Manuscript
Click here to download Manuscript: man-tib.tex
Click here to view linked References

Noname manuscript No.
(will be inserted by the editor)

Using Speaker Adaptive Training to Realize Mandarin-Tibetan Cross-Lingual Speech Synthesis

Hongwu YANG · Keiichiro OURA · Haiyan WANG · Zhenye GAN · Keiichi TOKUDA

Received: date / Accepted: date

Abstract This paper presents a method to realize the hidden Markov model (HMM)-based Mandarin-Tibetan cross-lingual statistical speech synthesis using speaker adaptive training. A set of Speech Assessment Methods Phonetic Alphabet (SAMPA) is designed to label the pronunciation of the initial and the final of Mandarin and Tibetan syllables according to the similarities in pronun-

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).

Hongwu YANG
Key Laboratory of Atomic and Molecular Physics and Functional Materials of Gansu Province, College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China.
Tel.: +86-931-7971503
Fax: +86-931-7971503
E-mail: yanghw@nwnu.edu.cn

Keiichiro OURA
Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan.
E-mail: uratec@sp.nitech.ac.jp Tel.: +81-52-735-5479

Haiyan WANG
College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China.
E-mail: why6715@163.com Tel.: +81-52-735-5479

Zhenye GAN
Key Laboratory of Atomic and Molecular Physics and Functional Materials of Gansu Province, College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China.
E-mail: ganzy@nwnu.edu.cn Tel.: +86-931-7971503

Keiichi TOKUDA
Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan.
E-mail: tokuda@nitech.ac.jp Tel.: +81-52-735-5479

ciation between Mandarin and Tibetan. A grapheme-to-phoneme conversion method is realized to convert Chinese or Tibetan sentences to SAMPA-based Pinyin sequences. A Mandarin statistical speech synthesis framework is employed to realize Mandarin-Tibetan cross-lingual speech synthesis. A set of context-dependent label format is designed to label the context information of Mandarin and Tibetan sentences. A question set is also realized for context dependent decision tree clustering. The initial and the final are used as the synthesis units with training using a set of average mixed-lingual models from a large Mandarin multi-speaker-based corpus and a small Tibetan one-speaker-based corpus using speaker adaptive training (SAT). Then, the speaker adaptation transformation is applied to the speaker dependent (SD) training data to obtain a set of speaker dependent Mandarin or Tibetan models from the average mixed-lingual models. The Mandarin speech or Tibetan speech is then synthesized from the speaker dependent Mandarin or Tibetan models. Tests show that this method outperforms the method using only Tibetan SD models when only a small number of Tibetan training utterances are available. When the number of training Tibetan utterances is increased, the performances of the two methods tend to be the same. Mixed Tibetan training sentences have a small effect on the quality of synthesized Mandarin speech.

Keywords HMM-based speech synthesis · speaker adaptive training · multi-lingual speech synthesis · Tibetan speech synthesis · Mandarin-Tibetan cross-lingual speech synthesis · grapheme-to-phoneme conversion

1 Introduction

Multi-lingual speech synthesis has been a hot topic of research in recent years [1]. Since multi-lingual speech synthesis can synthesize speech in 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 with 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 cross-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 a new language by language adaptation transformation [11]. There is still 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 the 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, 14] 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 language family [15, 16], Tibetan is close to Mandarin on linguists and phonetics. This enables us to focus on the realization of Mandarin-Tibetan cross-lingual speech synthesis by borrowing the speech synthesis framework and speech data of Mandarin, which takes advantage of small training Tibetan data and consistency of HMM-based Mandarin speech synthesis.

In this paper, we design a set of Speech Assessment Methods Phonetic Alphabet (SAMPA) to label the pronunciation for both Mandarin and Tibetan. A grapheme-to-phoneme procedure is used to obtain the SAMPA represented Pinyin sequences from Chinese sentences or Tibetan sentences. Then we modify a Mandarin statistical speech synthesis framework to realize Mandarin-Tibetan cross-lingual speech synthesis. A full context-dependent label format is designed to label the context information of Mandarin or Tibetan. The initial and the final form the synthesis units for both Mandarin and Tibetan. We also extend a set of Mandarin questions by adding Mandarin-specific and Tibetan-specific questions to perform the context dependent clustering of HMM states. An average mixed-lingual model is trained using the speaker adaptive training with a large Mandarin multi-speaker-based corpus and a small Tibetan one-speaker-based corpus. The Mandarin or Tibetan speech is then synthesized from a speaker adapted Mandarin model or 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 synthesis framework, the proposed method can be used to realize a cross-lingual speech synthesis system that can synthesis a new language which has scarcely speech resources and is similar to the majority language.

In following sections, we will introduce a SAMPA based grapheme-to-phoneme conversion in section 2. Our framework of Mandarin-Tibetan cross-lingual speech synthesis will be introduced in section 3. The full context-dependent label format is explained in section 4. Experiments are conducted in section 5 to show the results of our approach. We will bring our conclusion in section 6.

