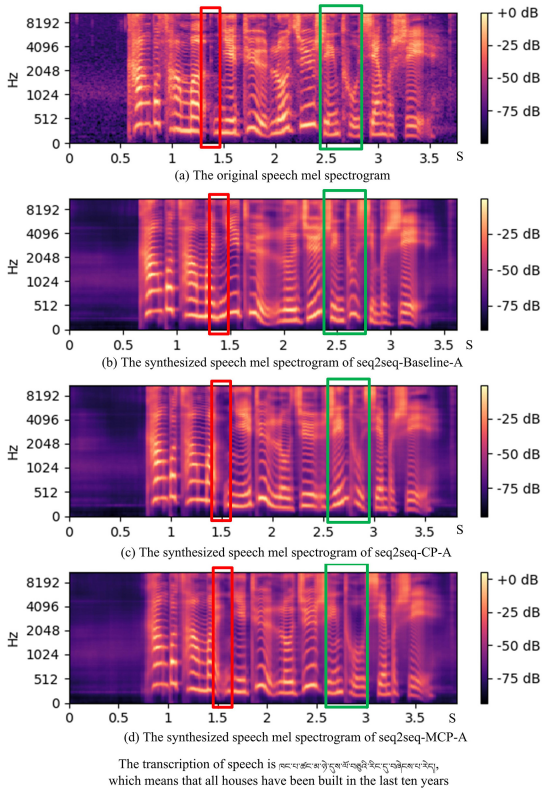


Fig. 5. The mel spectrogram of synthesized speech sample.

acoustic model is equivalent to that of seq2seq-Baseline-A. Thus, the subjective evaluation results are consistent with the objective evaluation results.

6.4 Experimental Results Discussion

In objective and subjective evaluations, the seq2seq TTS with prosodic information fusion can improve the voice quality of synthesized speech. Furthermore, the seq2seq-MCP model effectively reduces the training data required under a similar voice as the baseline, proving that model-level prosodic information fusion can help improve the modeling accuracy of seq2seq TTS.

In experiments, we also found that the improved cases of the seq2seq TTS with prosodic information fusion are usually from the synthesized samples with improper breaks and spectrogram details produced by the baseline model. For example, the mel spectrograms of a synthesized Tibetan utterance are shown in Figure 5. In Figure 5(b), we found a missing break (red rectangles)

in the seq2seq-Baseline-A model, while the proposed seq2seq-CP-A and seq2seq-MCP-A models produced a more appropriate break pattern. In addition, as shown by the green rectangle in Figure 5(c) and (d), the seq2seq-MCP-A model also improves details of the mel spectrogram of synthesized speech compared with seq2seq-CP-A. The above experimental results demonstrated that model-level prosodic information fusion has higher modeling accuracy than feature-level prosodic information fusion.

7 CONCLUSIONS AND FUTURE WORK

The article proposes seq2seq Tibetan speech synthesis with prosodic information fusion. We first constructed a large-scale Tibetan corpus for seq2seq TTS. Then we designed a prosody generator to extract prosodic information from the Tibetan sentences. Finally, we trained two seq2seq Tibetan TTS models by fusing prosodic information, including feature-level and model-level prosodic information fusion. The objective and subjective experiments demonstrated that the proposed two prosodic information fusion methods could effectively improve the voice quality of synthesized speech. Compared with feature-level prosodic information fusion, model-level prosodic information fusion has higher modeling accuracy. This advantage becomes more obvious as the utterance number of the training set decreases.

Furthermore, we also investigated how to use the least corpus for synthesizing Tibetan speech with satisfactory voice quality. The experiments showed that the model-level prosodic information fusion only needs 60% ∼ 70% of the training data to synthesize a voice similar to the baseline. Therefore, our methods would be valuable for constructing a TTS for low-resource languages with prosodic information fusion.

Future work will improve the prosodic expression of synthesized Tibetan speech with other linguistic features (e.g., BERT-derived features). We will also focus on realizing low-resource TTS through cross-language speech synthesis and data enhancement.

