Cross-lingual speaker adaptation based on factor analysis using bilingual speech data for HMM-based speech synthesis

Takenori Yoshimura, Kei Hashimoto, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda

Department of Scientific and Engineering Simulation, Nagoya Institute of Technology, Nagoya, Japan

Abstract

This paper proposes a cross-lingual speaker adaptation (CLSA) method based on factor analysis using bilingual speech data. A state-mapping-based method has recently been proposed for CLSA. However, the method cannot transform only speaker-dependent characteristics. Furthermore, there is no theoretical framework for adapting prosody. To solve these problems, this paper presents a CLSA framework based on factor analysis using bilingual speech data. In this proposed method, model parameters representing language-dependent acoustic features and factors representing speaker characteristics are simultaneously optimized within a unified (maximum likelihood) framework based on a single statistical model by using bilingual speech data. This simultaneous optimization is expected to deliver a better quality of synthesized speech for the desired speaker characteristics. Experimental results show that the proposed method can synthesize better speech than the state-mapping-based method.

Index Terms: cross-lingual speaker adaptation, factor analysis, HMM-based speech synthesis

1. Introduction

The advance of internationalization has rapidly increased opportunities to communicate with people who speak different languages. However, language barriers often give rise to insufficient communication due to the difficulty of fluently speaking foreign languages. To overcome language barriers, speech-to-speech translation (S2ST) systems that translate input speech into target language speech are required. S2ST typically consists of three techniques: automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS). Conventional S2ST systems output speech with certain speaker characteristics that are unchanged even if the input speaker changes. This results in unnatural communication (i.e., the input speaker cannot be identified from the output speech). To solve this problem, cross-lingual speaker adaptation (CLSA) techniques that output speech in another language with input speaker characteristics have been developed in TTS. Hidden Markov model (HMM)-based speech synthesis systems [1, 2] are widely used for CLSA [3, 4, 5, 6, 7, 8, 9] because they can synthesize speech with various characteristics by estimating model parameters from a small amount of adaptation data, thus making them ideal for the purposes of CLSA.

Intra-lingual speaker adaptation, which is usually just called "speaker adaptation," transforms a source model into an input speaker model using a limited amount of speech data of the input speaker. Maximum likelihood linear regression (MLLR) [10, 11], which is one of the most well-known speaker adaptation techniques in HMM-based speech synthesis, can change the speaker characteristics of synthesized speech by linear transformation of the source model parameters. In this method, it is assumed that the transforms represent only the target speaker characteristics. Recently, a state-mapping-based method has been proposed as a CLSA method using MLLR [4]. This method can adapt the speaker characteristics of an output language speech by applying linear transforms in the input language to the models in the output language according to the state-mapping information. However, since the transforms include both language-dependent characteristics and speaker-dependent characteristics, this method cannot transform only speaker-dependent characteristics. In addition, the mapping information established by minimizing the Kullback-Leibler divergence (KLD) between the two states of HMMs is not guaranteed to be optimal. That is, there is no theoretical framework for adapting prosody such as rhythm and accent that have a longer dependency than one state.

To overcome these problems, we propose a CLSA method based on factor analysis using bilingual speech data[†]. In an eigenvoice method based on factor analysis [12], model parameters and factors expressing speaker characteristics in a certain language are simultaneously optimized within a unified maximum likelihood (ML) framework based on a single statistical model. In this paper, we extend the factor analysis-based eigenvoice model to the bilingual eigenvoice model. The proposed method simultaneously optimizes the model parameters for each of the input and output languages and factors by using bilingual speech data. By assuming that the factors representing speaker characteristics are common in the two languages, the approach can estimate model parameters considering the relation between the acoustic features in the two languages. In the adaptation step, the factors estimated from the adaptation data in the input language are applied to the output language directly. As a result, the proposed method can synthesize high-quality speech with the desired speaker characteristics in the output language. Furthermore, since the estimated factors that also express prosody of the input speaker is not constrained by the state structure of HMM, unlike the state-mapping-based method, the proposed method is potentially able to adapt prosody of the input speaker.

The rest of this paper is organized as follows. Section 2 describes the state-mapping-based method. Section 3 presents the intra-lingual speaker adaptation based on factor analysis. Section 4 proposes the CLSA technique based on factor analysis, and subjective listening test results are discussed in Section 5. Finally, conclusions and future work are presented in Section 6.

[†]Bilingual data is speech data uttered by persons who are able to speak two languages equally well.