2 A SAMPA based Grapheme-to-phoneme conversion of Mandarin and Tibetan

Mandarin TTS systems almost use Pinyin system to label the pronunciation of Chinese sentences while Tibetan uses Tibetan Pinyin system to reflect the pronunciation of Tibetan sentences. Because these two Pinyin systems are

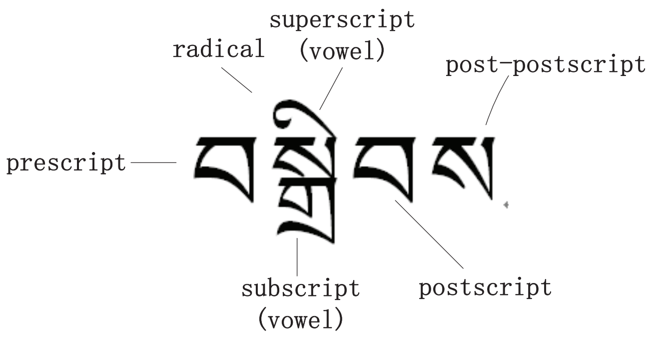


Fig. 1 A 2-dimensional structure of the Tibetan word.

incompatible, we cannot directly use them to realize Mandarin-Tibetan cross-lingual speech synthesis. To solve this problem, we proposed a set of Speech Assessment Methods Phonetic Alphabet (SAMPA) [17] as these two language's Pinyin system, and use the SAMPA-based Pinyin to label the pronunciation for both Mandarin and Tibetan.

2.1 Structure of Tibetan word

Tibetan is spoken primarily by Tibetan peoples who live across a wide area of eastern Central Asia especially in the Tibetan district of China as well as some parts of Nepal, India and other countries. Tibetan belongs to Burmese Tibetan branch of Sino-Tibetan family and set up in the early 7th century. It is a type of alphabetic writing developed on the basis of the Sanskrit. Tibetan has 3 different kinds of dialects, that are Tsang, Kang and Ando. While these dialects get a big difference on pronunciation, the scripts of Tibetan dialects are almost same. Because Lhasa dialect, which belongs to Tsang, is the most commonly used Tibetan dialect, it becomes the official dialect of Tibet.

A Tibetan word has a 2-dimensional structure as showed in figure 1. The Tibetan alphabet has 30 consonants, called radicals, which form the basis of the script. Each consonant letter can be regarded as a single syllable with an inherent vowel /a/. The consonants can be written either as radicals or in the form of superscripts and subscripts. The superscript position above a radical is reserved for the consonants /r/, /l/, and /s/, while the subscript position under a radical is for the consonants /y/, /r/, /l/, and /w/. Some consonants can also be placed in prescript, postscript, or post-postscript positions. In the Tibetan script, the syllables are written from left to right and are separated by a tseg (.). Because Tibetan words are monosyllabic, the mark tseg used as a space to divide words.

2.2 Pronunciation differences between Tibetan and Mandarin

Tibetan words are quite different from Chinese characters. However, since Mandarin and Tibetan are all belong to the Sino-Tibetan language family, these two languages have many similarities on linguistics and phonetics. Mandarin and Tibetan Lhasa dialect are syllabically paced tonal languages [15]. Each script can be regarded as a syllable which is a composition of an initial followed by a final. Each syllable carries its own tone to differentiate lexical or grammatical meaning. Tones are distinguished by the shape and the range of the pitch contour of syllables. Tibetan and Chinese have same part-of-speech and prosodic structure. 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.

In terms of pronunciation, the initials of Tibetan generally include single consonants and compound consonants. The initials of Lhasa dialect mainly refer to 28 single consonants, which including 8 plosives, 6 affricates, 6 fricatives, 1 approximant, 4 nasals, 1 lateral and 2 dullnesses. The Tibetan Lhasa dialect also has 6 unique compound consonants: /mb/, /nz/, /nd/, / nzh/, /nj/ and /ngy/. All 21 Mandarin initials are single consonants besides of a null initial. Mandarin has 2 unique initials, one is fricative /f/. Another is fricative /h/. Tibetan has 8 unique initials. Tibetan Lhasa dialect has 46 finals, which include 15 monophthongs, 2 diphthongs and 9 nasal vowels, as well as 20 finals that combined by basic vowels followed by /m/, /b/, /g/ or /r/. Mandarin has 39 finals that include 10 single finals, 13 compound finals and 16 nasal finals.

Like Mandarin, Tibetan Lhasa dialect is tonal language too. But there are no dedicated symbols for tone. However, since tones developed from segmental features they can be correctly predicted from Tibetan words. 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 word's pitch contour) are different from Mandarin. While tone values of four Mandarin tones are 55, 35, 214 respective, tone value of four Tibetan Lhasa dialect tones is 54, 55, 12 and 14, respectively.

2.3 SAMPA design for Mandarin and Tibetan

Since Tibetan and Mandarin have many similarities in pronunciation, we design a set of SAMPA by using a Chinese SAMPA set named SAMPA-SC [18] to label the pronunciation of initials and finals for both Mandarin and Tibetan. We discovered that parts of the International Phonetic Alphabet (IPA) of Mandarin are consistent with IPA of Tibetan. We take IPA as reference to design SAMPA for Mandarin and Tibetan. If IPAs of Tibetan and Mandarin are same we directly use SAMPA-SC to label both Mandarin and Tibetan; otherwise, we define new SAMPAs to label Tibetan. The design process s illustrated in figure 2. Tibetan IPAs are obtained from the Tibetan alphabet. Then the IPAs are compared with Chinese IPAs. For those Tibetan IPAs con-

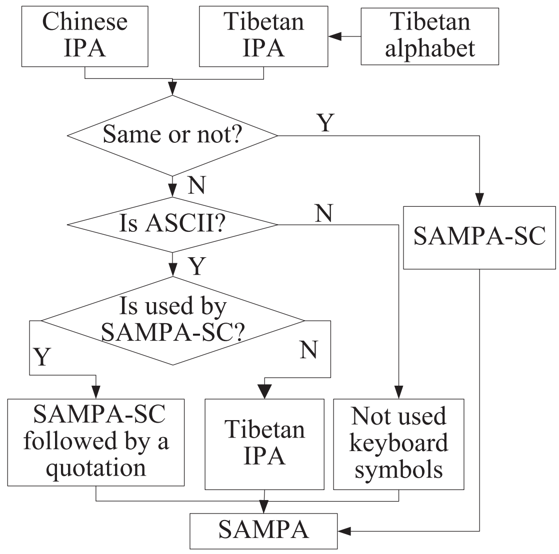


Fig. 2 Design procedure of SAMPA-T.

sistent with Chinese IPAs, we simply use SAMPA-SC to label both Mandarin and Tibetan. For those Tibetan IPAs that are different with Chinese IPAs, we use SAMPA-SC to label Mandarin and design new SAMPAs that can easily input from keyboard to label Tibetan.

