A DNN-based Mandarin-Tibetan cross-lingual speech synthesis

Weitong GUO*† and Hongwu YANG*‡§ and Zhenye GAN*‡
* College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China
† Ubiquitous Awareness and Intelligent Solutions Lab, Lanzhou University, Lanzhou 730070, China
‡ Engineering Research Center of Gansu Province for Intelligent Information Technology and Application, Lanzhou 730070, China
§ National and Provincial Joint Engineering Laboratory of Learning Analysis Technology in Online Education, Lanzhou 730070, China
E-mail: yanghw@nwnu.edu.cn

***Abstract*—The paper proposed a deep neural network (DNN)-based Mandarin-Tibetan cross-lingual speech synthesis by adopting speaker adaptation training. The initial and the final are used as the speech synthesis units for both Mandarin and Tibetan to train a set of average voice model(AVM) based on DNN from a large Mandarin multi-speaker corpus and a small Tibetan one-speaker corpus. The speaker adaption is adopted to train a set of speaker-dependent DNN models of Mandarin or Tibetan appended with AVM. The Mandarin speech or Tibetan speech is then synthesized by their respective speaker-dependent DNN acoustic models. Both subjective evaluations and objective tests show that synthesized Tibetan speech by the proposed method are not only better than the traditional Hidden Markov Model(HMM)-based method, but also better than the DNN-based Tibetan speech synthesis with only Tibetan training corpus. Mixed Tibetan training speech have little effect on the quality of synthesized Mandarin speech. Therefore, the proposed method can be applied to the speech synthesis of minority language with rare speech resources.**

I. Introduction

Multilingual speech synthesis has been a hot area of research on speech synthesis for several years [1]. Because multilingual speech synthesis can synthesize speech of same or different speaker with only a speech synthesis system, it has been extensively used in human-computer interaction, bilingual education, spoken dialogue system and so on. Generally speaking, there are two traditional methods to realize cross-lingual speech synthesis. One is large corpus-based unit selection waveform concatenation speech synthesis method [2] [3] [4], which is difficult to record different languages's speech of one speaker. The other is the hidden Markov model (HMM)-based statistical parametric speech synthesis method [5] [6] [7], which can easily synthesize voice of different speakers with speaker adaptation algorithms. However, this kind of method traditionally uses shallow gaussian mixture models (GMMs) to model the acoustic features with greedy algorithm, so that the acoustic feature models are suboptimum and the synthesized speech has a low naturalness.

Since 2010, deep learning techniques have been successfully applied to the modeling of speech signals [8]. Inspired by the successful application of deep neural networks(DNNs) to speech recognition [9], DNNs have also been applied to statistical parametric speech synthesis and achieved significant improvements [8]. The interaction between input context and output acoustic features is modeled by a DNN In [10], which can address some limitations of conventional HMM-based speech synthesis approaches. Wang et al. [11] achieved DNN-based speech synthesis on the basis of the HMM speech synthesis framework to set up the speech spectrum conversion model and the experiment proved that the method can effectively improve the synthesized speech quality. Lu et al. [12] used a vector-space representation of linguistic context as inputs of DNN-based speech synthesis and probability distributions over speech features as outputs of the network. Maximum Likelihood(ML) parameter generation was used to generate parameter contour, and then drove a vocoder to generate the speech waveform. Fernandez et al. [13] proposed a hybrid learning approach between DNN and Gaussian process(GP)-based regression to predict logf0. DNN was first trained from the raw inputs, and then the activations at the last hidden layer were used as inputs for GP-based nonparametric regression. Subjective and objective evaluation showed that this approach can implement parametric synthesis for the prediction of prosodic targets. Qian et al. [14] showed that DNN-based speech synthesis with a moderate size corpus outperform the HMM-based baseline, in particular the prosody. The weights of DNN were trained by using pairs of input context and output acoustic feature to minimize the errors between the mapped output from a given input and the target output. Subsequently, speaker adaptation techniques are used in DNN-based speech synthesis. [15] implemented a preliminary work on speaker adaptation for DNN-based speech synthesis was implement. Average voice models (AVM) from multiple speakers recordings were created to train DNNs through performing a speaker adaptive training (SAT). The result suggested that the adaptation technique in the DNN-based speech synthesis approach was superior to the standard feature transformations. Fan et al. [16] proposed a new unsupervised multi-speaker adaptation for DNN-based speech synthesis approach and it can achieve comparable performance with supervised adaptation. Wu et al. [17] augmented a low-dimensional speaker-specific vector with linguistic features as input to represent speaker identity.