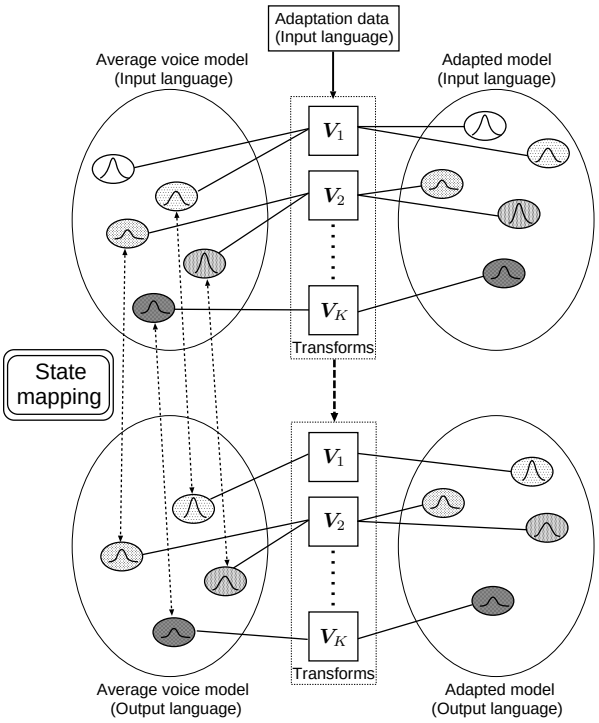

Figure 1: *Overview of the state-mapping-based method.*

2. State-mapping-based method

The basic idea of the state-mapping-based method for CLSA [4] is mapping the states of HMMs in an output language to ones in an input language. Figure 1 shows the overview of this method. First, average voice models in the input and output languages are respectively trained from speech data of various speakers by speaker adaptive training (SAT) [13]. State-mapping between these two models is then established. These mappings are estimated by searching for the state $\widehat{i}$ in the input language model that gives the minimum symmetric KLD $D_{\mathrm{KL}}(j,i)$ for each state j in the output language model:

$$\widehat{i} \;=\; \arg\min_{i} D_{\mathrm{KL}}(j,i). \tag{1}$$

Assuming that each state is represented by a single Gaussian pdf, the symmetric KLD between the two states in the input and output languages is calculated as

$$D_{\mathrm{KL}}(j,i) \;=\; D_{\mathrm{KL}}(j \,||\, i) + D_{\mathrm{KL}}(i \,||\, j), \tag{2}$$

$$D_{\mathrm{KL}}(i \,||\, j) \;=\; \frac{1}{2}\ln\left(\frac{\left|\boldsymbol{\Sigma}_j^{(O)}\right|}{\left|\boldsymbol{\Sigma}_i^{(I)}\right|}\right) + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_j^{(O)^{-1}}\boldsymbol{\Sigma}_i^{(I)}\right)$$
$$-\frac{D}{2} + \frac{1}{2}\left(\boldsymbol{\mu}_j^{(O)} - \boldsymbol{\mu}_i^{(I)}\right)^{\mathsf{T}}\boldsymbol{\Sigma}_j^{(O)^{-1}}\left(\boldsymbol{\mu}_j^{(O)} - \boldsymbol{\mu}_i^{(I)}\right), \tag{3}$$

where $(\cdot)^{(I)}$ and $(\cdot)^{(O)}$ respectively represent variables in the input and output languages, $\boldsymbol{\mu}^{(\cdot)}$ and $\boldsymbol{\Sigma}^{(\cdot)}$ denote the mean vector and covariance matrix of the Gaussian pdf associated with the state indicated by its subscript, and D is the dimension of observation vector. Then, MLLR [10] or constrained MLLR (CMLLR) [11] transforms $\boldsymbol{V} = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_K\}$ estimated from adaptation data in the input language are directly applied to the average voice models in the output language according to the established state-mapping information. By transforming all Gaussian pdfs in the output language in this way, CLSA is achieved. However, there are two theoretical problem: this method cannot transform only speaker-dependent characteristics, and cannot adapt prosody of an target speaker.

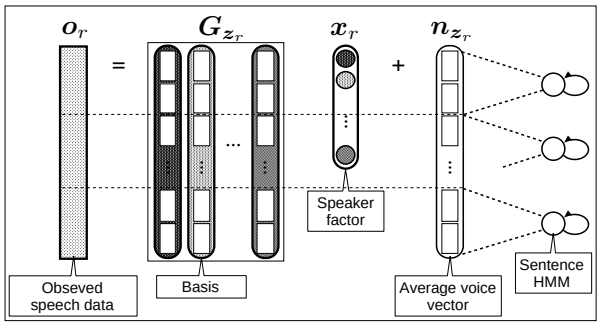

Figure 2: *Process of generating observation sequences in an eigenvoice model based on factor analysis.*

