IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 21, 2019, accepted November 8, 2019, date of publication November 19, 2019, date of current version December 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954342

Deep Learning for Mandarin-Tibetan Cross-Lingual Speech Synthesis

WEIZHAO ZHANG[1,2], HONGWU YANG[1,2,3], (Member, IEEE), XIAOLONG BU[1], AND LILI WANG[1]

[1]College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China
[2]National and Provincial Joint Engineering Laboratory of Learning Analysis Technology in Online Education, Lanzhou 730070, China
[3]School of Educational Technology, Northwest Normal University, Lanzhou 730070, China

Corresponding author: Hongwu Yang (yanghw@nwnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 11664036, and in part by the High School Science and Technology Innovation Team Project of Gansu under Grant 2017C-03.

ABSTRACT This paper proposes a deep learning-based Mandarin-Tibetan cross-lingual speech synthesis to realize both Mandarin speech synthesis and Tibetan speech synthesis under a unique framework. Because Tibetan training corpus is hard to record, we train the acoustic models with a large scale Mandarin multi-speaker corpus and a small scale Tibetan one-speaker corpus. The acoustic models are trained with deep neural network (DNN), hybrid long short-term memory (LSTM), and hybrid bi-directional long short-term memory (BLSTM). We also further extend our Chinese text analyzer by adding a Tibetan text analyzer for generating context-dependent labels from input Chinese or Tibetan sentences. The Tibetan text analyzer includes a text normalization, a novel Tibetan word segmentation that combines a BLSTM with conditional random field, a prosodic boundary prediction, and a grapheme-to-phoneme conversion. We select the initials and the finals of both Mandarin and Tibetan as the speech synthesis units to train a speaker-independent mixed language average voice model (AVM) with DNN, hybrid LSTM, and hybrid BLSTM from Mandarin and Tibetan mixed corpus. Then the speaker adaptation is applied to train speaker-dependent DNN, hybrid LSTM, or hybrid BLSTM models of Mandarin or Tibetan with a small target speaker corpus from an AVM. Finally, we synthesize the Mandarin speech, or Tibetan speech though the speaker-dependent Mandarin or Tibetan models. The experiments show that the hybrid BLSTM-based cross-lingual speech synthesis framework is better than the other two cross-lingual frameworks and the Tibetan monolingual framework. The mixed Tibetan training corpus does not influence the voice quality of synthesized Mandarin speech. Furthermore, the hybrid BLSTM-based cross-lingual speech synthesis framework only needs 60% of the training corpus to synthesize a similar voice as the Tibetan monolingual framework. Therefore, the proposed method can be used for speech synthesis of low resource languages by borrowing the same tremendous resource language's corpus.

INDEX TERMS Mandarin-Tibetan cross-lingual speech synthesis, Tibetan speech synthesis, minority language speech synthesis, deep learning, low resource languages.

I. INTRODUCTION

In recent years, cross-lingual speech synthesis has been a popular topic in text-to-speech synthesis (TTS) research [1], [2]. Since cross-lingual speech synthesis can synthesize speech in different languages with the same or a different speaker's voice, it has been widely used in human-computer interaction, bilingual teaching, oral dialog, and other aspects, as well as being of great significance for promoting language communication in multilingual areas.

Generating natural speech from text remains a challenging task in past decades. Speech synthesis technology has roughly gone through three phases over time, including unit selection-based concatenative speech synthesis, hidden Markov model (HMM)-based statistical parametric speech synthesis (SPSS), and end-to-end speech synthesis. The unit selection-based

The associate editor coordinating the review of this manuscript and approving it for publication was Jenny Mahoney.

concatenative speech synthesis [3], which concatenates small units of the pre-recorded waveforms together by unit selection algorithm, is state-of-the-art for many years. The HMM-based SPSS has been the most popular acoustic model since the 1990s [4], [5]. Generally speaking, SPSS pipeline usually consists of three parts [6] including a complex text frontend to extract various linguistic features from raw text, a duration model, an acoustic model to learn the transformation between linguistic features and acoustic features, and a complex signal-processing-based vocoder to reconstruct waveform from the predicted acoustic features. Many techniques such as adaptation, interpolation, eigen-voice, and multiple regression have been proposed for changing voice characteristics, speaking styles, and emotions in SPSS [5]. These methods not only have a better quality of synthesized speech than the unit selection-based concatenative speech synthesis, but it also can synthesize different speakers' speech. However, the synthesized speech of HMM-based SPSS is muffled compared with natural speech.

Deep learning technology has become a major method in SPSS since 2006. The main network architectures of deep learning include deep neural network (DNN) [7]–[9], long short-term memory (LSTM) and bi-directional long short-term memory (BLSTM) [10]. Compared with the traditional HMM-baed SPSS methods, deep learning-based methods can learn a better relationship between linguistic features and acoustic features. The voice quality of synthesized speech by the deep learning-based methods is better than that of the HMM-based speech synthesis methods. Recent research on attention-based sequence-to-sequence (seq2seq) models are demonstrated to outperform the SPSS by several end-to-end TTS systems [11], [12]. The attention-based end-to-end TTS can automatically learn alignments and mapping from linguistic features to acoustic features. These systems can be trained on <text, audio> pairs without complex language-dependent text frontend. However, such systems need a significant amount of corpus for training, while recording large-scale corpus is difficult for low resource languages.

Because it is tough to collect minority language corpus in China, the voice quality is reduced for the Tibetan TTS only using Tibetan training corpus. Therefore, using large scale majority language corpus to realize a cross-lingual speech synthesis can improve the voice quality of synthesized low resource minority language's speech. Besides, since there is still no complete Tibetan text analyzer [13], this also limits the application of the Tibetan TTS. The cross-lingual speech synthesis methods mainly include the unit selection-based methods [14], [15] and the SPSS-based methods [16], [17]. Since the SPSS-based techniques have been the most successful approach for cross-lingual speech synthesis, we have realized an HMM-based Mandarin-Tibetan cross-lingual speech synthesis [18], [19]. However, our previous work did not finish a Tibetan text analyzer leading to the poor voice quality of synthesized speech. Although DNN-based speaker adaptation [22], [23] has been adopted in various neural networks-based TTS applications [20], [21], the adaptation method has not been applied to the Mandarin-Tibetan cross-lingual speech synthesis. To address these limitations, we adopt a deep learning-based framework to realize the Mandarin-Tibetan cross-lingual speech synthesis. We have achieved a Chinese text analyzer in the previous work. In our new work, we firstly complete a Tibetan text analyzer for generating context-dependent labels from Tibetan sentences. The text analyzer includes a text normalization, a word segmentation, a prosody prediction, and a grapheme-to-phoneme conversion. Then we use the initials and the finals of Mandarin and Tibetan as the speech synthesis units to train a speaker-independent mixed language average voice model (AVM) with DNN, hybrid LSTM, and hybrid BLSTM from a large Mandarin multi-speaker corpus and a small Tibetan one-speaker corpus. Finally, the speaker adaptation is applied to train the speaker-dependent DNN, hybrid LSTM, or hybrid BLSTM models of Mandarin or Tibetan with a small target speaker corpus from an AVM. The Mandarin speech or Tibetan speech is synthesized by the speaker-dependent Mandarin or Tibetan models. Additionally, the paper examines the smallest corpus for synthesizing speech with satisfactory voice quality results. For speech synthesis of low-resource languages, it is significant to study the most small corpus for synthesizing speech with satisfactory voice quality.