REFERENCES

[1] Andrew J. Hunt and Alan W. Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing*. 373–376. DOI: http://dx.doi.org/10.1109/ICASSP.1996.541110

[2] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura. 2013. Speech synthesis based on hidden markov models. *Proc. IEEE* 101, 5 (2013), 1234–1252. DOI: http://dx.doi.org/10.1109/JPROC.2013.2251852

[3] Heiga Zen, Andrew Senior, and Mike Schuster. 2013. Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7962–7966. DOI: http://dx.doi.org/10.1109/ICASSP.2013.6639215

[4] Yuchen Fan, Yao Qian, Fenglong Xie, and Frank K. Soong. 2014. TTS synthesis with bidirectional LSTM based recurrent neural networks. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association*. 1964–1968.

[5] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards end-to-end speech synthesis. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*. 4006–4010. DOI: http://dx.doi.org/10.21437/Interspeech.2017-1452

[6] Makoto Tachibana, Shinsuke Izawa, Takashi Nose, and Takao Kobayashi. 2008. Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. 4633–4636. DOI: http://dx.doi.org/10.1109/ICASSP.2008.4518689

[7] Takayoshi Yoshimura, Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Tadashi Kitamura. 1997. Speaker interpolation in HMM-based speech synthesis system. *The Journal of The Acoustical Society of Japan (e)* 21 (1997), 199–206.

[8] Amir Mohammadi, Seyyed Saeed Sarfjoo, and Cenk Demiroğlu. 2014. Eigenvoice speaker adaptation with minimal data for statistical speech synthesis systems using a MAP approach and nearest-neighbors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22, 12 (2014), 2146–2157. DOI: http://dx.doi.org/10.1109/TASLP.2014.2362009

[9] Zhenhua Ling, Shiyin Kang, Heiga Zen, Andrew Senior, Mike Schuster, Xiaojun Qian, Helen Meng, and Li Deng. 2015. Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends. *IEEE Signal Processing Magazine* 32, 3 (2015), 35–52. DOI: http://dx.doi.org/10.1109/MSP.2014.2359987

[10] Zhenhua Ling, Li Deng, and Dong Yu. 2013. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 10 (2013), 2129–2139. DOI: http://dx.doi.org/10.1109/TASL.2013.2269291

[11] Shiyin Kang, Xiaojun Qian, and Helen Meng. 2013. Multi-distribution deep belief network for speech synthesis. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 8012–8016. DOI: http://dx.doi.org/10.1109/ICASSP.2013.6639225

[12] Eunwoo Song, Frank K. Soong, and Honggoo Kang. 2017. Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25, 11 (2017), 2152–2161. DOI: http://dx.doi.org/10.1109/TASLP.2017.2746264

[13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*. http://arxiv.org/abs/1409.0473

[14] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1412–1421.

[15] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: First results. In *Proceedings of the NIPS 2014 Workshop on Deep Learning*.

[16] Jan Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*. 577–585.

[17] Wenfu Wang, Shuang Xu, and Bo Xu. 2016. First step towards end-to-end parametric TTS synthesis: Generating spectral parameters with neural attention. In *Proceedings of the 17th Annual Conference of the International Speech Communication Association*. 2243–2247. DOI: http://dx.doi.org/10.21437/Interspeech.2016-134

[18] Jose Sotelo, Soroush Mehri, Kundan Kumar, Joao Felipe Santos, Kyle Kastner, Aaron Courville, and Yoshua Bengio. 2017. Char2Wav: End-to-end speech synthesis. In *Proceedings of the 5th International Conference on Learning Representations*.