3. Speaker adaptation based on factor analysis

Eigenvoice-based methods perform speaker adaptation by changing a small number of parameters that represent various characteristics, such as speaker and speaking style. As an eigenvoice-based method, HMM-based speech synthesis using factor analysis has been proposed [12]. This approach assumes that the process of generating observation sequences is based on factor analysis, which is a statistical method for modeling the covariance structure of high-dimensional static data using a small number of latent variables. Figure 2 shows the model structure. The observation sequences $\boldsymbol{o}_r$ of speaker r are represented by

$$\boldsymbol{o}_r \;=\; \boldsymbol{G}_{\boldsymbol{z}_r}\boldsymbol{x}_r + \boldsymbol{n}_{\boldsymbol{z}_r}, \tag{4}$$

where $\boldsymbol{x}_r$, $\boldsymbol{n}_{\boldsymbol{z}_r}$, and $\boldsymbol{G}_{\boldsymbol{z}_r}$ denote the factor, noise vector, and loading matrix, respectively. This model can express various speaker characteristics by changing the factor $\boldsymbol{x}_r$. Moreover, since the noise vector $\boldsymbol{n}_{\boldsymbol{z}_r}$ and loading matrix $\boldsymbol{G}_{\boldsymbol{z}_r}$ are generated stochastically according to state sequence $\boldsymbol{z}_r$, the model can represent variable-length observation directly.

The factor $\boldsymbol{x}_r$ and noise vector $\boldsymbol{n}_r$ are often given by the following Gaussian distribution:

$$\boldsymbol{x}_r \;\sim\; \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \tag{5}$$
$$\boldsymbol{n}_{\boldsymbol{z}_r} \;\sim\; \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{z}_r}, \boldsymbol{\Sigma}_{\boldsymbol{z}_r}). \tag{6}$$

Under this condition, the likelihood function for acoustic feature sequences $\boldsymbol{o}$ of all speakers $(r = 1, 2, \ldots, R)$ is written as

$$P(\boldsymbol{o} \mid \boldsymbol{\Lambda}) \;=\; \prod_{r=1}^{R}\sum_{\boldsymbol{z}_r}\int P(\boldsymbol{o}_r, \boldsymbol{z}_r, \boldsymbol{x}_r \mid \boldsymbol{\Lambda})\, d\boldsymbol{x}_r$$
$$= \prod_{r=1}^{R}\sum_{\boldsymbol{z}_r}\int P(\boldsymbol{o}_r \mid \boldsymbol{z}_r, \boldsymbol{x}_r, \boldsymbol{\Lambda}) P(\boldsymbol{z}_r \mid \boldsymbol{\Lambda})$$
$$\times P(\boldsymbol{x}_r)\, d\boldsymbol{x}_r, \tag{7}$$
$$P(\boldsymbol{o}_r \mid \boldsymbol{z}_r, \boldsymbol{x}_r, \boldsymbol{\Lambda}) \;=\; \mathcal{N}(\boldsymbol{o}_r \mid \boldsymbol{G}_{\boldsymbol{z}_r}\boldsymbol{x}_r + \boldsymbol{\mu}_{\boldsymbol{z}_r}, \boldsymbol{\Sigma}_{\boldsymbol{z}_r}), \tag{8}$$

where $\mathbf{\Lambda}$ is a set of model parameters that includes the loading matrix $\boldsymbol{G}_{\boldsymbol{z}_r}$, the noise mean vector $\boldsymbol{\mu}_{\boldsymbol{z}_r}$, and the noise covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{z}_r}$. These parameters and factors are simultaneously optimized within a unified ML framework. However, the parameter estimation is computationally intractable due to the combination of latent variables: factor and state sequence. Hence, the distribution of these variables is computed via the variational expectation-maximization (VEM) algorithm [12], which is an iterative algorithm that is closely related to the standard expectation-maximization (EM) algorithm. For speaker adaptation, the factor $\boldsymbol{x}_{r'}$ is estimated using the trained model and adaptation data of the target speaker r'. The estimated factor $\boldsymbol{x}_{r'}$ represents the characteristics of the target speaker r' so that synthesized speech with the characteristics of speaker r' can be obtained.

4. Cross-lingual speaker adaptation based on factor analysis

In the eigenvoice method based on factor analysis [12] described above, the model parameters and factors expressing speaker characteristics in a certain language are simultaneously optimized within a unified ML framework based on a single statistical model. In this paper, we extend the factor analysis-based eigenvoice model to the bilingual eigenvoice model. The proposed method simultaneously optimizes the model parameters for each of the input and output languages and factors by using bilingual speech data. By assuming that the factors representing speaker characteristics are common in the two language, the proposed method can estimate model parameters considering the relation between acoustic features in the two languages. In the adaptation step, the factors estimated using adaptation data in the input language are applied to the output language directly. Therefore, the same speaker characteristics can be obtained in another language. Furthermore, since the estimated factors that also express prosody of the input speaker is not influenced by language-dependent information such as context or the state structure of HMM, the proposed method can adapt the prosody of the input speaker in principle.