The rest of the paper is organized as follows. We firstly introduce the recurrent neural network-based modeling for temporal sequences In Section II. Then we give out our framework of deep learning-based Mandarin-Tibetan cross-lingual speech synthesis in Section III. In Section IV, we introduce the Tibetan text analyzer that includes text normalization, word segmentation, prosody boundary prediction, and grapheme-to-phoneme conversion. We explain the representation of the linguistic features, including mixed linguistic full context-dependent labels and question set in Section V. The experimental setup and experimental results are presented in Section VI, while the discussion of the results is given in Section VII. Finally, a brief conclusion and future work are provided in Section VIII.

II. RECURRENT NEURAL NETWORK-BASED MODELING FOR TEMPORAL SEQUENCES

A Recurrent Neural Network (RNN) is different from other feed-forward networks in its ability to model temporal sequences. However, the standard RNN structure is challenging to model long-term dependencies because of the vanishing gradient problem. The most effective way to address this issue is to use the LSTM variant of RNN. In this section, we present a brief description of the LSTM and BLSTM network.

A. LSTM

The critical component of the LSTM [24] is the memory cell and the gates. Fig. 1 illustrates a single LSTM memory cell. The LSTM memory cell is implemented as the following. This structure allows information to be retained across many

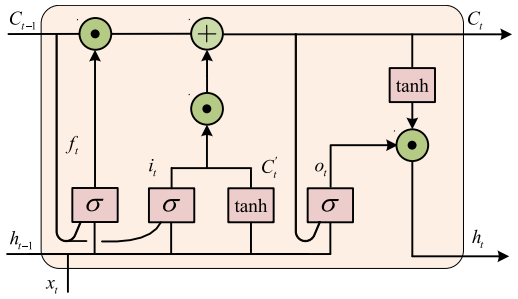


FIGURE 1. Diagram of an LSTM memory cell.

time-steps and also allows gradients to flow over many time-steps. Therefore, LSTM can effectively capture long-term temporal dependencies. It has been widely used in natural language processing and acoustic modeling for speech synthesis.

$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i\mathbf{x}_t + \mathbf{R}_i\mathbf{h}_{t-1} + \mathbf{p}_i\mathbf{c}_{t-1} + \mathbf{b}_i\right) \quad (1)$$

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f\mathbf{x}_t + \mathbf{R}_f\mathbf{h}_{t-1} + \mathbf{p}_f\mathbf{c}_{t-1} + \mathbf{b}_f\right) \quad (2)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh\left(\mathbf{W}_c\mathbf{x}_t + \mathbf{R}_c\mathbf{h}_{t-1} + \mathbf{b}_c\right) \quad (3)$$

$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o\mathbf{x}_t + \mathbf{R}_o\mathbf{h}_{t-1} + \mathbf{p}_o\mathbf{c}_t + \mathbf{b}_o\right) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right) \quad (5)$$

where σ is the sigmoid function; $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t$ and $\mathbf{c}_t$ are the outputs of input gate,forget gate, output gate and cell memory at t time, respectively. $\mathbf{h}_t$ and $\mathbf{x}_t$ stand for hidden layer outputs and input vectors at t time, respectively. $\mathbf{W}$, $\mathbf{R}$ are the weight matrices of input and recurrent units, respectively. $\mathbf{p}$, $\mathbf{b}$ are the peep-hole connections and biases, respectively. $\odot$ represents element-wise product.

B. BLSTM

The disadvantage of LSTM is that it can only access the previous inputs. BLSTM can access both the preceding and succeeding inputs utilizing the bidirectional architecture [25]. We unfold the BLSTM network forward and backward in time step, as shown in Fig. 2. Each orange box represents an LSTM memory cell. The feedforward of BLSTM includes forward state sequence $\overrightarrow{\mathbf{h}}$, and backward state sequence $\overleftarrow{\mathbf{h}}$. The hidden state sequence can be represented as $\mathbf{h} = [\overrightarrow{\mathbf{h}}, \overleftarrow{\mathbf{h}}]$.

III. THE FRAMEWORK OF DEEP LEARNING-BASED MANDARIN-TIBETAN CROSS-LINGUAL SPEECH SYNTHESIS

The framework of deep learning-based Mandarin-Tibetan cross-lingual speech synthesis is shown in Fig. 3.

In the training stage, a large multi-speaker Mandarin corpus and a small one-speaker Tibetan corpus are used to train a Mandarin-Tibetan cross-lingual speaker-independent AVM. Since speech assessment methods phonetic alphabet (SAMPA)-based grapheme-to-phoneme conversion of Mandarin and Tibetan has been designed to take into account the similarities and differences of pronunciation between Mandarin and Tibetan in generating

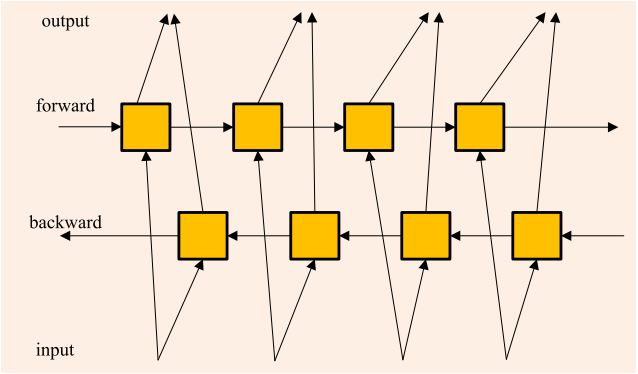


FIGURE 2. A BLSTM network.

context-dependent linguistic features [18], the AVM can be considered the transformation of all speakers' global linguistic features. In the speaker adaptation stage, a speaker-dependent target language acoustic model is re-trained using a small target language speaker corpus based on the speaker-independent AVM.

In the speech synthesis stage, the input sentence is firstly labeled by the text analyzer to generate the context-dependent labels. The context-dependent labels are then fed into the speaker-dependent acoustic model of the target language to generate the acoustic parameters. The WORLD [26] vocoder is finally used to generate the speech waveforms from the acoustic parameters.

IV. TIBETAN TEXT ANALYSIS

We further developed a Tibetan text analyzer based on the Mandarin text analyzer we have realized for the Mandarin-Tibetan cross-lingual speech synthesis. The framework of Tibetan text analysis is shown in Fig. 4. The input Tibetan sentence is first normalized. Then the word segmentation is subsequently implemented on the normalized sentence to obtain the word boundary. The prosodic boundary prediction is then carried out to get the prosodic word boundary and prosodic phrase boundary. Finally, the SAMPA-based pronunciation of Tibetan characters is obtained through a grapheme-to-phoneme conversion.