[19] Daniel W. Griffin and Jae S. Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32, 2 (1984), 236–243. DOI: http://dx.doi.org/10.1109/TASSP.1984.1164317

[20] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. 4779–4783. DOI: http://dx.doi.org/10.1109/ICASSP.2018.8461368

[21] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. In arXiv:1609.03499. Retrieved from https://arxiv.org/abs/1609.03499

[22] Hongwu Yang, Keiichiro Oura, Haiyan Wang, Zhenye Gan, and Keiichi Tokuda. 2015. Using speaker adaptive training to realize MandarinTibetan crosslingual speech synthesis. *Multimedia Tools & Applications* 74, 22 (2015), 1–16.

[23] Weizhang Zhang, Hongwu Yang, Xiaolong Bu, and Lili Wang. 2019. Deep learning for mandarin-tibetan cross-lingual speech synthesis. *IEEE Access* 7 (2019), 167884–167894. DOI: http://dx.doi.org/10.1109/ACCESS.2019.2954342

[24] R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *Proceedings of the 35th International Conference on Machine Learning*. 4700–4709. Retrieved from http://proceedings.mlr.press/v80/skerry-ryan18a.html

[25] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*. 5167–5176. Retrieved from http://proceedings.mlr.press/v80/wang18h.html

[26] Ya Jie Zhang, Shi feng Pan, Lei He, and Zhen Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6945–6949. DOI: http://dx.doi.org/10.1109/ICASSP.2019.8683623

[27] Rafael Valle, Jason Li, Ryan Prenger, and Bryan Catanzaro. 2020. Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6189–6193. DOI: http://dx.doi.org/10.1109/ICASSP40776.2020.9054556

[28] Younggun Lee and Taesu Kim. 2019. Robust and fine-grained prosody control of end-to-end speech synthesis. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. 5911–5915. DOI: http://dx.doi.org/10.1109/ICASSP.2019.8683501

[29] Guangzhi Sun, Yu Zhang, Ron J. Weiss, Yuan Cao, Heiga Zen, and Yonghui Wu. 2020. Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6264–6268. DOI: http://dx.doi.org/10.1109/ICASSP40776.2020.9053520

[30] Yanfeng Lu, Minghui Dong, and Ying Chen. 2019. Implementing prosodic phrasing in chinese end-to-end speech synthesis. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. 7050–7054. DOI: http://dx.doi.org/10.1109/ICASSP.2019.8682368

[31] Jingbei Li, Zhiyong Wu, Runnan Li, Pengpeng Zhi, Song Yang, and Helen Meng. 2019. Knowledge-based linguistic encoding for end-to-end mandarin text-to-speech synthesis. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*. 4494–4498. DOI: http://dx.doi.org/10.21437/Interspeech.2019-1118

[32] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Shubham Toshniwal, and Karen Livescu. 2019. Pre-trained text embeddings for enhanced text-to-speech synthesis. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association*. 4430–4434. DOI: http://dx.doi.org/10.21437/Interspeech.2019-3177

[33] Yujia Xiao, Lei He, Huaiping Ming, and Frank K. Soong. 2020. Improving prosody with linguistic and bert derived features in multi-speaker based mandarin chinese neural TTS. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. 6704–6708. DOI: http://dx.doi.org/10.1109/ICASSP40776.2020.9054337

[34] CAI Rangzhuoma and CAI Zhijie. 2017. Unit selection algorithm for corpus-based Tibetan speech synthesis. *Journal of Chinese Information Processing* 31, 5 (2017), 59–63.

[35] Zhiqiang Wu, Hongzhi Yu, Guanyu Li, and Shuhui Wan. 2013. HMM-based Tibetan Lhasa speech synthesis system. In *Proceedings of 2013 3rd International Conference on Computer Science and Network Technology*. 92–95. DOI: http://dx.doi.org/10.1109/ICCSNT.2013.6967070

[36] DOU Gecao, CAI Rangzhuoma, NAN Cuoji, and SUAN Taiben. 2019. Neural network based Tibetan speech synthesis. *Journal of Chinese Information Processing* 33, 2 (2019), 75–80.

Received 13 December 2020; revised 14 July 2022; accepted 2 August 2023