4.1. Model structure based on factor analysis in multilingual space

Figure 3 gives the model structure of the proposed method. Bilingual data $\boldsymbol{o}$, which consists of input language data $\boldsymbol{o}^{(I)}$ and output language data $\boldsymbol{o}^{(O)}$ uttered by bilingual speakers $(r = 1, 2, \ldots, R)$, is used for training the proposed model. The likelihood function is calculated as

$$
\begin{aligned}
P(\boldsymbol{o} \mid \mathbf{\Lambda}) &= \prod_{r=1}^{R} \int P\left(\boldsymbol{o}_r^{(O)}, \boldsymbol{o}_r^{(I)}, \boldsymbol{x}_r \,\middle|\, \mathbf{\Lambda}^{(O)}, \mathbf{\Lambda}^{(I)}\right) d\boldsymbol{x}_r \\
&= \prod_{r=1}^{R} \int P\left(\boldsymbol{o}_r^{(O)} \,\middle|\, \boldsymbol{x}_r, \mathbf{\Lambda}^{(O)}\right) P\left(\boldsymbol{o}_r^{(I)} \,\middle|\, \boldsymbol{x}_r, \mathbf{\Lambda}^{(I)}\right) \\
&\quad \times P\left(\boldsymbol{x}_r\right) d\boldsymbol{x}_r, \qquad (9)
\end{aligned}
$$

where $\mathbf{\Lambda}$ is composed of sets of model parameters for the input and output languages, $\mathbf{\Lambda}^{(I)}$ and $\mathbf{\Lambda}^{(O)}$. A set of model parameters $\mathbf{\Lambda}^{(\cdot)}$ includes the loading matrix $\boldsymbol{G}^{(\cdot)}$, the noise mean vector $\boldsymbol{\mu}^{(\cdot)}$, and the noise covariance matrix $\mathbf{\Sigma}^{(\cdot)}$. The factor $\boldsymbol{x}_r$ representing speaker characteristics is independent from languages, i.e., the common speaker factor is used in the input and output languages, as shown in Fig. 3. In Eq. (9),

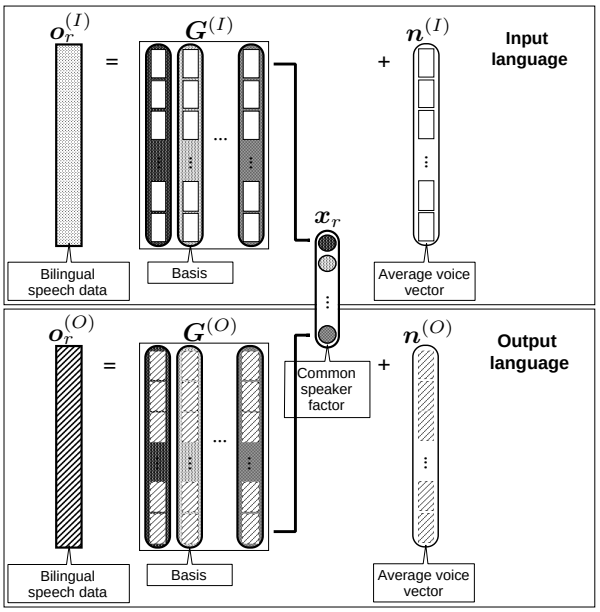

Figure 3: *Model structure of cross-lingual speaker adaptation based on factor analysis using bilingual speech data.*

$P\left(\boldsymbol{o}_r^{(I)} \,\middle|\, \boldsymbol{x}_r, \mathbf{\Lambda}^{(I)}\right)$ and $P\left(\boldsymbol{o}_r^{(O)} \,\middle|\, \boldsymbol{x}_r, \mathbf{\Lambda}^{(O)}\right)$ denote the output probability of the input and output language, respectively. This approach can simultaneously model the speech data of different languages. Therefore, the model parameters $\mathbf{\Lambda}^{(I)}$ and $\mathbf{\Lambda}^{(O)}$ are trained by a similar algorithm to the one using monolingual data (Eq. (7)) in the training step. Note that state sequence is omitted from above.

4.2. Variational expectation-maximization algorithm

A lower bound $\mathcal{F}$ of the log likelihood is defined by using Jensen's inequality:

$$
\begin{aligned}
& \log P(\boldsymbol{o} \mid \mathbf{\Lambda}) \\
= \; & \log \prod_{r=1}^{R} \sum_{\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}} \\
& \int P\left(\boldsymbol{o}_r^{(O)}, \boldsymbol{z}_r^{(O)}, \boldsymbol{o}_r^{(I)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r \,\middle|\, \mathbf{\Lambda}\right) d\boldsymbol{x}_r \\
\geq \; & \sum_{r=1}^{R} \sum_{\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}} \int Q\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r\right) \\
& \times \log \frac{P\left(\boldsymbol{o}_r^{(O)}, \boldsymbol{z}_r^{(O)}, \boldsymbol{o}_r^{(I)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r \,\middle|\, \mathbf{\Lambda}\right)}{Q\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r\right)} d\boldsymbol{x}_r \\
\equiv \; & \mathcal{F}, \qquad (10)
\end{aligned}
$$

where $\boldsymbol{z}_r^{(\cdot)}$ is a state sequence and $Q\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r\right)$ is an arbitrary function. Then, the relation between the log likelihood and the lower bound $\mathcal{F}$ is represented as

$$
\mathcal{F} = \log P(\boldsymbol{o} \mid \mathbf{\Lambda}) - D_{\mathrm{KL}}(Q \,||\, P). \qquad (11)
$$