A. STRUCTURE OF TIBETAN CHARACTER

A Tibetan character is a phonetic character, but its spelling structure is different from that of English with full linear spelling. Tibetan character has a unique two-dimensional structure because of its horizontal spelling and vertical spelling. The spelling order for Tibetan character is prescript, superscript, radical, subscript, vowel, postscript, and post-postscript. The Tibetan character is monosyllabic, and the radical is the core of the syllable. A syllable has at least one radical. The longest syllable consists of seven parts, as shown in Fig. 5.

The Tibetan alphabet has 30 consonants. Each consonant letter can be regarded as a single syllable with a vowel /a/. When consonants become syllables, they all have fixed tones:

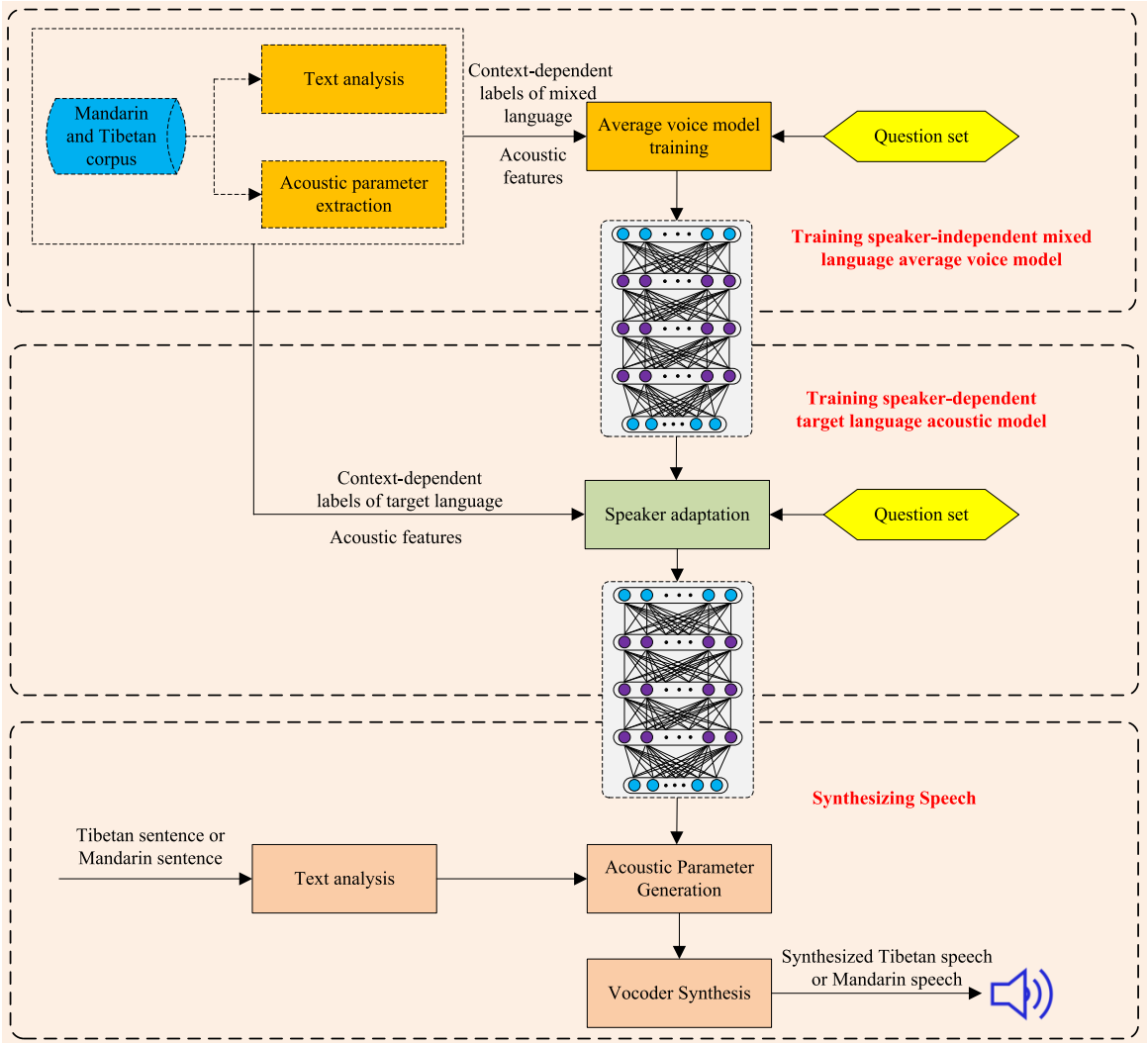


FIGURE 3. The framework of deep learning-based Mandarin-Tibetan cross-lingual speech synthesis.

14 high-pitch tones (labeled as f) and 16 with low-pitch tones (labeled as v). The consonants /rav/, /lav/ and /saf/ can be used as superscripts, /yav/, /rav/, /lav/ and /wav/ can be used as subscript, /kav/, /tav/, /pav/, /mav/, /av/ can be used as prescripts, /kav/, /ngav/, /tav/, /nav/, /pav/, /mav/, /av/, /rav/, /lav/, /saf/ can be used as postscripts, and /taf/, /saf/ can be used as post-postscripts. In modern Tibetan, the use of post-postscript /taf/ has decreased, and there is a trend of gradual loss.

B. TEXT NORMALIZATION

Text normalization is the process of converting non-Tibetan characters and abbreviated words into Tibetan characters to determine pronunciation. Non-Tibetan characters in input text mainly include Chinese words, English characters, numbers, and other symbols. These non-Tibetan characters need to be converted into corresponding Tibetan characters for further text analysis. Normalizing abbreviated words is key to text normalization due to the high frequency of these words in Tibetan sentences. We designed a set of rules for recognizing abbreviated words and adopt the add-restore method to normalize the abbreviated words according to [27], [28].

C. WORD SEGMENTATION

A Tibetan paragraph has apparent signs of separation between syllables and syllables, as well as sentences and sentences. Therefore, it is not difficult to segment a text into sentences and segment a sentence into syllables. Tibetan sentence is very like the Chinese sentence that has no distinct separator between Tibetan words. Therefore, to obtain the Tibetan word boundary for speech synthesis, we need to perform the word segmentation on Tibetan sentence. Traditional Tibetan word segmentation mainly relies on the combination of rule-based methods and statistics-based methods [29]. Here we proposed a BLSTM with conditional random field (BLSTM_CRF)-based method, as shown in Fig. 6 to obtain the Tibetan word boundary [30].