Model adaptation was performed to scale the hidden activation weights. Then a feature space transformation was conducted at the output layer to modify generated acoustic features.

However, the state-of-the-art researches on speech synthesis are focusing on major language [15] [16] [17], which obtain plenty of data resources for model training, while it lacks of studies on minority nationality language in cross-lingual speech synthesis due to scarce speech resources such as Tibetan [18] [19]. To the best of our knowledge, speaker adaptation for DNN-based speech synthesis has not yet been applied on the Mandarin-Tibetan cross-lingual speech synthesis.

In this paper, we achieve a DNN-based Mandarin-Tibetan cross-lingual speech synthesis by adopting speaker adaptation training. According to the HMM-based speech synthesis framework, DNN acoustic models are trained with a large Mandarin multi-speaker-based corpus and a small Tibetan one-speaker-based corpus to replace HMM-based acoustic models. Firstly, we use a set of designed Speech Assessment Methods Phonetic Alphabet(SAMPA) to label the pronunciation for both Mandarin and Tibetan. A set of context-dependent label format is designed to label the context information of Mandarin and Tibetan. The initial and the final form the synthesis units for both Mandarin and Tibetan. Secondly, average voice model(AVM) based on DNN is trained using speaker adaptive training from a large Mandarin multi-speaker corpus and a small Tibetan one-speaker corpus. Finally, the speaker adaption is adopted to train a set of speaker-dependent DNN models of Mandarin or Tibetan with AVM. The Mandarin speech or Tibetan speech is then synthesized by their respective speaker-dependent DNN acoustic models. The results show that the proposed method can achieve a better Tibetan voice quality than the method of paper [19].

II. Framework of DNN-based Mandarin-Tibetan Cross-lingual Speech Synthesis

The framework of the proposed DNN-based speech synthesis consists of three parts including preparing training data module, training DNN acoustic models module and synthesizing speech module, as shown in Fig. 1. The Training DNN acoustic module includes average voice model (AVM) training and speaker adaptation.

III. Preparing Training Data

We use linguistic features as input and acoustics features as output to train the DNN-based acoustic models. Chinese sentences and Tibetan sentences are selected to be the text corpus. Speech corpus are recorded according to the corresponding sentences by Mandarin subjects or Tibetan subjects in sound-proof studio. Then we extract linguistic features from the sentences and acoustic features from the recorders.

A. linguistic features

The paper adopt all initials and finals of Mandarin and Tibetan, including silence and pause, as the synthesis unit.

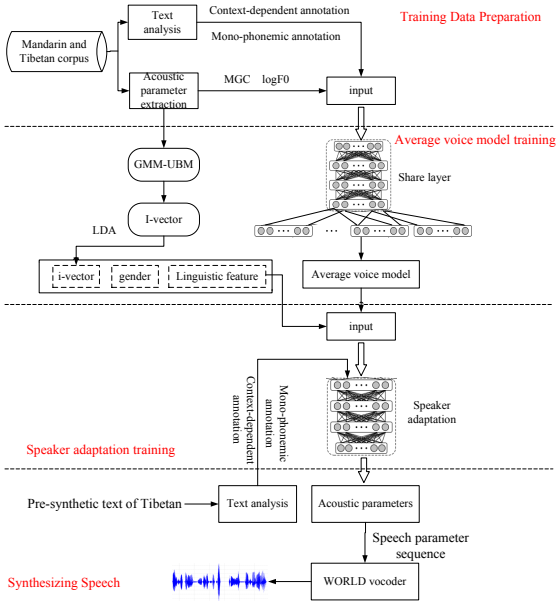

Fig. 1. Framework of DNN-based Mandarin-Tibetan cross-lingual speech synthesis.