Therefore, maximizing the lower bound $\mathcal{F}$ is equivalent to minimizing the KLD $D_{\mathrm{KL}}(Q \,||\, P)$. The arbitrary function

$Q\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r\right)$ is then regarded as an approximate posterior distribution because $Q\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r\right)$ approximates the true posterior distribution $P\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r \,\middle|\, \boldsymbol{o}_r^{(O)}, \boldsymbol{o}_r^{(I)}, \boldsymbol{\Lambda}\right)$ when $D_{\mathrm{KL}}(Q \mid\mid P)$ is reduced. The approximate posterior distribution $Q\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r\right)$ is estimated by maximizing the lower bound $\mathcal{F}$. However, it is difficult to estimate $Q\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r\right)$ directly. To compute it easily, the VEM method assumes that the latent variables are independent under the condition that observation sequences $\boldsymbol{o}_r$ is given:

$$Q\left(\boldsymbol{z}_r^{(O)}, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r\right) = Q\left(\boldsymbol{z}_r^{(O)}\right) Q\left(\boldsymbol{z}_r^{(I)}\right) Q(\boldsymbol{x}_r). \quad (12)$$

Under this assumption, the optimal posterior distributions can be computed by a similar procedure to the EM algorithm, which increases the value of the objective function $\mathcal{F}$ at each iteration until convergence. The posterior distributions of state sequences $Q\left(\boldsymbol{z}_r^{(\cdot)}\right)$ and factor $Q(\boldsymbol{x}_r)$ are obtained by adapting the variational method to the lower bound $\mathcal{F}$:

$$Q\left(\boldsymbol{z}_r^{(O)}\right) = C_{\boldsymbol{z}_r^{(O)}} P\left(\boldsymbol{z}_r^{(O)} \,\middle|\, \boldsymbol{\Lambda}^{(O)}\right) \exp\left[\int Q(\boldsymbol{x}_r) \times \log P\left(\boldsymbol{o}_r^{(O)} \,\middle|\, \boldsymbol{z}_r^{(O)}, \boldsymbol{x}_r, \boldsymbol{\Lambda}^{(O)}\right) d\boldsymbol{x}_r\right], \quad (13)$$

$$Q\left(\boldsymbol{z}_r^{(I)}\right) = C_{\boldsymbol{z}_r^{(I)}} P\left(\boldsymbol{z}_r^{(I)} \,\middle|\, \boldsymbol{\Lambda}^{(I)}\right) \exp\left[\int Q(\boldsymbol{x}_r) \times \log P\left(\boldsymbol{o}_r^{(I)} \,\middle|\, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r, \boldsymbol{\Lambda}^{(I)}\right) d\boldsymbol{x}_r\right], \quad (14)$$

$$Q(\boldsymbol{x}_r) = C_{\boldsymbol{x}_r} P(\boldsymbol{x}_r) \exp\left[\sum_{\boldsymbol{z}_r^{(O)}} Q\left(\boldsymbol{z}_r^{(O)}\right) \log P\left(\boldsymbol{o}_r^{(O)} \,\middle|\, \boldsymbol{z}_r^{(O)}, \boldsymbol{x}_r, \boldsymbol{\Lambda}^{(O)}\right) + \sum_{\boldsymbol{z}_r^{(I)}} Q\left(\boldsymbol{z}_r^{(I)}\right) \log P\left(\boldsymbol{o}_r^{(I)} \,\middle|\, \boldsymbol{z}_r^{(I)}, \boldsymbol{x}_r, \boldsymbol{\Lambda}^{(I)}\right)\right], \quad (15)$$

where $C_{\boldsymbol{z}_r^{(I)}}$, $C_{\boldsymbol{z}_r^{(O)}}$, and $C_{\boldsymbol{x}_r}$ are the normalization terms to ensure $\sum_{\boldsymbol{z}_r^{(\cdot)}} Q\left(\boldsymbol{z}_r^{(\cdot)}\right) = 1$ and $\int Q\left(\boldsymbol{x}_r\right) d\boldsymbol{x}_r = 1$. Since the posterior distributions depend on each other, they should be optimized simultaneously by iterative procedures. If the factor $\boldsymbol{x}_r$ and the noise vector are given by Gaussian distributions as Eqs. (5) and (6), the posterior distribution $Q\left(\boldsymbol{x}_r\right)$ is expressed as

$$Q\left(\boldsymbol{x}_r\right) = \mathcal{N}\left(\boldsymbol{x}_r \,\middle|\, \hat{\boldsymbol{\mu}}_{\boldsymbol{x}_r}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_r}\right), \quad (16)$$

where $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}_r}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_r}$ denote the mean vector and covariance matrix of factor $\boldsymbol{x}_r$, respectively.