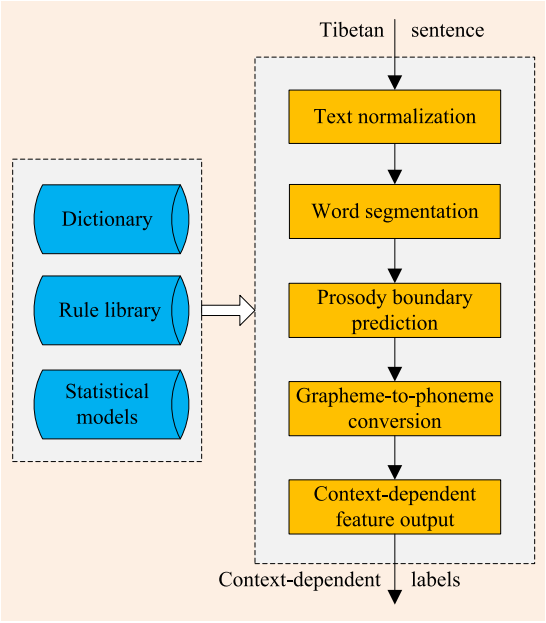


FIGURE 4. The framework of the Tibetan text analysis.

We use four word-position labeling sets (B, M, E, S) to label Tibetan words, B (begin) to label the beginning of the Tibetan word, M (middle) to label the middle position of the Tibetan word, E (End) to label the end of the Tibetan word, and S (single) to label a single word.

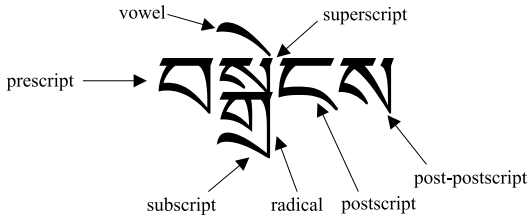


FIGURE 5. The longest syllable structure of a Tibetan character.

CRF models are typical discriminatively trained models for sequence segmentation and labeling. CRFs do not have the strict assumption of independence as HMMs so that they can contain arbitrary, overlapping, and agglomerative observation features from both the past and future. CRFs calculate the joint probability of the entire sequence of labels given the observation sequence. Maximum entropy Markov models (MEMMs) use per-state distribution exponential models for the conditional probabilities of the next states given the current state. Therefore, CRFs solve the label bias problem, and, more significantly, that CRFs perform better than HMMs and MEMMs when the true data distribution has higher-order dependencies than the model [31].

Although BLSTM learns the context-dependent information, its independent classification decisions are limiting when there are strong dependencies across output labels. In word segmentation tasks, the "word-formation"

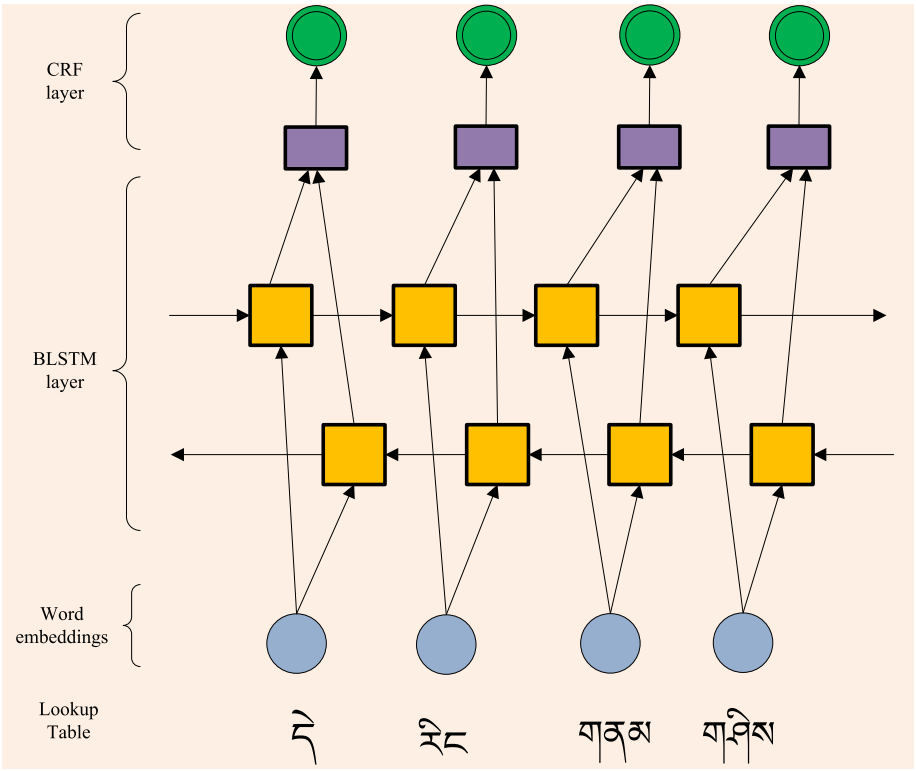


FIGURE 6. The framework of BLSTM_CRF-based Tibetan Tibetan word segmentation.

characterizes interpretable words of tags imposes several hard constraints (e.g., B cannot follow B) that are impossible to model with independent assumptions. Therefore, the output of the BLSTM layer is linearly projected onto a layer whose size is equal to the number of tags. Instead of using the softmax output from this layer, we use a CRF layer to take into account neighboring tags in Fig. 6. For a normalized input sentence $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_n}\}$ containing n words, and a tag sequence of sentence $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, each word is represented as a d-dimensional vector by word2vec. We define its prediction score $s(\mathbf{X}, \mathbf{y})$ to be

$$s(\mathbf{X}, \mathbf{y}) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=0}^{n} A_{y_i, y_{i+1}} \tag{6}$$

where $\mathbf{P}$ is the matrix of scores output by the BLSTM network. P_{i,y_i} corresponds to the score of the y_i tag of the i^{th} word in a sentence. $\mathbf{A}$ is the transition scores matrix of the CRF layer, $A_{y_i,y_{i+1}}$ corresponds to the score from the tag y_i to tag y_{i+1}.

In the training, we maximize the following log-likelihood functions

$$\log(p(\mathbf{y}|\mathbf{X})) = s(\mathbf{X}, \mathbf{y}) - \log\left(\sum_{\widehat{\mathbf{y}} \in \mathbf{Y_X}} e^{s(\mathbf{X}, \widehat{\mathbf{y}})}\right) \tag{7}$$

where $\mathbf{Y_X}$ represents all possible tag sequences for an input text $\mathbf{X}$.

In the decoding, the optimal sequence y^* is given as follows

$$\mathbf{y}^* = \underset{\widetilde{\mathbf{y}} \in \mathbf{Y_X}}{\operatorname{argmax}} \quad s(\mathbf{X}, \widehat{\mathbf{y}}) \tag{8}$$

D. PROSODIC BOUNDARY PREDICTION

It is necessary to acquire the prosodic features from input Tibetan sentences for synthesizing higher quality speech. Similar to Mandarin, the prosodic hierarchy of Tibetan can also be divided into prosodic words, prosodic phrases, and intonation phrases. The boundary of intonation phrases is easy to determine. Tibetan punctuation marks are the boundary of intonation phrases. For prosodic word boundary and prosodic phrase boundary, we use a new feature named adjacent degree to describe the relationship between grammatical structure and prosodic structure. A transform-based error-driven learning algorithm is applied to predict the boundary of prosodic words and prosodic phrases [32].