Since Tibetan and Mandarin have many similarities in pronunciation [20], we design a set of SAMPA to label the pronunciation of initials and finals for both Mandarin and Tibetan. We also design a six level context-dependent label format for labeling the context information of the speech synthesis unit of Mandarin and Tibetan. The context-dependent label format reflects the unit level, syllable level, word level, prosodic word level, phrase level, and sentence level context information. We developed a mix-lingual text analyzer to obtain the context-dependent labels from both Chinese sentences and Tibetan sentences. The linguistic features for the DNN acoustic model training is obtained from the context-dependent label.

B. acoustic features

The acoustic features are extracted from the recorded speech files. We use the WORLD vocoder to extract acoustic features including the fundamental frequency (F0), the mel-generalized cepstral (MGC), the band a periodical (BAP), and the voiced /unvoiced decision. These acoustic features are used as the output for DNN acoustic models training.

IV. Training the DNN Acoustic Models

In this work, a speaker adaptation framework is employed for training the DNN-based acoustic models proposed in [21]. Speaker-independent DNN model is trained as an average voice model (AVM) with augmented i-vector to capture speaker identity. An i-vector represents speaker identity [22]. At the adaptation phase, the target speaker's i-vector is first estimated by using the adaptation data, and then the i-vector, a gender code and the context-dependent linguistic features are used as the network input to generate the target speaker's speech. DNN

is set up via lots of hidden units hierarchically structured into a sequence of layers for generating speaker-dependent acoustic parameters. The linear discriminant analysis(LDA) is adopted to obtain the best results.

A. training the average voice model

The paper trained a set of DNN models as the cross-lingual AVM by using the linguistic features (binary and digital) as input and acoustic features as output. The linguistic features were obtained from context-dependent label of the text corpus. The acoustic features were extracted from speech corpus including MGC, BAP, F0 and voice/unvoiced (V/UV). The DNN models share various hidden layers between different emotional speakers to model its language parameters. Duration models and acoustic feature models were trained by a stochastic gradient descent (SGD) [19] of back propagation (BP) algorithm. Finally, a set of speaker indenpent AVM were trained by the Mandarin and Tibetan multi-speaker corpus.

The paper used a DNN structure including an input layer, a output layer and 3 hidden layer to train the AVM. The $tanh$ is used in the hidden layer and the linear activation function is used in the output layer. All speakers' training corpus share the hidden layer, so the hidden layer is a global linguistic feature mapping shared by all speakers. Each speaker has its own regression layer to model its own specific acoustic space. After multiple batches of SGD training, a set of optimal multi-speaker AVM model (average duration models and average acoustic feature models) is obtained.

B. Speaker adaptation

In the speaker adaptation stage, a small corpus of the target languages(Mandarin or Tibetan) of the target speaker is used to extract acoustic features in the same way as AVM training, including LogF0, MGC, BAP and V/UV. Firstly, the speaker adaptation is performed by multi-speaker AVM model with the DNN models of the target language to obtain a set of speaker-dependent adaptation models including duration models and acoustic feature models. The speaker-dependent model has the same DNN structure as the AVM, using six hidden layer structures, and the mapping function is the same as the AVM.

C. synthesizing speech

In the speech synthesis stage, we firstly obtain the context-dependent labels from Mandarin sentences or Tibetan sentences by the mix-lingual text analyzer. At the same time, the maximum likelihood parameter generation (MLPG) algorithm [23] is used to generate the speech parameters from acoustic models. At last, The WORLD vocoder is used to synthesize speech from speech parameters.

V. Experiments

A. Corpus and experiment conditions

We select seven female speakers' recordings (169 sentences per person) from EMIME bi-lingual speech database [24] as Mandarin corpus, and record 800 Tibetan speech of a native female Tibetan Lhasa dialect speaker as Tibetan training corpus. Tibetan sentences are chosen from recent years Tibetan newspapers. All recordings are saved in the Microsoft Windows WAV format as sound files (mono-channel, signed 16 bits, sampled at 48 kHz). There are 591- dimensional input features of DNN in all, 482-dimensional features of which are obtained from the context-dependent information, including the contextual information of the synthesized units (initials and finals) and their positions information in syllables, words, prosodic words, prosodic phrases and sentences. 9-dimensional features are internal locations information of the synthesis units, including the position of the frame in the HMM state and the synthesis units, the position of the state in the forward-backward synthesis units, and the duration of the HMM state and the synthesis units. The remaining 100-dimensional features are the MGC and the logF0 as well as their first order difference(delta) and second-order difference(delta-delta) extracted from the speech. State information and frame alignment are obtained from force alignment with five states per synthesis unit. The output features of neural network are 109-dimension vectors, containing the MGC, logF0 and their first order and second-order differences, as well as the binary features of synthesis units.