2.3.1 Consonants

Since 18 Tibetan consonants have the same IPA with Mandarin, we directly use these IPA's SAMPA-SC to label these consonants. The IPA of the other 12 Tibetan consonants are inconsistent with Mandarin. We adopt the following rules to design their SAMPAs.

1. If the IPAs of Tibetan consonants is consist of ASCII characters, and not be used by Mandarin, we directly use the IPAs as the SAMPAs of these Tibetan consonants.
2. If the ASCII character based IPA of Tibetan consonant has been used by Mandarin, we add a single quotation symbol after the IPA to be the SAMPA of Tibetan consonant.
3. For the rest of IPAs of Tibetan consonants that difficult to input from keyboard, we use never used ASCII characters similar with the IPA be the Tibetan SAMPAs.

2.3.2 vowels

Tibetan has 4 vowel symbols that can change the pronunciation of consonants. Because vowel symbols cannot be divided into single syllables, they must be compounded with consonants. So we customarily call them vowel symbols instead of vowels. Name of vowel symbols does not match its pronunciation. Thus, the vowels in Tibetan mainly act as the finals of syllables. /i/, /e/ and /u/ directly appeared as superscripts, and /o/ appeared as subscript. Therefore, vowels have different pronunciation in different condition. When we determine the pronunciation of a Tibetan word, we should firstly lookup the postscript, then obtain the pronunciation according to the effect of vowels.

For 4 Tibetan vowels, since /i/, /o/ and /u/ share the same IPA with Mandarin, we can use SAMPA-SC to label them. While IPA of /e/ is not coincident with Mandarin IPA, we simply use /e/ as its SAMPA.

2.3.3 tones

Tone is important for the pronunciation of Mandarin and modern Lhasa Tibetan. Tone gets the function of distinct meaning of words and grammar. Mandarin has 4 tones and a light tone. The number of tone is different in different Tibetan dialects. Tibetan Lhasa dialect also has 4 tones. We use tone value as the tone of SAMPA.

2.4 SAMPA based grapheme-to-phoneme conversion of Tibetan

We have developed a Mandarin TTS system, which uses a text analyzer to convert Chinese sentences to Mandarin Pinyin based initial and final sequences. Therefore we can easily obtain Mandarin SAMPA from Mandarin Pinyin by a SAMPA-SC lookup table. However, there is lack of Tibetan text analyzer, so we developed a grapheme-to-phoneme conversion module for Tibetan to obtain the Tibetan SAMPA of the initial and the final from Tibetan sentences. Figure 3 shows the flowchart of Tibetan grapheme-to-phoneme conversion. The Tibetan sentence is firstly segmented to syllables. Then, the radical is located from syllable. Next, the initial and the final of the syllable are obtained by decomposing the radical with a radical decomposition table. Finally, the SAMPAsof the initial and the final are acquired by searching an initial SAMPA table and a final SAMPA table.

2.4.1 The initial and final separation of Tibetan mono-syllable

Like Mandarin, Tibetan syllable constitute with an initial followed by a final. When syllables are segmented from Tibetan sentence, we can obtain the radical, superscript, subscript, prescript, postscript and post-postscript of each syllable. The initial and the final can be obtained by combining different part of a syllable, and then the SAMPA of the initial and the final can be obtained by searching initial SAMPA table or final SAMPA table.The initial and the final of Tibetan can be obtained from syllable by the following rules:

$$initial = prescript + superscript + radical + subscript$$
$$final = vowel + postscript + (post - postscript)$$

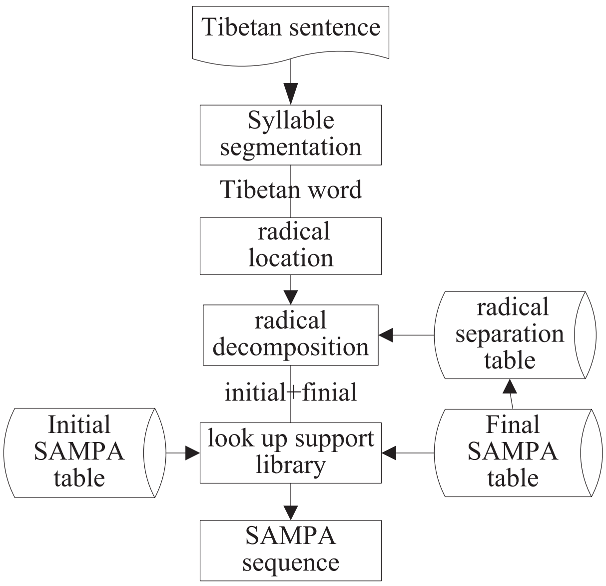


Fig. 3 grapheme-to-phoneme conversion of Tibetan.