4.3. Adaptation step

In the adaptation step (Fig. 4), the optimal speech parameter sequence in the output language $\hat{\boldsymbol{o}}_{r'}^{(O)}$ of an input speaker r', who is not a bilingual speaker, is generated by maximizing the output probability:

$$\hat{\boldsymbol{o}}_{r'}^{(O)} = \arg\max_{\boldsymbol{o}_{r'}^{(O)}} \int P\left(\boldsymbol{o}_{r'}^{(O)}, \boldsymbol{o}_{r'}^{(I)}, \boldsymbol{x}_{r'} \,\middle|\, \boldsymbol{\Lambda}^{(O)}, \boldsymbol{\Lambda}^{(I)}\right) d\boldsymbol{x}_{r'} = \arg\max_{\boldsymbol{o}_{r'}^{(O)}} \int P\left(\boldsymbol{o}_{r'}^{(O)} \,\middle|\, \boldsymbol{x}_{r'}, \boldsymbol{\Lambda}^{(O)}\right) \times P\left(\boldsymbol{x}_{r'} \,\middle|\, \boldsymbol{o}_{r'}^{(I)}, \boldsymbol{\Lambda}^{(I)}\right) d\boldsymbol{x}_{r'}. \quad (17)$$

It is difficult to estimate the true posterior distribution of the factor $\boldsymbol{x}_{r'}$ because the likelihood function includes multiple latent variables. Therefore, the same approximation as the training step, i.e., the VEM algorithm, is applied to estimate the approximate posterior distribution of $\boldsymbol{x}_{r'}$. In addition, the maximum a posterior approximation is applied. Consequently, the optimal speech parameter sequence is generated by using the mean vector of the posterior distribution $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}_{r'}}$, which obtains the maximum posterior probability, as

$$\hat{\boldsymbol{o}}_{r'}^{(O)} \approx \arg\max_{\boldsymbol{o}_{r'}^{(O)}} \int P\left(\boldsymbol{o}_{r'}^{(O)} \,\middle|\, \boldsymbol{x}_{r'}, \boldsymbol{\Lambda}^{(O)}\right) Q\left(\boldsymbol{x}_{r'}\right) d\boldsymbol{x}_{r'} \approx \arg\max_{\boldsymbol{o}_{r'}^{(O)}} P\left(\boldsymbol{o}_{r'}^{(O)} \,\middle|\, \hat{\boldsymbol{\mu}}_{\boldsymbol{x}_{r'}}, \boldsymbol{\Lambda}^{(O)}\right). \quad (18)$$

From this equation, the estimated factor is immediately applied to the output probability distribution in the output language. As a result, the synthesized speech does not depend on the input language-dependent characteristics, so that the high-performance sound is expected.

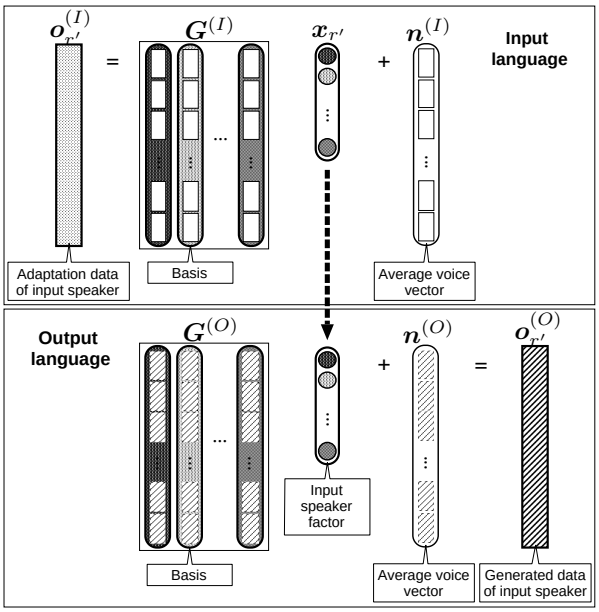

Figure 4: *Overview of the adaptation step in cross-lingual speaker adaptation based on factor analysis using bilingual speech data.*