Before the training, the input features are normalized to [0.01 0.09] by the min-max method, and the output features are normalized to the zero mean unit variance. In the stage of synthesis, the maximum likelihood parameter generation (MLPG) [25] is used to generate the smooth speech parameters from the output of the non-normalized neural network, and then the MGC is enhanced in the cepstrum domain by adopting the spectrum enhancement method to improve the naturalness of synthesized speech.

DNN is set to 5 layers, having 3 feed-forward hidden layers, and 1024 neurons set for each hidden layer. Hyperbolic tangent function is adopted in hidden layers and linear activation is finished at the output layer. In the training AVM and speaker adaptation stage, mini-epoch size is set to 256 and momentum is used to accelerate convergence. For the first 10 epochs, the momentum is fixed to 0.6 and then increased to 0.9. The fixed learning rate of 0.0008 is used in the first 10 epochs of AVM-DNN training. The speaker's learning rate is set to 0.02 during the speaker adaptation. The learning rate halved in each epoch after 10 epochs and $L2$ regularization is applied to the weight with 0.00001 penalty factor. Maximum number of epochs are set to 25 for AVM-DNN training and the speaker adaptation.

In the experiments, 100 Tibetan speech are randomly selected from 800 Tibetan speech corpus as test set, 10,100 and 700 Tibetan utterances are randomly selected respectively from the remaining 700 Tibetan corpus to form 3 Tibetan training sets. Simultaneously, all Mandarin recordings of seven female speakers(7 people $\times$ 169 sentences) andTibetan training sets are used to train the neural network.

In order to evaluate the quality of synthesized Mandarin speeches and Tibetan speeches, we carry out 3 different strategies with different training sets as shown below:

1) T: T indicates that DNN and HMM acoustic models

are trained with three Tibetan training sets (10, 100, 700 Tibetan utterances) respectively. The DNN models of Tibetan are labeled as: 10 (DNN), 100 (DNN), 700 (DNN). The HMM models of Tibetan are labeled as: 10T (HMM), 100T (HMM), 700T (HMM).

2) M: it means that all training corpora of only 7 female Mandarin speakers are trained in DNN and HMM acoustic models. The DNN models are labeled as M (DNN). The HMM models are labeled as M (HMM).
3) TM: TM represents that DNN and HMM acoustic models are trained with all training corpora of 3 Tibetan training sets (10, 100, 700 Tibetan utterances) respectively and 7 female Mandarin speakers. 10TM (DNN), 100TM (DNN), 700TM (DNN) are the Tibetan DNN models. The HMM acoustic models of Tibetan are labeled as: 10TM (HMM), 100TM (HMM), 700TM (HMM).

The HMM models in three different training sets (T, M, TM) trained in this paper are the same as the three HMM models (SD, SI, SAT) trained in [19] (T, M, TM are separately equal with SD, SI, SAT). Each of the above mentioned 14 models synthesizes 100 Tibetan utterances, totaling 1400 sentences (100 sentences * 14 models). 20 sentences are randomly selected from 100 synthesized Tibetan utterances of each model for evaluation, amounting to 280 sentences (20 sentences * 14 models) as Tibetan testing set of evaluation. 20 Mandarin speeches are synthesized form each of the M and TM models. All synthesized Mandarin utterances are used as the Mandarin testing set to evaluate the influence of synthesized Mandarin with DNN and HMM acoustic models.10 native Tibetan Lhasa dialect speakers and 10 Mandarin speakers are invited as evaluation subjects.