E. TIBETAN GRAPHEME-TO-PHONEME CONVERSION

Mandarin and Tibetan use different Pinyin systems to label pronunciation. However, since Mandarin and Tibetan belong to the Sino-Tibetan language family, these two languages have many similarities concerning linguistics and phonetics. We designed a SAMPA-based Mandarin and Tibetan grapheme-to-phoneme conversion based on analyzing the similarities and differences of pronunciation between Mandarin and Tibetan [18].

V. LINGUISTIC FEATURES

A. MIXED LINGUISTIC CONTEXT-DEPENDENT LABELS

All initials and finals of Mandarin and Tibetan, including silence and pause, are used as the speech synthesis unit. A six-level HMM-based speech synthesis system (HTS) context-dependent label format is designed by taking into account the following context-dependent features [18].

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

B. QUESTION SET

The essence of TTS based on deep learning is to use deep neural networks to realize the nonlinear mapping between input context-dependent labels and output acoustic features. Therefore, we need to transform the HTS-based context-dependent labels into continuous or binary vectors based on the designed question set to serve as the input of deep neural networks. Different from context-dependent labels, the question set consists of 44 classes of pronunciation information and 31 classes of prosody information, which mainly classify the basic features of pronunciation units and prosodic information.

We use a context-dependent question set of Mandarin to design a common question set for both Chinese and Tibetan. The question set extends the related questions of Tibetan-specific synthesis units to reflect the special pronunciation of Tibetan. Therefore, the question set covers the prosodic features and pronunciation information of Mandarin and Tibetan. We designed two kinds of questions starting with QS and starting with CQS, respectively, as shown below.

- QS question expression {answer1, answer2,$\cdots$ }
  For example, QS "C==Alveolar" {*-d+*,*-t+*,*-n+*,*-l+*,*-lh+*}. The expression " C==Alveolar" indicates whether the current syllable is Alveolar? The answer is {*-d+*,*-t+*,*-n+*,*-l+*,*-lh+*}.
- CQS question expression {answer}
  For example, CQS "C-Word_Num-Syls" {_(\d+)+}. The expression " C-Word_Num-Syls" indicates how many syllables the current word has. The answer is d.

TABLE 1. Architectures of DNN, hybrid LSTM and hybrid BLSTM.

Type	Type of layer(s)	Number of layer(s)	Number of units (memory blocks)
DNN	DNN	4	512
hybrid LSTM	DNN	3	512
	LSTM	1	256
hybrid BLSTM	DNN	3	512
	BLSTM	1	256

The question set starting with QS is similar to the question set in HTS format. It is used to transform the context-dependent labels into binary vectors. The question starting with CQS is used to transform the context-dependent labels into continuous vectors.

VI. EXPERIMENTS

A. CORPUS

In the experiment, we used one male and seven female speakers' recordings (169 sentences per person, total 1352 sentences) from the EMIME Bilingual Speech Database [33] as the Mandarin corpus. We selected a female speaker's 800 utterances of Tibetan Lhasa dialects as the Tibetan corpus. All recordings and synthesized speeches were saved in the Microsoft Windows WAV format as sound files (mono-channel, signed 16 bits, sampled at 16 kHz).

B. EXPERIMENTAL SETUP

Three kinds of TTS frameworks, including DNN, hybrid LSTM, and hybrid BLSTM, were compared in the experiments. The architectures of these frameworks are shown in Table 1.

The DNN-based framework consists of four feed-forward hidden layers with 512 units. We used the mini-batch stochastic gradient descent (SGD) algorithm to train the acoustic model. The mini-batch size was 128. In the training AVM, the momentum was fixed to 0.9. The learning rate was fixed to 0.002 for the first 10 epochs and then halved for the remaining epoches. In the speaker adaptation, the momentum was also fixed to 0.9. The learning rate was fixed to 0.001 for the first 10 epochs and then halved for the remaining epochs.

The hybrid LSTM-based framework consists of three hidden DNN layers with 512 units and a single LSTM layer with 256 memory blocks. The upper hidden LSTM layer follows the lower hidden DNN layers. The parameters of models were randomly initialized and trained using the backpropagation through time (BPTT) algorithm. In the training AVM, the learning rate was fixed to 0.0015 for the first 20 epochs and then halved for the remaining epochs. In the speaker adaptation, the learning rate was fixed to 0.002 for the first 5 epochs and then halved for the remaining epochs. Adam optimizer was performed in the training AVM and speaker adaptation stage.

The hybrid BLSTM-based framework consists of three hidden DNN layers with 512 units and a single BLSTM layer with 256 memory blocks. The upper hidden BLSTM layer follows the lower hidden DNN layers. The training schedule was the same as the hybrid LSTM-based network.

Each framework realizes the monolingual synthesis of Mandarin and Tibetan as well as the cross-lingual synthesis of Mandarin and Tibetan. In the monolingual synthesis of Mandarin, the input vectors of all neural networks consist of 425 dimension features, of which 416 are derived from context-dependent labels to reflect linguistic features. Due to the more complex linguistic features of Tibetan compared to Mandarin, the input vectors of all neural networks consist of 1,183 dimension features, of which 1,174 are derived from context-dependent labels to reflect the linguistic features in the monolingual synthesis of Tibetan. For the cross-lingual synthesis of Mandarin and Tibetan, the input feature vectors of all neural networks include 1,255 dimensions, 1,246 of which are derived from context-dependent labels. The remaining 9 of all neural network input vectors are frame position information in phoneme and HMM states. The frame alignment and state information were obtained from the forced alignment using a mono phone HMM-based system with 5 emitting states per phone.

We use the WORLD vocoder to extract 60-dimensional Mel-generalized cepstral (MGC) coefficients, 5-dimensional distortion of band aperiodicities (BAP) coefficients, and 1-dimensional logF0. The output vectors of the neural network include MGC, BAP, logF0, and their deltas and delta-deltas in addition to one-dimensional voiced/unvoiced binary features, totaling 187 dimensions.

The input features are normalized by min-max to [0.01, 0.99], while the output features are normalized by 0 mean and unit variance before training. In the speech synthesis stage, the WORLD vocoder is also used to generate waveforms.

To evaluate the quality of synthesized speech, we trained the following models for three TTS frameworks.

Tibetan speaker-dependent (TSD) model: We use one female speaker's Tibetan corpus (800 utterances) to train the TSD model. We compare the quality of synthesized speech in three frameworks, including TSD-DNN, TSD-LSTM, and TSD-BLSTM.

Mandarin speaker-dependent (MSD) model: 7 female speakers' Mandarin corpus (a total of 1,183 utterances) are firstly used to train the Mandarin speaker-independent average voice model. Then one male speaker's corpus (a total of 169 utterances) is used to train the MSD module. We compare the quality of synthesized speech in three frameworks, including MSD-DNN, MSD-LSTM, and MSD-BLSTM.

