Realizing Tibetan Speech Synthesis by Speaker Adaptive Training

Hong-wu YANG* , Keiichiro OURA† , Zhen-ye GAN* and Keiichi TOKUDA†
*College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China
E-mail: yanghw@nwnu.edu.cn Tel: +86-931-7971503
†Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan.
E-mail: tokuda@nitech.ac.jp Tel: +81-52-735-5479

***Abstract*—This paper presents a method to realize HMM-based Tibetan speech synthesis using a Mandarin speech synthesis framework. A Mandarin context-dependent label format is adopted to label Tibetan sentences. A Mandarin question set is also extended for Tibetan by adding language-specific questions. A Mandarin speech synthesis framework is utilized to train an average mixed-lingual model from a large Mandarin multi-speaker-based corpus and a small Tibetan one-speaker-based corpus using the speaker adaptive training. Then the speaker adaptation transformation is applied to the average mixed-lingual model to obtain a speaker adapted Tibetan model. Experimental results show that this method outperforms the method using speaker dependent Tibetan model when only a small amount of training Tibetan utterances are available. When the number of training Tibetan utterances is increased, the performances of the two methods tend to be the same.**

I. Introduction

Multi-lingual speech synthesis has been a hot topic of research in recent years [1]. Since multi-lingual speech synthesis can synthesize speech of different languages with same or different speaker's voice, it has been widely used in multi-lingual spoken dialogue systems especially in the areas where many languages are spoken. The hidden Markov model-based (HMM-based) speech synthesis [2], which can easily synthesizes voice of different speakers by speaker adaptation transformation [3], has been a main technology for realizing multi-lingual speech synthesis system. The HMM-based multi-lingual speech synthesis uses mixed language methods [4], phoneme mapping methods [5] or state mapping methods [6][7] to achieve multi-lingual speech synthesis. To improve the quality of synthesized speech, the language dependent questions [8] are designed for model clustering. The KL distance is also employed [9][10] to measure the difference between the states of different languages. To overcome degradation of voice quality caused by different language resources, a set of language independent models are proposed to synthesize speech of new language by language adaptation transformation [11]. There is still a room for synthesizing speech for languages lacking of speech resources.

The development of speech synthesis technology is closely related to languages. Mandarin and Tibetan are the official languages in Tibetan region of China. While state-of-the-art researches are focusing on speech synthesis for major languages [2]-[12], which have fully developed speech synthesis frameworks and use plenty of data resources for model training, there is still very few studies on Tibetan speech synthesis [13] due to scarce speech resources of Tibetan. In HMM-based speech synthesis, we found that contexts can be shared for a new language if the new language is comparable with a major language. Since Mandarin and Tibetan belong to the Sino-Tibetan family [14]-[15], Tibetan is close to Mandarin on linguists and phonetics. This enables us to focus on the realization of Tibetan speech synthesis by borrowing the speech synthesis framework and speech data of Mandarin, which takes advantage of small training Tibetan data and consistence of HMM-based Mandarin speech synthesis.

In this paper, we use a small Tibetan training corpus to realize the Tibetan speech synthesis with a Mandarin speech synthesis framework. A full context-dependent label format designed for Mandarin is adopted to label the Tibetan sentences. The initial and the final are used as the synthesis units for both Mandarin and Tibetan. We also extend a set of Mandarin questions by adding Mandarin-specific and Tibetan-specific questions. A Mandarin speech synthesis framework is employed to train an average mixed-lingual model by using the speaker adaptive training with a large Mandarin multi-speaker-based corpus and a small Tibetan one-speaker-based corpus. The Tibetan speech is then synthesized from a speaker adapted Tibetan model which is transformed from the average mixed-lingual model by the speaker adaptation transformation. Therefore, by using small training speech data and a major language's speech of synthesis framework, the proposed method can be used to realize a speech synthesis system for a new language which has scarce speech resources and is similar to the major language.

In following sections, we will introduce our framework of Tibetan speech synthesis in section II. The full context-dependent label format is explained in section III. Experiments are conducted in section IV to show the results of our approach. We will bring our conclusion in section V.

II. Mixed lingual framework

Our framework of the Tibetan speech synthesis is shown in Fig. 1. We firstly used a large Mandarin multi-speaker-based speech corpus and a small Tibetan one-speaker-based speech corpus to train an average mixed-lingual voice model using the speaker adaptive training. The Tibetan speech corpus is

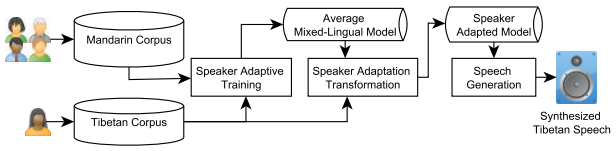

Fig. 1. Framework of Tibetan Speech Synthesis.

then used to perform the speaker adaptation transformation to obtain a speaker adapted Tibetan model for synthesizing the Tibetan speech.