According to word decomposition method [19], to separate the initial and the final, the radical should be located correctly. Tibetan has a strict and complete alphabetical combination rules. All 30 consonants can be radical but vowels only can be superscripts or subscripts. Therefore the radical can be located according to Tibetan grammar. The Tibetan alphabet that takes a vowel symbol, a superscript or a subscript can be regarded as a radical. According to statistics, 79.92% Tibetan words have a vowel symbol, and 93.16% Tibetan words have a superscript or a subscript [20]. This helps us to locate a radical according to vowel, superscript position or subscript position around the radical. Remain radicals can be determined with a lookup table.

2.4.2 SAMPA conversion from the initial and final

Modern Tibetan has 213 initials and 77 finials in total. We established Tibetan initial SAMPA dictionaries and Tibetan final SAMPA dictionaries for different dialects. After separating the initial and final from Tibetan mono-syllable, we obtain the SAMPA of the initial and the final by searching the initial dictionary or final dictionary. Since Tibetan Lhasa dialect has tone, we put the tone SAMPA after the final SAMPA when the initial and the final are converted to SAMPAs.

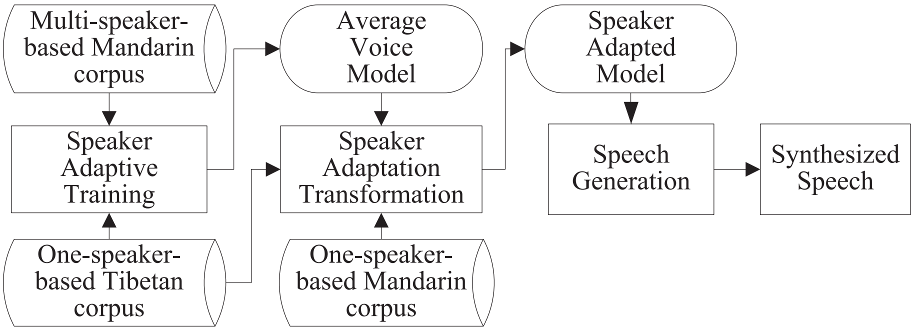


Fig. 4 Framework of Mandarin - Tibetan cross-lingual speech synthesis.

3 Framework of Mandarin-Tibetan cross-lingual speech synthesis

Our framework of the Mandarin-Tibetan cross-lingual speech synthesis is shown in Fig. 4. 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 Mandarin speech corpus or Tibetan speech corpus is then used to perform the speaker adaptation transformation to obtain a speaker adapted Mandarin model or Tibetan model for synthesizing the Mandarin Speech or Tibetan speech.

We adopt the speaker adaptive training (SAT) [21] to train the average mixed-lingual model. The SAT normalize the difference of speakers among the training speakers with linear regression functions of state outputs and duration distributions as shown in Eq. 1 and Eq. 2,

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

$$\hat{\mu}_i^s = \alpha^s \mu_i + \beta^s = \chi^s \phi_i, \tag{2}$$

where, s is the index of speakers $1 \cdots S$. $\mathbf{o}_i$ is the mean vector of state output. $\hat{\mathbf{o}}_i^s$ is the speaker s's mean vector of state output. $\xi_i = [\mathbf{o}_i \quad 1]^T$. $\mathbf{W}^s = [\mathbf{A}^s \quad \mathbf{b}^s]$ is the state output transformation matrices of the speaker s. $.\mu_i$ is the mean of duration distributions. $\hat{\mu}_i^s$ is the speaker s's duration distributions. $\phi_i = [\mu_i \quad 1]^T$. $\chi^s = [\alpha \quad \beta]$ is the duration distribution transformation matrices of the speaker s.

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) [**?**] 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 [21] to the Mandarin or Tibetan training speech data so that the speaker dependent Mandarin model or Tibetan model are trained from the average mixed-lingual model. The HSMM-based CMLLR adaptation can estimate the state output and duration distribution simultaneously by a linear transformation as shown in Eq. 3 and Eq. 4,

$$\begin{aligned} b_i(\mathbf{o}) &= \mathcal{N}(\mathbf{o}; \mathbf{A}\mu_i - \mathbf{b}, \mathbf{A}\mathbf{\Sigma}_i\mathbf{A}^T) \\ &= |\mathbf{A^{-1}}|\mathcal{N}(\mathbf{W}\xi; \mu_i, \mathbf{\Sigma}_i), \end{aligned} \tag{3}$$

$$\begin{aligned} p_i(d) &= \mathcal{N}(d; \alpha d_i - \beta, \alpha\sigma_i\alpha A^T) \\ &= |\chi^{-1}|\mathcal{N}(\chi\mathbf{\Phi}; \mu_i, \sigma_i^2), \end{aligned} \tag{4}$$

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.

The transformation matrices $\mathbf{W}$ can be estimated by maximizing the likelihood of adaptation data $\mathbf{O}$ as shown in Eq. 5

$$\hat{\mathbf{\Lambda}} = \arg\max_{\Lambda} P(\mathbf{O}|\lambda, \Lambda) \tag{5}$$

Furthermore, we adopt MAP (Maximum A Posteriori) modification algorithm [24] to further modify and upgrade the speaker adapted Mandarin model or Tibetan model. The state occupancy probability $\gamma_t^d(i)$ of being in the state i at the period of time from $t-d+1$ to t is defined as

$$\gamma_t^d(i) = \frac{1}{P(O|\lambda)} \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_{t-d}(j)p(d) \prod_{s=t-d+1}^{t} b_i(o_s)\beta_t(i) \tag{6}$$

The MAP adaptation of mean vectors of the Gaussian pdfs transformed by the CSMAPLR algorithm can be simply estimated as follows:

$$u_i = \frac{\omega\bar{u}_i + \sum_{t=1}^{T}\sum_{d=1}^{t}\kappa_t^d(i)\sum_{s=t-d+1}^{t} o_s}{\omega + \sum_{t=1}^{T}\sum_{d=1}^{t}\kappa_t^d(i)d} \tag{7}$$

4 Mixed lingual Full Context-dependent labels

We adopt a full context-dependent label format of Mandarin to label Mandarin sentences and Tibetan sentences. A set of SAMPA is designed for labeling the initial and the final of Mandarin and Tibetan. 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 levels 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. Tibetan-specific units and Mandarin-specific 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.