Speaker adapted target language model: First, we use a one female Tibetan speaker's corpus and seven female Mandarin speakers' corpus to train the Mandarin-Tibetan cross-lingual AVM model. Then we train the Tibetan speaker adapted-target (SATT) language model and the Mandarin speaker adapted-target (SATM) language model from the AVM model by using speaker adaptation techniques. The SATT model is transformed from the AVM model by using 800 Tibetan training utterances. The SATM model is

TABLE 2. Objective results of TSD model.

Type	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	V/UV (%)	MODEL SIZE (MByte)
DNN	7.812	0.174	33.648	6.832	54.078
LSTM	7.619	0.172	33.234	6.708	60.166
BLSTM	7.606	0.170	33.044	6.658	79.292

TABLE 3. Objective results of MSD model.

	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	V/UV (%)	MODEL SIZE (MByte)
DNN	5.050	0.170	13.423	5.778	40.443
LSTM	4.984	0.170	12.969	5.940	46.626
BLSTM	4.959	0.167	12.671	5.920	65.821

TABLE 4. Objective results of SATM model for Mandarin.

	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	V/UV (%)	MODEL SIZE (MByte)
DNN	5.053	0.174	13.554	5.829	55.267
LSTM	5.033	0.173	13.519	6.321	70.600
BLSTM	4.959	0.169	12.374	6.108	93.585

TABLE 5. Objective results of SATT model for Tibetan.

	MCD (dB)	BAP (dB)	F0 RMSE (Hz)	V/UV (%)	MODEL SIZE (MByte)
DNN	7.357	0.174	31.243	6.484	55.225
LSTM	7.202	0.176	30.639	6.012	70.597
BLSTM	7.090	0.173	30.089	5.391	93.586

transformed from the AVM model by using one male Mandarin training utterances. We also compare the quality of synthesized speech in three frameworks, including SATT-DNN, SATM-DNN, SATT-LSTM, SATM-LSTM, SATT-BLSTM, and SATM-BLSTM.

In the experiment, 10% of the utterances were randomly selected as the test set, 10% of the utterances were used as the development set, and the remaining were used as the training set.

C. OBJECTIVE EVALUATIONS

In the objective evaluations, we calculated the distortions between the acoustic parameters of the original speeches and the predicted parameters of each model, including Mel-cepstral distortion (MCD), distortion of BAP, F0 root-mean-square error (RMSE) and voiced/unvoiced error (V/UV). The objective evaluation results of the TSD model and MSD model are shown in Table 2 and Table 3, respectively. The objective evaluation results of the SAT for Mandarin (SATM) and Tibetan (SATT) are shown in Table 4 and Table 5, respectively.

D. SUBJECTIVE EVALUATIONS

For subjective evaluations, 20 utterances were randomly selected from the test set. We conducted a mean opinion score (MOS) test, degradation mean opinion score (DMOS) test, and AB preference test to evaluate the quality of synthesized speech. We invited 30 native Mandarin and 30 native Tibetan listeners as subjects. Mandarin subjects were invited to evaluate the MSD and SATM models, while Tibetan subjects were invited to assess the TSD and SATT models. In the MOS test, subjects were asked to rate the naturalness of the synthesized speech using a 5-point scale score. The average MOS scores of synthesized Tibetan and Mandarin speech are shown in Fig. 7 and Fig. 8.

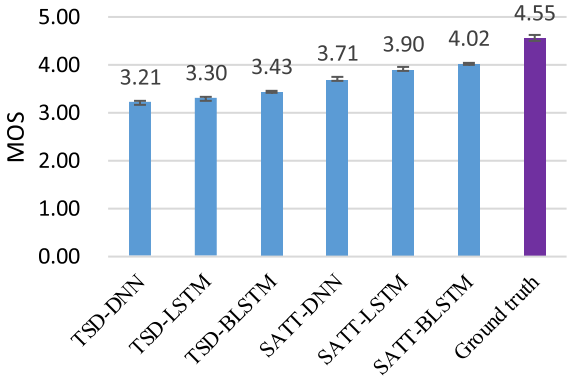


FIGURE 7. The average MOS scores of synthesized Tibetan speech under 95% confidence intervals.

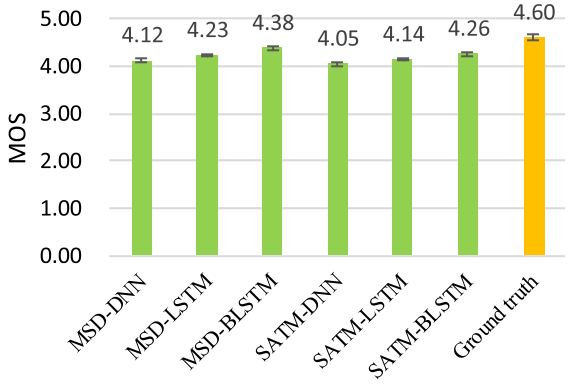


FIGURE 8. The average MOS scores of synthesized Mandarin speech under 95% confidence intervals.

In the DMOS test, the synthesized utterances and their corresponding original recording formed a pair of speech files for each model. We randomly played each pair of speech files to the subjects with the order of the original speech after synthesized speech. The subjects were asked to carefully compare these two files and evaluate the degree of similarity of the synthesized speech to the original speech on a 5-point scale score. A 5-point score represents synthesized speech that is very close to the original speech, while a 1-point score represents synthesized speech that is very different from the original speech. The average DMOS scores of synthesized Tibetan and Mandarin speech are shown in Fig. 9 and Fig. 10.

In the AB preference test, the sentences were the same in both items within a pair. 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''

TABLE 6. Subjective AB preference score(%) of Tibetan with $p < 0.01$.

	TSD-DNN	TSD-LSTM	TSD-BLSTM	SATT-DNN	SATT-LSTM	SATT-BLSTM	Neutral
1	22.2	67.5	-	-	-	-	10.3
2	19.0	-	70.7	-	-	-	10.3
3	-	20.3	69.2	-	-	-	10.5
4	-	-	-	18.8	69.5	-	11.7
5	-	-	-	16.8	-	71.3	11.8
6	-	-	-	-	17.1	71.0	11.8
7	-	-	14.8	-	-	75.0	10.2

TABLE 7. Subjective AB preference score(%) of Mandarin with $p < 0.01$.