We adopt the speaker adaptive training (SAT) [16] to train the average mixed-lingual model. The SAT normalize the difference of speakers among the training speakers with a linear regression function of state output distributions as shown in Eq. 1,

$$\hat{\mathbf{o}}_i^s(t) = \mathbf{A}^s\mathbf{o}(t) + \mathbf{b}^s = \mathbf{W}_i^s\xi_i(t), \tag{1}$$

where, s is the index of speakers $1 \cdots S$, t is the index of frame $1 \cdots T$. $\mathbf{W} = [\mathbf{A} \quad \mathbf{b}]$ is the transformation matrices of the speaker s. $\mathbf{o}(t)$ is the average observation vector of frame t. $\hat{\mathbf{o}}(t)$ is the speaker s's observation vector of frame t. $\xi_i(t) = [\mathbf{o}(t) \quad 1]^T$.

The average mixed-lingual model is trained from the Mandarin multi-speaker-based corpus and Tibetan one-speaker-based corpus. In particular, we use the constrained maximum likelihood linear regression (CMLLR) [17] to train the average mixed-lingual model on the context-dependent multi-space distribution hidden semi-Markov models (MSD-HSMMs).

After the speaker adaptive training, we apply the HSMM-based CMLLR adaptation [16] to the Tibetan training speech data so that the speaker dependent Tibetan models are trained from the average mixed-lingual model. We also adopt the maximum a-posteriori (MAP) algorithm [18] to further modify and upgrade the speaker adapted Tibetan models. The HSMM-based CMLLR adaptation can estimate the state output and duration distribution simultaneously by a linear transformation as shown in Eq. 2,

$$\begin{aligned} b_i(\mathbf{o}) &= \mathcal{N}(\mathbf{o}; \mathbf{A}\mu_i - \mathbf{b}, \mathbf{A}\mathbf{\Sigma}_i\mathbf{A}^T) \\ &= |\mathbf{A}^{-1}|\mathcal{N}(\mathbf{W}\xi; \mu_i, \mathbf{\Sigma}_i), \end{aligned} \tag{2}$$

where,$\mathbf{W} = [\mathbf{A}^{-1} \quad \mathbf{b}^{-1}]$ is the transformation matrices of the target Tibetan speaker. $\xi_i(t) = [\mathbf{o}^T \quad 1]^T$ is the extended vector of observations, μ_i is the mean of observations, and $\mathbf{\Sigma}_i$ is the covariance of observations.

III. Mixed Lingual Full Context-dependent Labels

Since Mandarin and Tibetan Lhasa dialect are syllabically paced tonal languages [14], each character can be regarded as a syllable which is a composition of an initial followed by a final. Each syllable carries its own tone to distinguish lexical or grammatical meaning. Tones are distinguished by the shape and the range of pitch contour of syllables. Mandarin uses Pinyin to reflect pronunciation of Chinese while Tibetan Lhasa dialect uses Tibetan Pinyin to reflect pronunciation of Tibetan. Each Pinyin system includes an initial set, a final set and a tone set. Mandarin has 22 initials and 39 finals while Tibetan Lhasa Dialect has 36 initials and 45 finals. Two languages can share 20 initials and 13 finals. Mandarin has four tones and one light tone while Tibetan Lhasa dialect has 4 tones but the tone values (tone value reflects the shape and range of a pitch contour) are different from Mandarin. Two languages have same part-of-speech and prosodic structure.

We adopt a full context-dependent label format of Mandarin to label Mandarin sentences and Tibetan sentences. A set of Speech Assessment Methods Phonetic Alphabet (SAMPA) is designed for labeling the initial and the final of Mandarin and Tibetan. The shared initials or finals by tow languages are labeled with same SAMPAs. All initials and finals of Mandarin and Tibetan, including silence and pause, are used as the synthesis unit of the context-dependent MSD-HSMMs. A six level context-dependent label format is designed by taking into account the following contextual features.

- **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.

We extend a question set designed for the HMM-based Mandarin speech synthesis by adding the language-specific questions. The Tibetan-specific units and Mandarin-specifics units are asked in the question set. We also design the questions to reflect the special pronunciation of Tibetan. Finally we get more than 3000 questions. These questions cover all features of the full context-dependent labels.