4.4. Cross-lingual speaker adaptation based on speaker interpolation

In the proposed framework, CLSA based on speaker interpolation can be represented. This method replaces bases of loading matrices $\boldsymbol{G}^{(\cdot)}$ with the mean vectors $\bar{\boldsymbol{\mu}}_r^{(\cdot)}$ of the speaker-dependent model of each training speaker:

$$\boldsymbol{G}^{(I)} = \left[\begin{array}{cccc} \bar{\boldsymbol{\mu}}_1^{(I)} & \bar{\boldsymbol{\mu}}_2^{(I)} & \cdots & \bar{\boldsymbol{\mu}}_R^{(I)} \end{array} \right], \quad (19)$$

$$\boldsymbol{G}^{(O)} = \left[\begin{array}{cccc} \bar{\boldsymbol{\mu}}_1^{(O)} & \bar{\boldsymbol{\mu}}_2^{(O)} & \cdots & \bar{\boldsymbol{\mu}}_R^{(O)} \end{array} \right], \quad (20)$$

and substitutes noise mean vectors $\boldsymbol{\mu}^{(\cdot)}$ with zero vectors:

$$\boldsymbol{\mu}^{(I)} = \boldsymbol{\mu}^{(O)} = \boldsymbol{0}. \quad (21)$$

Since the training speaker models are weighted by the factor $\boldsymbol{x}_{r'}$ of the input speaker r', this method is regarded as speaker interpolation. By applying the factor $\boldsymbol{x}_{r'}$ estimated from the adaptation data $\boldsymbol{o}_{r'}^{(I)}$ of the input speaker r' to the output language directly, speaker interpolation in the output language is achieved. Hence, the method is considered an approach that approximates ML criterion based optimization in the proposed method. To evaluate the effectiveness of model parameter optimization in the proposed method, we included this interpolation approach in the following experiments.

5. Experiments

5.1. Experimental setups

We used a Japanese-English and Japanese-Chinese bilingual monologue speech database [14] in these experiments. The speech signals were sampled at a rate of 48 kHz and windowed by a 25-ms Hamming window with a 5-ms shift. Feature vectors consisted of 34 mel-cepstral coefficients including the zeroth coefficient, fundamental frequency (F_0), and their first and second time derivatives. A five-state left-to-right no-skip context-dependent MSD-HSMM [15, 16] was used. The input language was English and the output language was Japanese. For training the models, 2,700 sentences uttered by five female speakers were used in each language. Two English sentences were used as adaptation data. The target speakers were four female speakers who were not included in the training speakers.

Two subjective listening tests were conducted. The first test evaluated the naturalness of the synthesized speech by the mean opinion score (MOS) test method, and the second one evaluated the speaker similarity between the target speech and the synthesized speech for each model by the differential MOS (DMOS) test method. In the MOS test, after the subjects had listened to a test sample, they were asked to assign it a naturalness score on a five-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). In the DMOS test, after the subjects had listened to Japanese natural speech of the target speaker and a test sample, they were asked to assign it a similarity score on a five-point scale (5: very close, 4: close, 3: fair, 2: far, 1: very far). Ten subjects evaluated 15 and 12 sentences, which were randomly chosen from 45 sentences, in the MOS and DMOS tests, respectively.

The following methods were compared.

- **SD**: The speaker-dependent model of each target speaker.
- **SM**: CLSA based on state-mapping (conventional method).
- **IP**: CLSA based on speaker interpolation (approximation method).
- **FA**: CLSA based on factor analysis (proposed method).

Table 1: *The number of sentences for training the speaker-dependent model.*

Speaker	EJF04	EJF06	EJF10	EJF11
No. of sentences	1400	1200	167	153

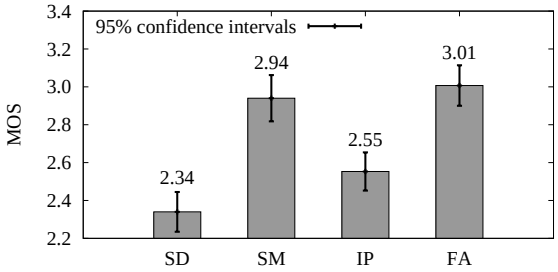

Figure 5: *Experimental results (naturalness).*

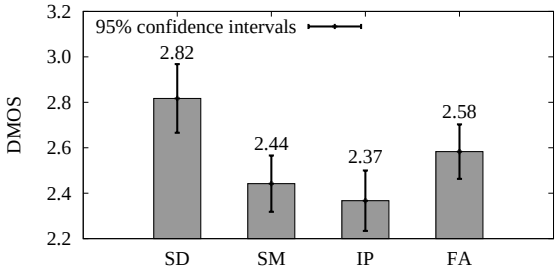

Figure 6: *Experimental results (speaker similarity).*

The number of sentences used for training the **SD** models is listed in Table 1. The parameters of the spectral, F_0, and duration models were adapted to the target speaker in **IP** and **FA**, but adaptation of durations was not performed in **SM** because it is considered unsuitable [4]. For constructing **IP** models using Eqs. (19), (20), and (21), speaker-dependent models of training speakers were constructed by applying CMLLR transformations for each training speaker to average voice models. The average voice models for the input and output languages were the same as the ones used in **SM**. The number of bases for **FA** was three in these experiments.