5 Experiments

5.1 Experimental conditions

In our work, we use the EMIME Mandarin bilingual speech database [22] and a female Tibetan speech database as the training data. The EMIME Mandarin bilingual speech database is a Mandarin-English bilingual database. The database has 7 male Mandarin speakers and 7 female Mandarin speakers. Each speaker records 169 Mandarin training sentences and 18 Mandarin testing 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 asked to record the Tibetan speech database in a studio. 800 Tibetan sentences are chosen from recent year's Tibetan newspapers. All recordings are saved in the Microsoft Windows WAV format as sound files (mono-channel, signed 16 bits, 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 [23] 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 Tibetan 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. These Tibetan training sets and all 7 female Mandarin recordings are used to train the average mixed-lingual model. First Mandarin female speaker's training sentences are employed in the speaker adaptation

Table 1 The coverage of Tibetan synthesis units for different Tibetan training sets

number of Tibetan sentences	coverage (%)
10	69.4
100	91.7
700	100

transformation to train speaker adapted Mandarin model, and all Tibetan training sets are used to train speaker adapted Tibetan model. The coverage of Tibetan synthesis units (initials and finals) for different number of Tibetan training sentences is given in table 1. We can see from table 1 that for 10 Tibetan training sentences only less than 70% synthesis units are included in training sentences. Uncovered Tibetan synthesis units will be synthesized with Mandarin units.

5.2 Experimental results

To evaluate the synthesized Mandarin speeches and Tibetan speeches, we trained 3 sets of different MSD-HSMMs as showed in below. Each set of models synthesizes 100 Tibetan testing sentences, from which we randomly select 20 Tibetan utterances be the Tibetan testing set of evaluation. We also synthesize 18 Mandarin female testing sentences using of SI model or SAT model be the Mandarin testing set.

- 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 or Mandarin model. The Tibetan SAT model is transformed from the average mixed-lingual model by using {10, 100 or 700} Tibetan training utterances respectively. The Mandarin SAT model is transformed from the average mixed-lingual model by utilizing first Mandarin female speaker's training sentences. The average mixed-lingual model is trained from 7 female Mandarin speaker's training utterances and {10, 100 or 700} Tibetan training utterances respectively.

5.2.1 Speech quality

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

For observing the effect of mixed Tibetan training utterances on Mandarin models, we also invite 8 native Mandarin speakers to be the Mandarin subjects for evaluating synthesized Mandarin utterances. The evaluation method is consistent with Tibetan evaluation. Each of the SI model and the SAT model synthesizes 54 Mandarin utterances respectively. These Mandarin utterances are synthesized using 18 Chinese testing sentences from 3 Tibetan training set trained models)

Figure 5 shows the average scores and their 95% confidence intervals, in which the Tibetan SAT model, Tibetan SD model and Mandarin SAT model are compared on different Tibetan training sets. From the results, we can see that the Tibetan SAT model outperform the Tibetan SD model with 10 and 100 utterances of training sets. For 10 training Tibetan utterances, the Tibetan SD model synthesized speech gets the lowest score of 1.31 while the Tibetan SAT model gets 1.99 score. Meanwhile, the subjects feel that the Tibetan SD model synthesized utterances are unintelligible but the Tibetan SAT model synthesized utterances are understandable. When the number of training Tibetan utterances is raised to 100, the score and intelligibility of both Tibetan models are improved. The Tibetan SAT model still is obviously better than the Tibetan SD model. Score of two Tibetan models is basically the same when the training utterances are brought to 700. In this case, all subjects think they can easily understand all synthesized utterances. Therefore, the voice quality of the Tibetan SAT model synthesized speech is significantly superior to those of the Tibetan SD model synthesized speech in the case of the small amount of Tibetan training utterances. When Tibetan training utterances are increased, the voice quality of different Tibetan model synthesized speech will tend to be the same.

For synthesized Mandarin utterances, we can clearly see from figure 5 that the MOS scores of Mandarin SAT model synthesized utterances are all great than 4.0 in each Tibetan training set. This means that mixing Tibetan training utterances into Mandarin training utterances has little impact on the results of Mandarin speech synthesis.