IV. Experiments

A. Experimental conditions

In our work, we use the EMIME Mandarin bilingual speech database [19] and a female Tibetan speech database as the training data. The EMIME Mandarin bilingual speech database is a Mandarin-English bilingual database aiming for personalized speech-to-speech translation. The database has 7 male Mandarin speakers and 7 female Mandarin speakers. Each speaker records 169 Mandarin sentences. The sentences are translated from a set of English sentences which include 25

European sentences, 100 news sentences and 20 semantically unpredictable sentences. We select all 7 female speaker's recordings as the Mandarin training data. A native female Tibetan Lhasa dialect speaker is invited to record the Tibetan speech database in a studio. 800 Tibetan sentences are selected from recent year's Tibetan newspapers. All recordings are saved in the Microsoft Windows WAV format as sound files (mono-channel, signed 16 bit, sampled at 16 kHz). We use 5-state left-to-right context-dependent multi-stream MSD-HSMMs. The TTS feature vectors are comprised of 138-dimensions: 39-dimension STRAIGHT [20] mel-Cepstral coefficients, log F0, 5 band-filtered aperiodicity measures, and their delta and delta delta coefficients.

We randomly select 100 sentences from 800 Tibetan sentences as the testing sentences. 10, 100 and 700 Tibetan utterances are randomly selected respectively from the left 700 Tibetan recordings to set up 3 Tibetan training sets. The initial/final coverage for different number of Tibetan sentences selected is 69.4 %, 91.7% and 100%, respectively. These Tibetan training sets and all 7 female Mandarin recordings are used to train the average mixed-lingual model. The Tibetan training sets are used in the speaker adaptation transformation.

B. Experimental results

To evaluate the synthesized Tibetan speeches, we trained 3 sets of different MSD-HSMMs as shown in blow. Each set of models synthesizes 100 testing sentences, from which we randomly select 20 utterances be the testing set of evaluation.

- SD model: Speaker dependent Tibetan model trained directly from {10,100 or 700} of Tibetan training utterances respectively.
- SI model: Speaker independent model trained only from 169*7=1183 Mandarin utterances.
- SAT model: Speaker adapted Tibetan model transformed from the average mixed-lingual model by using {10,100 or 700} Tibetan training utterances respectively. The average mixed-lingual model is trained from 1183 Mandarin utterances and {10,100 or 700} Tibetan training utterances respectively.

1) Speech quality: We invite 8 native speakers of Tibetan to be our subjects in a listening evaluation. We adopt mean opinion score (MOS) test to evaluate the naturalness of synthesized speech. We randomly play the testing set of all models except the SI model to the subjects. There are (20 utterances)*(3 Tibetan training sets)*(2 models)=120 testing speech files in total. The subjects are asked to carefully listen to these 120 utterances and score the naturalness of every utterance by a 5-point score. We also ask subjects the intelligibility they impressed after the test.

Fig. 2 shows the average scores and their 95% confidence intervals, in which the SAT model is compared with the SD model on different training sets. From the results we can see that the SAT model outperform the SD model on 10 and 100 utterances of training sets. For 10 training Tibetan utterances, the SD model synthesized speech get the lowest score of 1.31 while the SAT model get 1.99 of score. Meanwhile, the subjects feel that the SD model synthesized utterances are unintelligible but the SAT model synthesized utterances are understandable. When the number of training Tibetan utterances are increased to 100, the score and intelligibility of both models are improved. The SAT model still are obviously better than the SD model. The score of two models are basically the same when the training utterances are increased to 700. In this case, all subjects feel that they can easily understand all synthesized utterances. Therefore, the voice quality of the SAT model synthesized speech is significantly superior to those of the SD model synthesized speech in the case of the small amount of Tibetan training utterances. When the Tibetan training utterances are increased, the voice quality of different model synthesized speech will tend to be the same.

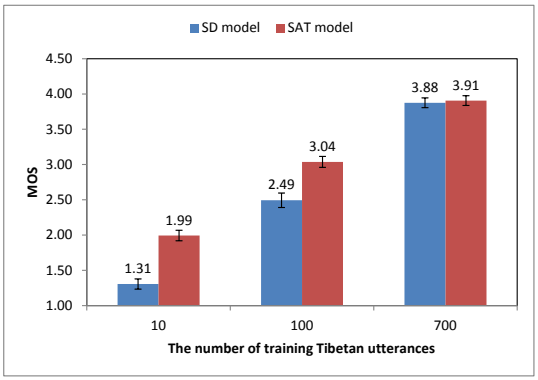

Fig. 2. MOS evaluation of synthesized speech by using different training Tibetan utterances.