B. Speech quality evaluation

The mean opinion score (MOS) is used to evaluate the naturalness of synthesized Tibetan and Mandarin speech. All selected testing sets are used to be evaluated. The synthesized Tibetan test utterances of each model are randomly played to the 10 Tibetan subjects except M models. We ask the subjects to carefully listen to these utterances and score the naturalness of every utterance by a 5-point score. Mandarin evaluation method is the same as Tibetan evaluation.

Fig. 2 compares the average MOS scores of synthesized Tibetan speech and synthesized Mandarin speech with different methods. We can see from the results that the MOS score of synthesized Mandarin speech with both the HMM-based method and the DNN-based method in TM strategy tend to be stable with the increase of Tibetan training corpus. This means that mixing Tibetan training corpus does not affect the quality of synthesized Mandarin speech. The scores of synthesized Tibetan speech by each TM strategy and T strategy increase with the increase of Tibetan training sentences. At the same time, The scores of Tibetan with TM strategy outperform that of the T strategy. This means that mixing Mandarin training sentences is helpful for the naturalness of

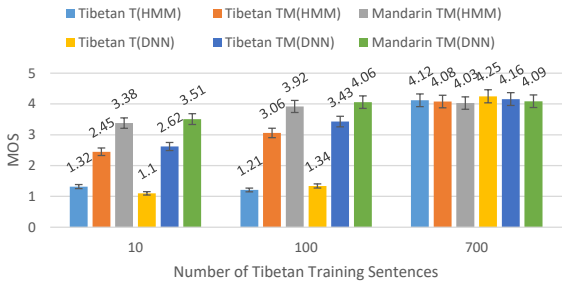

Fig. 2. The average MOS scores of synthesized Tibetan speech and Mandarin speech with different methods and training set under 95% confidence intervals.

synthesized Tibetan speech. The scores of DNN-based synthesized Tibetan with TM strategy are higher than that of the HMM-based method, which indicates that the proposed DNN-based Mandarin-Tibetan bilingual speech synthesis method is superior to the traditional HMM-based method in improving synthesized speech quality. For 10 Tibetan training utterances, the MOS score of synthesized Tibetan speech with the DNN-based method is greater than that of the HMM-based method in TM strategy. It indicates that the naturalness of synthesized Tibetan speech with the DNN-based method is better than that of the HMM-based method when mixing a small amount of Tibetan training corpus to the mixing corpus. It shows that proposed method is superior to the conventional HMM-based speech synthesis method in improving the speech quality of synthesized speech especially in the case of small Tibetan training corpus.

C. Speech similarity evaluation

We use the degradation mean opinion score(DMOS) to evaluate the speech similarity of synthesized Tibetan utterances and Mandarin utterances. In the DMOS evaluation, all the original recordings and the synthesized testing utterances (Tibetan and Mandarin) of each model are used. Each of the synthesized speech and its corresponding original recording form a pair of speech file. We randomly play each pair of speech files to the corresponding subjects with the order of the original speech followed by synthesized speech. The subjects are asked to carefully compare two speech files and evaluate the similarity of the synthesized speech to the original speech at 5-point score. The score 5 represents the synthesized speech is very close to the original speech while the score 1 represents the synthesized speech is very different from the original speech.

Fig. 3 compares the DMOS scores of synthesized speech by using different Tibetan training sets, in which the synthesized Tibetan speech and synthesized Mandarin speech are selected from each model of the 3 strategies (T, M, TM). We can see from Fig. 3 that the DMOS scores of synthesized Tibetan speech without mixing Tibetan training corpus are close to those of the synthesized Tibetan speech by only 100 Tibetan training sentences. The DMOS score

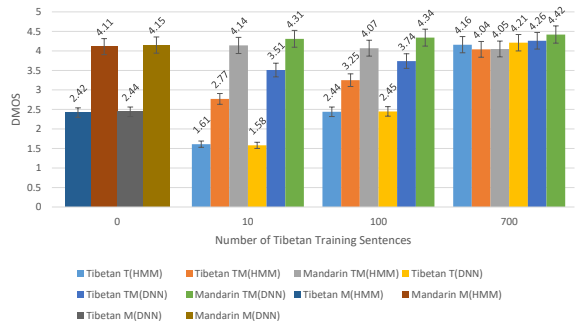

Fig. 3. DMOS evaluation of synthesized speech by using different Tibetan training sets.