5.2.2 Speaker similarity

We also carry out 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 Tibetan 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 the 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 the 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. We also perform the DMOS evaluation on

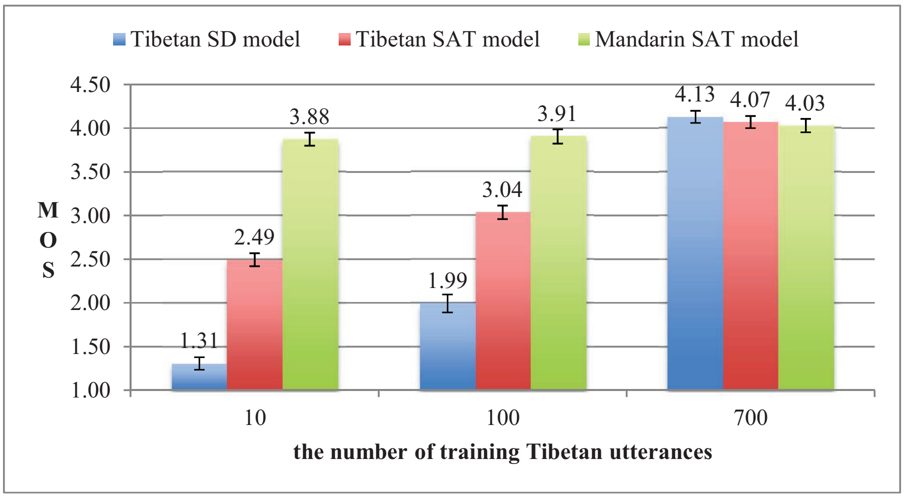


Fig. 5 MOS evaluation of synthesized speech by using different training Tibetan utterances.

the Mandarin SAT model synthesized 54 Mandarin utterances to test the effect of different number of Tibetan training sentences on Mandarin SAT model.

Figure 6 shows the average score and their 95% confidence intervals in which we compare the Tibetan utterances synthesized from the Tibetan SAT model, the SI model and the Tibetan SD model. In figure figure:03, the Tibetan SI model means the DMOS score on Tibetan utterances synthesized from the SI model, which is trained only using Mandarin training utterances. The Mandarin SI model means the DMOS results on the SI model synthesized Mandarin utterances. It is interesting that the 2.41 of score on the Tibetan SI model is better than those of the 10 Tibetan utterances trained Tibetan SD model, and is close to those of the 10 Tibetan utterances trained Tibetan SAT model. We also request the impression of subjects 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 using only Mandarin model. When we mix in more Tibetan training utterances, the Tibetan SAT model synthesized utterances are more close to Tibetan than the Tibetan SD model synthesized utterances. When training Tibetan utterances are increased to 700, the score of the Tibetan SD model is close to the score of the Tibetan SAT model. This again indicates that our method is preferable to the Tibetan SD model based method when the amount of training Tibetan utterances is small.

Figure 6 also compare the DMOS scores of synthesized Mandarin utterances. We can see from figure 6 that the score on the Mandarin SI model and 3 different Tibetan training sets trained Mandarin SAT model is all great than 4.0. The DMOS score has slight decrease when mixed in more Tibetan

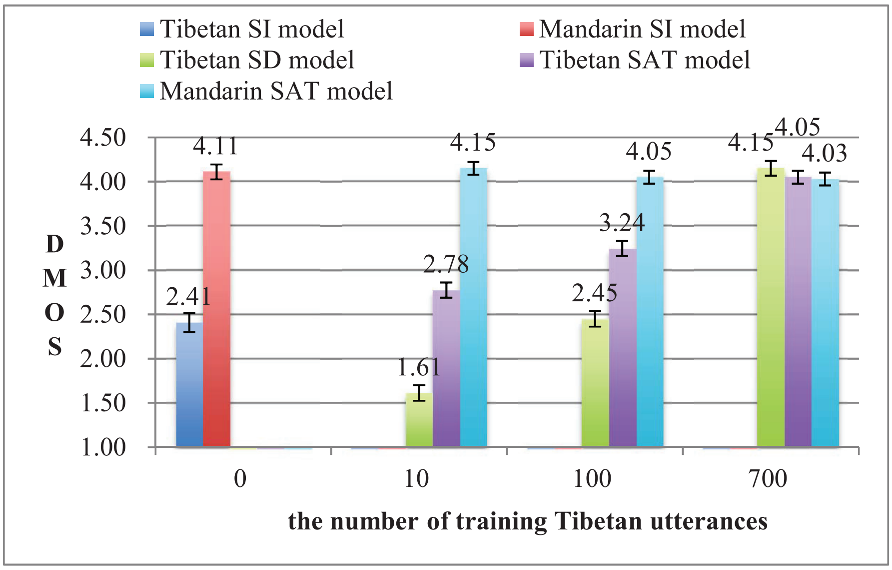


Fig. 6 DMOS evaluation of synthesized speech by using different training Tibetan utterances. The Tibetan SI model is trained by using only Mandarin utterances, which can synthesize Tibetan speech with 2.41 of score.

training utterances. This means mixed Tibetan training utterances have less effect on synthesized Mandarin speech.

6 Conclusions

In the paper, we presented a method for realizing Mandarin-Tibetan cross-lingual speech synthesis by using a HMM-based Mandarin speech synthesis framework. A Mandarin context-dependent label format was adopted in order to label both Mandarin and 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 or Mandarin model was transformed from the average mixed-lingual voice model by using the speaker adaptation transformation. The Mandarin or Tibetan speech is then synthesized from the Mandarin speaker adapted model or Tibetan adapted model. Experimental results demonstrated that our method outperforms the Tibetan 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.