2) Speaker similarity: We also perform a degradation mean opinion score (DMOS) test for the Speaker similarity evaluation. In the DMOS test all testing utterances and their original recordings are used. There are (20 utterances)* {(3 Tibetan training sets)*(2 models)+(1 SI model)}=140 synthesized speech files in total. Each synthesized utterance and its corresponding original recording form a pair of speech files. We randomly play each pair of speech files to the subjects with the order of original speech after synthesized speech. The subjects are asked to carefully compare these two files and evaluate the degree of similarity of synthesized speech to original speech. The 5-point score is used in which 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 shows the average score and their 95% confidence intervals in which we compare the SAT model with the SI model and the SD model. The results are interesting that the 2.41 of score of the SI model is better than those of the 10 Tibetan utterances trained SD model, and is close to those of the 10 Tibetan utterances trained SAT model. We also ask the subjects the impression on the SI model synthesized

speech. The subjects feel that these utterances are similar to the Tibetan voice uttered by foreigners. This is due to Mandarin and Tibetan not only share 33 synthesis units but also have the same syllabic structure and prosodic structure. Therefore, we can synthesize Tibetan-like voice by only using Mandarin model. When we mix in more Tibetan training utterances, the SAT model synthesized speech is more close to Tibetan than the SD model synthesized speech. When the training Tibetan utterances are increased to 700, the score of the SD model is close to the score of the SAT model. This again indicates that our method is better than the SD model based method when the amount of training Tibetan utterances is small.

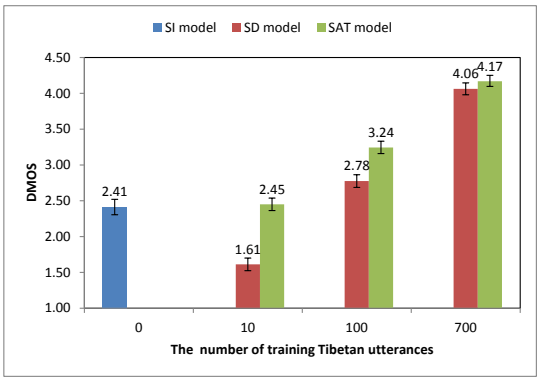

Fig. 3. DMOS evaluation of synthesized speech by using different training Tibetan utterances. The SI model for Tibetan is trained by using only Mandarin utterances, which can synthesize Tibetan speech with 2.41 of score.

V. Conclusions

In the paper, we presented a method for synthesizing Tibetan speech by using a HMM-based Mandarin speech synthesis framework. A Mandarin context-dependent label format was adopted to label Tibetan sentences. We also added language-specific questions into a Mandarin question set. The speaker adaptive training was used to train an average mixed-lingual model by mixing in a large Mandarin multi-speaker-based corpus and a small Tibetan one-speaker-based corpus. The speaker adapted Tibetan model was transformed from the average mixed-lingual voice model by using the speaker adaptation transform. Experimental results demonstrated that our method outperforms the SD model based method in the case of the small amount of training Tibetan utterances. Therefore, proposed method can be applied to realize the speech synthesis system for languages of scarce speech resources by using a speech synthesis framework of similar major language. Future work will attempt to improve the synthesized speech quality of our method by using a small deliberately designed Tibetan multi-speaker-based speech database.

Acknowledgements

The research leading to these results was partly funded by the National Natural Science Foundation of China (Grant No. 61263036, 61262055) , Gansu Science Fund for Distinguished Young Scholars (Grant No. 1210RJDA007) and the Core Research for Evolutional Science and Technology (CREST) from Japan Science and Technology Agency (JST).

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

[3] 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.

[4] J. Latorre, K. Iwano, and S. Furui, "New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer," *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, 2006.

[5] Y. J. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis," in *ISCSLP 2008*, 2008, pp. 9–12.

[6] 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.

[7] Y. Qian, H. Liang, and F. K. Soong, "A cross-language state sharing and mapping approach to bilingual (Mandarin/English) TTS," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1231–1239, 2009.

[8] H. Liang, Y. Qian, F. K. Soong, and G. Liu, "A cross-language state mapping approach to bilingual (Mandarin-English) TTS," in *ICASSP 2008*, 2008, pp. 4641–4644.

[9] X. L. Peng, K. Oura, Y. Nankaku, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices," in *IEEE 10th International Conference on Signal Processing*, 2010, pp. 605–608.

[10] Y. N. Chen, Y. Jiao, Y. Qian, and F. K. Soong, "State mapping for cross-language speaker adaptation in TTS," in *ICSP 2010*, 2009, pp. 4273–4276.

[11] 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.

[12] Y. Qian, F. K. Soong, Y. Chen, and M. Chu, "An HMM-based Mandarin Chinese Text-To-Speech system," in *ISCSLP 2006*, pp. 223–232. 2006.

[13] 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.

[14] 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.

[15] M.C. Goldstein, Gelek Rimpoche, and L. Phuntshog, "Essentials of modern literary Tibetan," *University of California Press*, 1991.

[16] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[17] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[18] O. Siohan, T. A. Myrvoll, and C. H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech & Language*, vol. 16, no. 1, pp. 5–24, 2002.

[19] W. Mirjam, "The EMIME bilingual database," *Technical Report EDI-INF-RR-1388, The University of Edinburgh*, 2010.

[20] H. Kawahara, I. Masuda-Katsuse, and A. Cheveign de, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.