	MSD-DNN	MSD-LSTM	MSD-BLSTM	SATM-DNN	SATM-LSTM	SATM-BLSTM	Neutral
1	21.2	68.3	-	-	-	-	10.5
2	18.7	-	72.1	-	-	-	9.2
3	-	19.2	70.3	-	-	-	10.5
4	-	-	-	20.5	68.8	-	10.7
5	-	-	-	19.0	-	70.8	10.2
6	-	-	-	-	19.5	71.0	9.5
7[a]	-	-	41.0	-	-	39.2	19.8

[a] $p = 0.04$

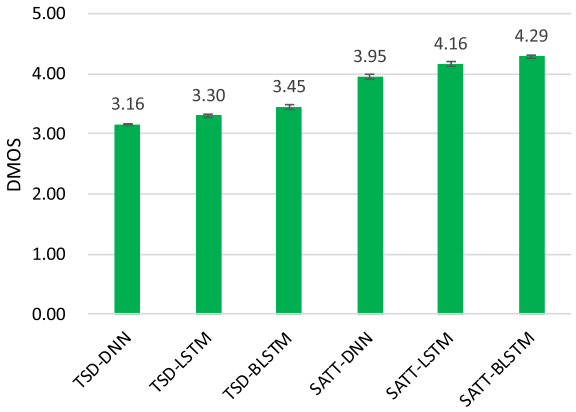


FIGURE 9. The average DMOS scores of synthesized Tibetan speech under 95% confidence intervals.

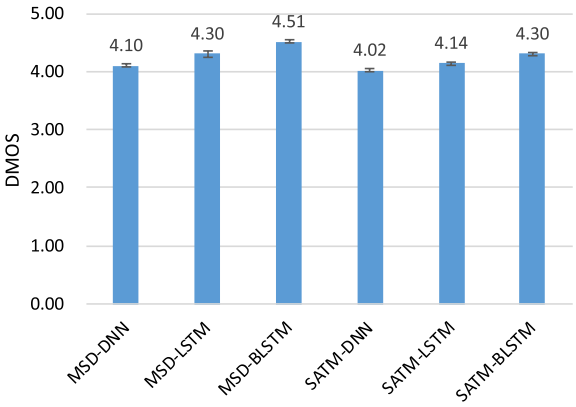


FIGURE 10. The average DMOS scores of synthesized Mandarin speech under 95% confidence intervals.

means that the subjects had no preference). The preference results of the synthesized Tibetan and Mandarin speech are shown in Table 6 and Table 7.

In the speech synthesis of low-resource languages, it is significant to study the smallest corpus for synthesizing speech with satisfactory voice quality. Based on the above experimental setup, we used {100, 200, 350, 500} Tibetan training utterances of one female speaker and a seven-female Mandarin corpus to train the BLSTM-based AVM model. In speaker adaptive transformation, the training Tibetan utterances were {100, 200, 350, 500}, respectively. The objective measures of different settings are shown in Table 8. When the number of training utterances was approximately 500 utterances, the MCD was 7.388, BAP was 0.173, F0 RMSE was 33.038, and V/UV swapping errors were 6.708. These measurements are similar to those of the monolingual BLSTM framework trained with 800 utterances.

VII. DISCUSSION

In objective evaluations, the findings can be analyzed as follows.

Firstly, although the DNN-based TTS framework map linguistic features to acoustic features frame by frame through multiple hidden layers, the temporal information of speech is not explicitly modeled.

Secondly, the hybrid LSTM network can directly model temporal information of the speech. However, the hybrid BLSTM network can capture the forward and backward input features for a given frame. Therefore, the acoustic model of the hybrid BLSTM-based TTS framework is better than other frameworks.

Thirdly, the size of the hybrid BLSTM model of each TTS framework is the largest, and the size of the DNN model is the smallest, which is also consistent with the training time of the model. The training time of the hybrid BLSTM model is about three times that of the DNN model. At the same time, because the SATM model and SATT model have the

TABLE 8. Objective evaluation results on different number of the training utterances.

Utterances	MCD(dB)	BAP(dB)	F0 RMSE(Hz)	V/UV(%)
100	8.677	0.188	40.163	8.199
200	8.295	0.185	36.917	7.304
350	8.006	0.176	36.161	6.957
500	7.388	0.173	33.038	6.708

same architecture, the model size of each category is almost the same. Finally, for Tibetan, the objective results of the SATT model are better than that of the TSD model. That is because Mandarin and Tibetan belong to the Sino-Tibetan language family, so they have many internal similarities. A set of SAMPA is designed to label the pronunciation of the initial and the final of both Mandarin and Tibetan syllables according to the similarities in pronunciation between Mandarin and Tibetan. The same pronunciation of Mandarin and Tibetan is represented as the same symbol in SAMPA. Interestingly, the deep neural network can learn these similarities. Therefore, we find that the quality of synthesized Tibetan speech can be improved by adding a Mandarin corpus in the TTS experiment of low resource language. If there are similarities between the two languages (such as Mandarin, Tibetan, and Yi), our framework can be applied to those language's cross-lingual speech synthesis.

All subjective evaluations are consistent with objective evaluations in several ways.

- The hybrid BLSTM-based TTS framework provides the best quality of speech in naturalness, speaker similarity and preference of synthesized speech.
- Due to the supplement of the Mandarin corpus, the voice quality of synthesized Tibetan speech by the cross-lingual TTS frameworks is better than that of the monolingual TTS frameworks.
- In the AB preference test, we can further confirm that the synthesized Mandarin speeches by the cross-lingual TTS frameworks are not significantly different from the speech synthesized by the monolingual Mandarin TTS framework.

From the results of Table 8, we can also find that the hybrid BLSTM-based cross-lingual speech synthesis framework only needs 60% of the training corpus to synthesize a similar voice as the Tibetan monolingual framework. It may be of guiding significance to the recording of the low resource TTS corpus.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we further our previous work to realize a deep-learning-based Mandarin-Tibetan cross-lingual speech synthesis. We adopted the DNN, hybrid LSTM, and hybrid BLSTM in work to improve the voice quality of synthesized Mandarin and Tibetan speech. We also finished a complete Tibetan text analyzer to improve our previous work. Therefore, the work realized a integral Mandarin-Tibetan cross-lingual text-to-speech synthesis that can convert any Chinese or Tibetan sentence to corresponding Mandarin speech or Tibetan speech. Objective and subjective experiments demonstrated that the hybrid BLSTM-based cross-lingual speech synthesis framework was not only better than the other two cross-lingual frameworks, but also the Tibetan monolingual framework. The mixed Tibetan training corpus did not influence the voice quality of synthesized Mandarin speech. Furthermore, we also investigated how to use the least corpus for synthesizing Tibetan speech with satisfactory voice quality. Therefore our method would be valuable to construct a speech synthesis system for low resource minority languages by borrowing a similar majority language's speech resources. Future work will focus on expanding the corpus and realize speech synthesis of other minority languages (such as Yi, and Kham dialect) by borrowing large-scale Mandarin corpus. Further work will also attempt to study how to use new methods such as end-to-end framework and reinforcement method to realize a cross-lingual speech synthesis for improving the voice quality and expressiveness of synthesized speech.

REFERENCES

[1] H. Bourlard, J. Dines, P. N. Garner, D. Imseng, P. Motlicek, H. Liang, L. Saheer, F. Valente, and M. Magimai-Doss, "Current trends in multilingual speech processing," *Sadhana*, vol. 36, no. 5, pp. 885–915, Oct. 2011.

[2] F.-L. Xie, F. K. Soong, and H. Li, "A KL divergence and DNN approach to cross-lingual TTS," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5515–5519.