References

1. Bourlard H, Dines J, Magimai-Doss M, Garner P, Imseng D, Motlicek P, Liang H, Saheer L, Valente Fm (2011) Current trends in multilingual speech processing. Sadhana 36: 885–915.
2. Zen H, Tokuda K, Black AW (2009) Statistical parametric speech synthesis. Speech Communication 51(11):1039–1064
3. Yamagishi J, Tamura M, Masuko T, Tokuda K, Kobayashi T (2003) A training method of average voice model for HMM-based speech synthesis. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences E86-A:1956–1963
4. Latorre J, Iwano K, Furui S (2006) New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer. Speech Communication 48(10):1227–1242
5. Wu YJ, King S, Tokuda K (2008) Cross-lingual speaker adaptation for HMM-based speech synthesis. In: ISCSLP 2008, pp 9–12
6. Wu YJ, Nankaku Y, Tokuda K (2009) State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis. In: Interspeech 2009, pp 528–531
7. Qian Y, Liang H, Soong FK (2009) A cross-language state sharing and mapping approach to bilingual (Mandarin/English) TTS. IEEE Transactions on Audio, Speech, and Language Processing 17(6):1231–1239
8. Liang H, Qian Y, Soong FK, Liu G (2008) A cross-language state mapping approach to bilingual (Mandarin-English) TTS. In: ICASSP 2008, pp 4641–4644
9. Peng XL, Oura K, Nankaku Y, Tokuda K (2010) Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices. In: IEEE 10th International Conference on Signal Processing, pp 605–608
10. Chen YN, Jiao Y, Qian Y, Soong FK (2009) State mapping for cross-language speaker adaptation in TTS. In: ICSP 2010, pp 4273–4276
11. Zen H, Braunschweiler N, Buchholz S, Knill K, Krstulovic S, Latorre J (2010) Speaker and language adaptive training for HMM-based polyglot speech synthesis. In: Interspeech 2010, pp 186–191
12. Qian Y, Soong FK, Chen Y, Chu M (2006) An HMM-based Mandarin Chinese Text-To-Speech system. In: ISCSLP 2006, pp 223–232
13. Schrder M, Hunecke A (2007) MARY TTS participation in the Blizzard Challenge 2007. In: Blizzard Challenge 2007, Bonn, Germany.
14. Gao L, Yu H, Li Y, Liu J (2010) A research on text analysis in tibetan speech synthesis. In: IEEE International Conference on Information and Automation (ICIA) 2010, pp 817–822
15. Handel Z (2008) What is Sino-Tibetan? snapshot of a field and a language family in flux. Language and Linguistics Compass 2(3):422–441
16. Goldstein M, Rimpoche G, Phuntshog L (1991) Essentials of modern literary Tibetan. University of California Press

17. Wells J (1997) SAMPA computer readable phonetic alphabet. Gibbon, D. and Moore, R. and Winski, R. Handbook of Standards and Resourcesfor Spoken Language Systems
18. Zhang J (2009) Machine readable phonetic sampa-sc of chinese mandarin. ACTA ACUSTICA 34(1):81–86
19. Li Y, Kong J, Yu H (2008) Conversion and realization of tibetan text to pronunciation automatic rules. Journal of Tsinghua university (natural science edition) 48(S1):621–626
20. Gao D, Gong Y (2005) A statistically study on the qualities of all modern tibetan character set. Journal of Chinese Information Processing 19(1):71–75
21. Yamagishi J, Kobayashi T, Nakano Y, Ogata K, Isogai J (2009) Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm. IEEE Transactions on Audio, Speech, and Language Processing 17(1):66–83
22. Mirjam W (2010) The EMIME bilingual database. Technical Report EDI-INF-RR-1388, The University of Edinburgh
23. Kawahara H, Masuda-Katsuse I, Cheveign de A (1999) 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 27(3):187–207
24. Siohan O, Myrvoll TA, Lee CH (2002) Structural maximum a posteriori linear regression for fast HMM adaptation. Computer Speech & Language 16(1):5–24

Author Biographies
Click here to download Author Biographies: Author biographies.docx

Hongwu Yang received the M.S. degree in physics from Northwest Normal University, Lanzhou, China, in 1995 and the Ph.D. degree in computer science and technology from Tsinghua University, Beijing, China, in 2007.
He is currently a professor with the College of Physics and Electronic Engineering, Northwest Normal University. His research interests include multi-lingual speech synthesis, expressive speech synthesis and recognition, audio content-based information retrieval, and multimedia processing.
He is a member of the Institute of Electronics, Information, and Communication Engineers and a member of the IEEE Signal Processing Society.

Keiichiro Oura received his Ph.D. degree in Computer Science and Engineering from Nagoya Institute of Technology, Nagoya, Japan, in 2010.
He was an intern/co-op researcher at ATR Spoken Language Translation Research Laboratories, Kyoto, Japan, from Sep. to Dec. 2007.
From Mar. 2008 to Nov. 2009, he was a postdoctoral fellow of the EMIME project at the Nagoya Institute of Technology.
From Dec. 2009 to Mar. 2012, he was a postdoctoral fellow of the SCOPE project at Nagoya Institute of Technology.
He is currently a postdoctoral fellow of the CREST project at the Nagoya Institute of Technology.
His research interests include statistical speech recognition and synthesis.
He received the ISCSLP Best Student Paper Award, the IPSJ YAMASHITA SIG Research Award, the ASJ ITAKURA Award, and IPSJ KIYASU Special Industrial Achievement Award, in 2008, 2010, 2013, and 2013, respectively.

Haiyan Wang received the B. E. degree in electronic science and technology from Northwest Normal University, Lanzhou, China, in 2012. She is now a master student in Northwest Normal University.
Her main research interests include speech synthesis and speech man-machine interaction.