5.2. Experimental results

The results of the MOS and DMOS listening tests are shown in Figs. 5 and 6, respectively, with the vertical line indicating the 95% confidence intervals. Figure 5 shows that **FA** significantly improved the naturalness of the synthesized speech compared with **SD**. This is because **FA** can estimate appropriate acoustic models by efficiently using training data of various speakers while **SD** uses the data of only the target speaker. **IP** also obtained a higher score than **SD**. However, **FA** achieved a greater improvement than **IP**, indicating that **FA** can estimate the bases and noise vector properly and that simultaneous op-

Table 2: *DMOS for each target speaker.*

	SD	SM	IP	FA
EJF04	2.52	2.43	2.81	2.38
EJF06	3.14	2.71	2.60	2.54
EJF10	2.90	2.35	2.00	2.77
EJF11	2.58	2.24	2.18	2.58
Mean	2.82	2.44	2.37	2.58

timization based on the proposed method is effective. **FA** and **SM** had similar results. **SM** often occurred discontinuity errors due to not considering the contextual factors in the output language. In contrast, **FA** generated smooth synthesized speech, although the naturalness of the speech was often degraded by over-smoothing. To overcome the over-smoothing problem, a training algorithm for the proposed method will be investigated in the future.

It can be seen from Fig. 6 that **FA** delivered suitable speaker similarity. **FA** obtained a higher score than **SM**, i.e., **FA** synthesized speech with more similar speaker characteristics to the target speaker ones than **SM**. Furthermore, **FA** outperformed **IP**, although **FA** had a smaller number of bases than **IP**. These results suggest that the proposed model can represent various speaker characteristics and that simultaneous optimization is effective to estimate the appropriate model parameters in terms of representation of speaker variations. Table 2 gives the detailed DMOS results for each target speaker. From Tables 1 and 2, **FA** showed a nearly identical performance to **SD** in the case of speaker-dependent models trained by a small amount of training data, such as EJF10 and EJF11. This indicates that the proposed method has potential for delivering more suitable speaker similarity by investigating the model structure such as the number of bases.

6. Conclusion

We proposed a cross-lingual speaker adaptation method based on factor analysis using bilingual speech data for HMM-based speech synthesis. In the proposed method, model parameters representing language-dependent acoustic features and factors representing speaker characteristics are simultaneously optimized within a unified framework based on a single statistical model by using bilingual speech data. The results of subjective listening tests indicated that the proposed method can generate natural synthesized speech with suitable speaker similarity. In addition, the effectiveness of the simultaneous optimization was shown by the experimental results. Our future work is to investigate the model structure, such as the number of bases and the utilization of monolingual speech data.

7. Acknowledgments

This research was partly funded by Core Research for Evolutional Science and Technology (CREST) from Japan Science and Technology Agency (JST).

8. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc. of Eurospeech 1999*, pp. 2347–2350, 1999.

[2] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," *Proc. of ICASSP 2000*, vol. 3, pp. 1315–1318, 2000.

[3] Y. J. Wu, S. King, and K. Tokuda, "Cross-Lingual speaker adaptation for HMM-based speech synthesis," *Proc. of ISCSLP 2008*, pp. 1–4, 2008.

[4] Y. J. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," *Proc. of Interspeech 2009*, pp. 528–531, 2009.

[5] K. Oura, J. Yamagishi, M. Wester, S. King, and K. Tokuda, "Unsupervised cross-lingual speaker adaptation for speech-to-speech translation system," *Proc. of ICASSP 2010*, pp. 4594–4597, 2010.

[6] M. Gibson, T. Hirsimäki, R. Karhila, M. Kurimo, and W. Byrne, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," *Proc. of ICASSP 2010*, pp. 4642–4645, 2010

[7] X. Peng, K. Oura, Y. Nankaku, and, K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis considering differences between language-dependent average voices," *Proc. of ICSP 2010*, pp. 605–608, 2010.

[8] N. Hattori, T. Toda, H. Kawai, H. Saruwatari, and K. Shikano, "Speaker-adaptive speech synthesis based on eigenvoice conversion and language-dependent prosodic conversion in speech-to-speech translation," *Proc. of Interspeech 2011*, pp. 2769–2772, 2011.

[9] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulović, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Trans.*, vol. 20, no. 6, pp. 1713–1724, 2012.

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[12] K. Kazumi, Y. Nankaku, and K. Tokuda, "Factor analyzed voice models for HMM-based speech synthesis," *Proc. of ICASSP 2010*, pp. 4234–4237, 2010.

[13] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans.*, vol. E90–D, no. 2, pp. 533–543, 2007.

[14] ALAGIN language/voice resources site, Online: https://alaginrc.nict.go.jp/resources/nictmastar/menuspeechlist/speechoutline.html, accessed on 2 Mar 2013.

[15] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," *Proc. of ICASSP 1999*, vol. 1, pp. 229–232, 1999.

[16] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans.*, vol. E90–D, no. 5, pp. 825–834, 2007.