[3] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Atlanta, GA, USA, May 1996, pp. 373–376.

[4] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, Nov. 2009.

[5] 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.

[6] S. Yang, H. Lu, S. Kang, D. Yu, and L. Xie, "Enhancing hybrid self-attention structure with relative-position-aware bias for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 6910–6914.

[7] Z.-H. Ling, S.-Y. Kang, A. Senior, M. Schuster, X.-J. Qian, H. M. Meng, L. Deng, H. Zen, "Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 1234–1252, Apr. 2015.

[8] S. Yang, Z. Wu, and L. Xie, "On the training of DNN-based average voice model for speech synthesis," in *Proc. Signal Inf. Process. Assoc. Summit Conf.*, Jeju-do, South Korea, Jan. 2017, pp. 1–6.

[9] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 4475–4479.

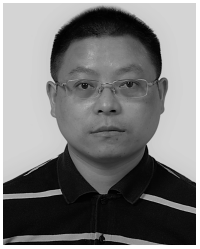
[10] E. Song, F. K. Soong, and H.-G. Kang, "Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2152–2161, Nov. 2017.

[11] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," Apr. 2017, *arXiv:1703.10135*. [Online]. Available: https://arxiv.org/abs/1703.10135

[12] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 4479–4783.

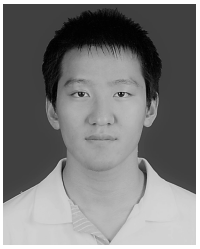
[13] S. Xu, H. Yu, and G. Li, "The influence of context on Tibetan Lhasa speech synthesis," in *Proc. IEEE 2nd Adv. Inf. Technol., Electron. Automat. Control Conf. (IAEAC)*, Chongqing, China, Mar. 2017, pp. 625–629.
[14] F. Deprez, F. Odijk, and J. De Moortel, "Introduction to multilingual corpus-based concatenative speech synthesis," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, Antwerp, Belgium, 2007, pp. 2129–2132.
[15] W. Zhiyong, C. Guangqi, M. H. Meng, and L. Cai, "A unified framework for multilingual text-to-speech synthesis with SSML specification as interface," *Tsinghua Sci. Technol.*, vol. 14, no. 5, pp. 623–630, Oct. 2009.
[16] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Cross-lingual speaker adaptation for statistical speech synthesis using limited data," in *Proc. 17th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, San Francisco, CA, USA, 2016, pp. 317–321.
[17] S. S. Sarfjoo, C. Demiroğlu, and S. King, "Using eigenvoices and nearest-neighbors in HMM-Based cross-lingual speaker adaptation with limited data," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 4, pp. 839–851, Apr. 2017.
[18] H. Yang, K. Oura, Z. Gan, K. Tokuda, and H. Wang, "Using speaker adaptive training to realize Mandarin-Tibetan cross-lingual speech synthesis," *Multimed Tools Appl.*, vol. 74, pp. 9927–9942, Nov. 2015.
[19] H. Wang, H. Yang, and Z. Gan, "Realizing Mandarin-Tibetan bilingual speech synthesis by speaker adaptive training," *J. Tsinghua Univ., Sci. Technol.*, vol. 53, no. 6, pp. 776–780, Jun. 2013.
[20] Q. Yu, P. Liu, S. K. Ang, H. Meng, L. Cai, and Z. Wu, "Learning cross-lingual information with multilingual BLSTM for speech synthesis of low-resource languages," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 5545–5549.
[21] W. Guo, H. Yang, and Z. Gan, "A DNN-based Mandarin-Tibetan cross-lingual speech synthesis," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Honolulu, HI, USA, 2018, pp. 1702–1707.
[22] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, May 2013, pp. 7962–7966.
[23] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Dresden, Germany, 2015, pp. 879–883.
[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[25] Y. Fan, Y. Qian, F. L. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc. (Interspeech)*, Singapore, 2014, pp. 1964–1968.
[26] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.
[27] N. Wanmezhaxi, "Research on several key issues in Tibetan word segmentation," *J. Chin. Inf. Process.*, vol. 28, no. 4, pp. 132–139, Jul. 2014.
[28] Y. Li, Y. Jia, X. He, and H. Yu, "Study on fusion of unsupervised features for Tibetan word segmentation," *J. Chin. Inf. Process.*, vol. 31, no. 2, pp. 71–85, Mar. 2017.
[29] Y. Li, Y. Jia, X. He, and H. Yu, "An open source toolkit for Tibetan word segmentation and POS tagging," *J. Chin. Inf. Process.*, vol. 29, no. 6, pp. 203–207, Nov. 2015.
[30] G. Lample, M. Ballesteros, K. Kawakami, C. Dyer, and S. Subramanian, "Neural architectures for named entity recognition," Apr. 2016, *arXiv:1603.01360*. [Online]. Available: https://arxiv.org/abs/1603.01360
[31] J. Lafferty, A. McCallum, F. Xie, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, San Francisco, CA, USA, Jun. 2001, pp. 282–289.
[32] H. Yang and L. Zhu, "Predicting Chinese prosodic boundary based on syntactic features," *J. Northwest Normal Univ., Natural Sci.*, vol. 49, no. 1, pp. 41–45, 2013.
[33] M. Wester, "The EMIME bilingual database," Univ. Edinburgh, Edinburgh, U.K., Tech. Rep. EDI-INF-RR-1388, 2010.

WEIZHAO ZHANG was born in Qingyang, Gansu, China, in 1987. He received the B.S. degree in electronic and information engineering and the master's degree in circuit and system from Northwest Normal University, in 2008 and 2011, respectively, where he is currently pursuing the Ph.D. degree with the College of Physics and Electronic Engineering. He is currently a Lecturer with the College of Physics and Electronic Engineering, Northwest Normal University. His current research interests include multilingual speech synthesis and expressive speech synthesis and recognition.



HONGWU YANG (M'05) was born in Hezuo, Gansu, China, in 1969. He received the B.S. degree in physics and the M.S. degree in education from Northwest Normal University, Lanzhou, China, in 1992 and 1995, respectively, and the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2007.

From 1995 to 2009, he was a Teaching Assistant, a Lecturer, and an Associate Professor with Northwest Normal University. Since 2009, he has been a Professor with Northwest Normal University. He is the author of two books, more than 50 articles, and more than 30 inventions. His research interests include speech signal processing, speech recognition, speech synthesis, and artificial education.



XIAOLONG BU was born in Lanzhou, Gansu, China, in 1993. He received the B.S. degree in electronic and information engineering from Nanjing Agricultural University, in 2016. He is currently pursuing the master's degree with the College of Physics and Electronic Engineering, Northwest Normal University. His main research interests include speech synthesis and speech man–machine interaction.

LILI WANG was born in Lanzhou, Gansu, China, in 1994. She received the B.S. degree in electronic information science and technology from Chongqing Normal University, in 2017. She is currently pursuing the master's degree with the College of Physics and Electronic Engineering, Northwest Normal University. Her main research interests include natural language processing and speech synthesis.