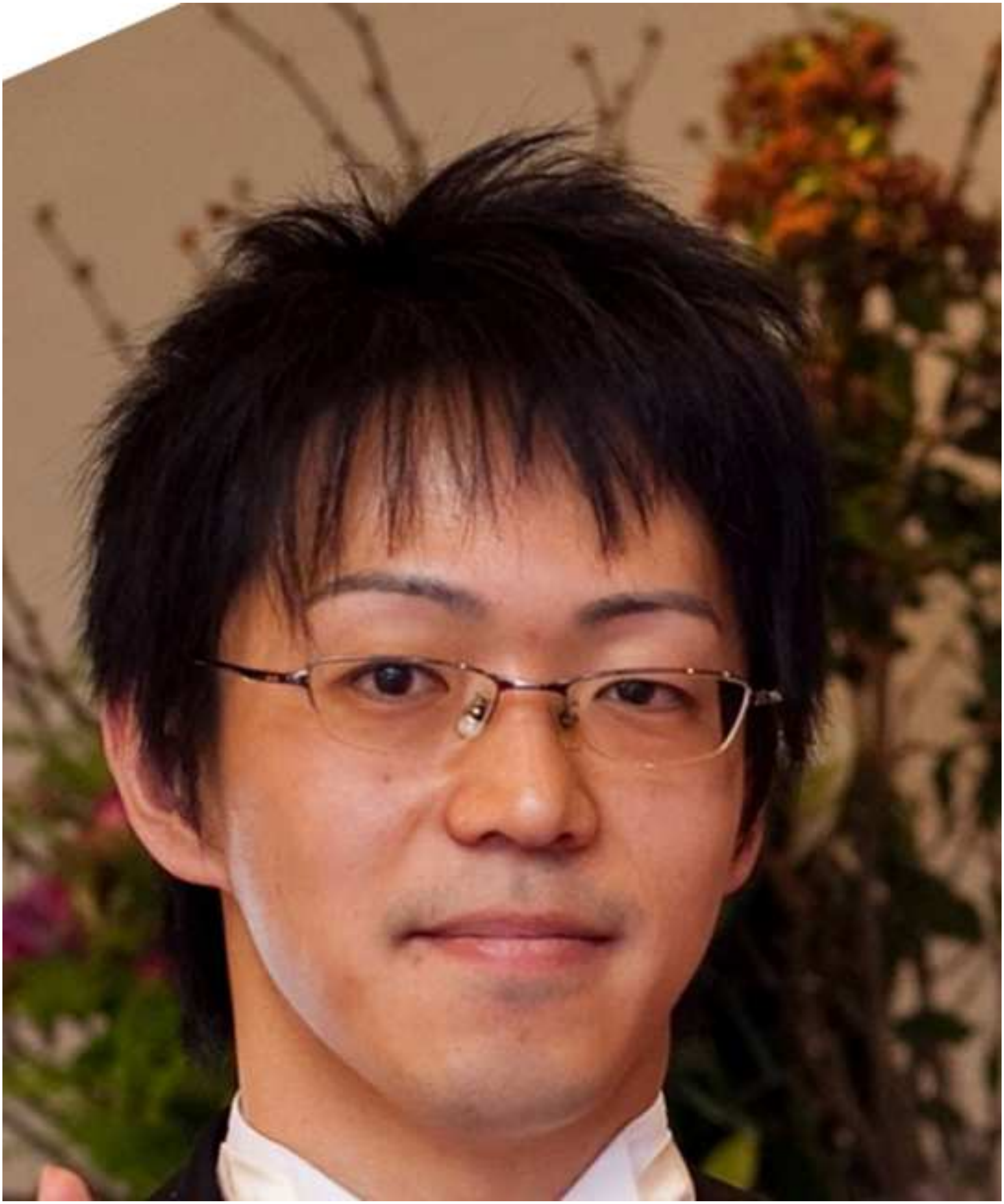
Zhenye Gan received the B.S. degree in application electronic technology, the M. E. degree in circuit and system from Northwest Normal University, Lanzhou, China, and the Ph.D. degrees in Chinese information processing from the Northwest University for Nationalities, Lanzhou, China, in 1999, 2007, and 2011, respectively. From 1999 to 2004 he was a teaching assistant at the College of Physics and Electronic Engineering, Northwest Normal University. From 2004 to 2011 he was a lecturer at the College of Physics and Electronic Engineering, Northwest Normal University. Now he is an associate professor at the same university.

Keiichi Tokuda received the B.E. degree in electrical and electronic engineering from Nagoya Institute of Technology, Nagoya, Japan, the M.E. and Dr. Eng. degrees in information processing from the Tokyo Institute of Technology, Tokyo, Japan, in 1984, 1986, and 1989, respectively. From 1989 to 1996 he was a Research Associate at the Department of Electronic and Electric Engineering, Tokyo Institute of Technology. From 1996 to 2004 he was a Associate Professor at

the Department of Computer Science, Nagoya Institute of Technology as Associate Professor, and now he is a Professor at the same institute. He is also an Honorary Professor at the University of Edinburgh.
He was an Invited Researcher at ATR Spoken Language Translation Research Laboratories, Japan from 2000 to 2013 and was a Visiting Researcher at Carnegie Mellon University from 2001 to 2002. He published over 80 journal papers and over 190 conference papers, and received six paper awards and two achievement awards. He was a member of the Speech Technical Committee of the IEEE Signal Processing Society from 2000 to 2003, a member of ISCA Advisory Council and an associate editor of IEEE Transactions on Audio, Speech & Language Processing, and acts as organizer and reviewer for many major speech conferences, workshops and journals. He is a IEEE Fellow and ISCA Fellow. His research interests include speech coding, speech synthesis and recognition, and statistical machine learning.

Hongwu Yang
Click here to download high resolution image

Keiichiro Oura
Click here to download high resolution image



Haiyan Wang
Click here to download high resolution image

zhenye Gan
Click here to download high resolution image

Keiichi Tokuda
Click here to download high resolution image