of synthesized Tibetan with TM strategy is higher than that of T strategy for a certain Tibetan corpus, which indicates that when mixed Mandarin-Tibetan bilingual corpus is used to synthesize Tibetan, the speech similarity of synthesized Tibetan speech can be improved by sharing the units of Mandarin. For the 10 and 100 Tibetan training sentences, the score of DNN-based synthesized Tibetan is higher than that of the HMM-based method in TM strategy, which shows that the proposed method is preferable to the HMM-based method with a small amount of Tibetan training corpus. With the increase of Tibetan training corpus, the DMOS scores of synthesized Mandarin speech have not changed much. This means mixed Tibetan training corpus has little effect on the naturalness of synthesized Mandarin speech.

D. Objective evaluation

Root Mean Square Error (RMSE) analysis of duration and fundamental frequency are performed on all synthesized Tibetan and Mandarin speech. Table I and Table II show the RMSE of the duration and fundamental frequency of synthesized Tibetan speech by each model of the 3 strategies. Because the duration of the DNN-based method is consistent with that of the HMM-based method, we can see from the two tables that the RMSE of the duration are same on the synthesized speech for both Tibetan and Mandarin by the two methods. The RMSE of fundamental frequency whit the DNN-based method are small than that of the HMM-based method for both Tibetan and Mandarin, which indicates that DNN-based method is superior to the HMM-based method in modeling acoustic features. We also can see from Table I that with the increase of Tibetan training sentences, the RMSE of duration and fundamental frequency of synthesized Tibetan speech with mixed Mandarin and Tibetan training sentences(for example for TM strategy) is lower than those of synthesized Tibetan speech only with Tibetan training sentences(for example T strategy). This means that the mixing Mandarin training sentences can improve the voice quality of the synthesized Tibetan speech. From Table II, we can see that with the increase of Tibetan training sentences, the RMSE of duration and fundamental frequency of synthesized

TABLE I
RMSE OF DURATION AND FUNDAMENTAL FREQUENCY FOR SYNTHESIZED TIBETAN SPEECH.

Tibetan models	*durRMSE (s)*		*f0RMSE (Hz)*	
	HMM	*DNN*	*HMM*	*DNN*
M	12.49	12.49	17.78	17.58
10T	15.89	15.89	25.90	25.73
100T	15.03	15.03	28.72	25.82
700T	14.99	14.99	24.45	23.11
10TM	15.47	15.47	25.58	12.28
100TM	14.86	14.86	16.94	11.82
700TM	13.01	13.01	16.18	11.44

TABLE II
RMSE OF DURATION AND FUNDAMENTAL FREQUENCY FOR SYNTHESIZED MANDARIN SPEECH.

Mandarin models	*durRMSE (s)*		*f0RMSE (Hz)*	
	HMM	*DNN*	*HMM*	*DNN*
M	7.46	7.46	15.21	10.12
10TM	7.96	7.96	21.28	17.65
100TM	7.70	7.70	19.50	17.62
700TM	7.41	7.41	19.45	17.50

Mandarin speech on TM strategy is gradually reduced, which indicates that mixing Tibetan training corpus is also can improve Mandarin speech synthesis.

VI. CONCLUSIONS

On the basis of the traditional HMM-based Mandarin-Tibetan bilingual speech synthesis, the paper proposes a DNN-based method to realize Mandarin-Tibetan cross-lingual speech synthesis. The experimental results show that synthesized speech by the method is better than that of the HMM-based method. At the same time, adding Mandarin training corpus can improve the quality of the synthesized Tibetan speech. It indicates that this method can be used to realize a Mandarin-Minority language cross-lingual speech synthesis with a small amount of training corpus of Minority language by the mature Mandarin speech synthesis framework. Further work will attempt to improve the synthesized speech quality of DNN-based method by using multi-speaker-based speech database including males and females speakers and implementing speaker adaptive transformation in DNN model.

ACKNOWLEDGMENT

The research leading to these results was partly funded by the National Natural Science Foundation of China (Grant No. 11664036, 61263036) and high school science and technology innovation team project of Gansu (2017C-03), Natural Science Foundation of Gansu (Grant No. 1506RJYA126), and Student Innovation Project of Northwest Normal University (CX2018Y162).

REFERENCES

[1] H. Bourlard, J. Dines, M. Magimai-Doss, P. Garner, D. Imseng, P. Motlicek, H. Liang, L. Saheer, and F. Valente, "Current trends in multilingual speech processing," *Sadhana*, vol. 36, pp. 885–915, 2011.

[2] F. Deprez, J. Odijk, and J. D. Moortel, "Introduction to multilingual corpus-based concatenative speech synthesis," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.
[3] Z.Y. Wu, G.Q. Cao, M. L. Meng, and L. H. Cai, "A unified framework for multilingual text-to-speech synthesis with ssml specification as interface," *Tsinghua Science and Technology*, vol. 14, no. 5, pp. 623–630, 2009.
[4] Y. Zhang and J. Tao, "Prosody modification on mixed-language speech synthesis," in *ISCSLP'08*, 2008, pp. 1–4.
[5] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.
[6] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7962–7966.
[7] J. Wang and Y.Y. Zhang, "Title research on deep neural network based chinese speech synthesis," *Computer Science*, vol. 42, no. S1, pp. 75–78, 2015.
[8] H. Lu, S. King, and O. Watts, "Combining a vector space representation of linguistic context with a deep neural network for text-to-speech synthesis," in *Proc.ISCA SSW8*, 2013, pp. 261–265.
[9] R. Fernandez, A. Rendel, and R. Ramabhadran, B.and Hoory, "f0 contour prediction with a deep belief network-gaussian process hybrid model," in *Acoustics, Speech and Signal Processing*, 2013, p. 68856889.
[10] Y. Qian, Y.-C. Fan, W.-P . Hu, and F. K. Soong, "On the training aspects of deep neural network (dnn) for parametric tts synthesis," in *ICASSP2014*, 2014, pp. 3829–3833.
[11] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *ICASSP2015*, 2015, pp. 4460–4464.
[12] P. Potard, B.and Motlicek and D. Imseng, "Preliminary work on speaker adaptation for dnn-based speech synthesis," Tech. Rep., Idiap, 2015.
[13] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Unsupervised speaker adaptation for dnn-based tts synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5135–5139.
[14] Z. Wu, P. Swietojanski, C. Veaux, and S. Renals, S.and King, "A study of speaker adaptation for dnn-based speech synthesis," in *Proceedings interspeech*, 2015.
[15] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E86-A, pp. 1956–1963, 2003.
[16] H. Zen, N. Braunschweiler, S. Buchholz, K. Knill, S. Krstulovic, and J. Latorre, "Speaker and language adaptive training for HMM-based polyglot speech synthesis," in *Interspeech 2010*, 2010, pp. 186–191.
[17] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Interspeech 2009*, 2009, pp. 528–531.
[18] L. Gao, H. Yu, Y. Li, and J. Liu, "A research on text analysis in tibetan speech synthesis," in *IEEE International Conference on Information and Automation (ICIA) 2010*, 2010, pp. 817–822.
[19] H. W. Yang, K. Oura, H.Y. Wang, and Z.Y. Gan, "Using speaker adaptive training to realize mandarin-tibetan cross-lingual speech synthesis," *Multimedia Tools and Applications*, vol. 74, no. 22, pp. 9927–9942, 2015.
[20] Zev Handel, "What is Sino-Tibetan? snapshot of a field and a language family in flux," *Language and Linguistics Compass*, vol. 2, no. 3, pp. 422–441, 2008.
[21] S. Yang, Z. Wu, and L. Xie, "On the training of dnn-based average voice model for speech synthesis," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.
[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
[23] L. Deng and D. Yu, "Deep learning: methods and applications," *Foundations and Trends in Signal Processing*, vol. 7, no. 3, pp. 197–387, 2014.
[24] W. Mirjam, "The EMIME bilingual database," *Technical Report EDI-INF-RR-1388, The University of Edinburgh*, 2010.
[25] P. C. Loizou, *Speech quality assessment*, chapter Multimedia analysis, processing and communications, pp. 623–654, Springer Berlin Heidelberg, 2011.