# Speaker-Adaptive Speech Synthesis Based on Eigenvoice Conversion and Language-Dependent Prosodic Conversion in Speech-to-Speech Translation

*Nobuhiko Hattori*[1], *Tomoki Toda*[1,2], *Hisashi Kawai*[2], *Hiroshi Saruwatari*[1], *Kiyohiro Shikano*[1]

[1]Graduate School of Information Science, Nara Institute of Science and Technology, Japan
[2]National Institute of Information and Communications Technology, Japan

tomoki@is.naist.jp

## Abstract

This paper describes a novel approach based on voice conversion (VC) to speaker-adaptive speech synthesis for speech-to-speech translation. Voice quality of translated speech in an output language is usually different from that of an input speaker of the translation system since a text-to-speech system is developed with another speaker's voices in the output language. To render the input speaker's voice quality in the translated speech, we propose a voice quality control method based on one-to-many eigenvoice conversion (EVC) and language-dependent prosodic conversion. Spectral parameters of the translated speech are effectively converted by one-to-many EVC enabling unsupervised speaker adaptation. Moreover, prosodic parameters are modified considering their global differences between the input and output languages. The effectiveness of the proposed method is confirmed by experimental evaluations on cross-lingual VC among Japanese, English, and Chinese.

**Index Terms**: speech-to-speech translation, speech synthesis, speaker adaptation, eigenvoice conversion, prosodic conversion

## 1. Introduction

Speech-to-speech translation is an effective technique to make it possible for us to communicate with each other beyond language barriers. Voices of an input speaker of a speech-to-speech translation system are translated into voices in an output language with three main techniques, automatic speech recognition (ASR), machine translation (MT), and text-to-speech (TTS) [1]. Voice quality of the translated speech is usually different from that of the input speaker since a TTS system needs to be developed with voices of another speaker in the output language. It is more effective if not only linguistic information but also non-linguistic information such as speaker individuality is conveyed by the translated speech.

To render the input speaker's voice quality in the translated voice, cross-lingual speech synthesis techniques have been studied. Recently a speech synthesis technique based on a hidden Markov model (HMM) [2] has attracted attention due to its flexible framework capable of voice quality control with model adaptation techniques. Unsupervised model adaptation and a mapping of model (or adaptation) parameters between different languages are essential techniques to achieve cross-lingual speaker adaptation. King *et al.* [3] proposed an unsupervised adaptation method based on a mapping of transforms between triphone units to be used in recognition and fullcontext units to be used in synthesis. Chen *et al.* [4] proposed a HMM state mapping method between different languages exploiting bilingual speech data sets. Gibson and Byrne [5] proposed a two-pass decision tree clustering technique to effectively cope with a model mapping problem and applied it to unsupervised model adaptation using a different language. These methods need a decoding process to perform model adaptation since linguistic units such as phonemes are used in the HMM. Therefore, the effect of decoding errors on the adaptation performance needs to be reduced.

As another approach, voice conversion (VC) techniques have been studied. The most popular method is to define a conversion function based on a Gaussian mixture model (GMM) [6, 7], which is usually developed with a parallel data set consisting of utterance pairs of source and target speakers. One approach to cross-lingual VC is to produce a parallel data set between speakers in different languages in some way. Abe *et al.* [8] proposed the use of a TTS system to generate voices in a different language based on a mapping of phoneme sets. Mashimo *et al.* [9] proposed the use of bilingual speaker's data. Erro *et al.* [10] proposed to generate *pseudo* parallel data from non-parallel data based on frame alignment between voices of different languages.

Recently another approach to cross-lingual VC has been proposed inspired by the model adaptation techniques. Eigenvoice conversion (EVC) [11], one of the effective methods for adaptive VC, uses multiple parallel data sets between a single speaker and multiple speakers to effectively achieve unsupervised adaptation of a GMM to an arbitrary speaker. Because specific linguistic units are not used in the GMM, voices of any language are straightforwardly accepted as adaptation data in the unsupervised adaptation. Malorie *et al.* [12] applied EVC to cross-lingual VC and reported its effectiveness.

In this paper, we propose VC techniques to develop speaker-adaptive speech synthesis in speech-to-speech translation. The EVC technique is used to convert spectral parameters of the translated speech into those of the input speaker. Moreover, to improve naturalness of the converted speech, a language-dependent prosodic conversion method is used to globally modify prosodic parameters considering their global differences between input and output languages. The effectiveness of the proposed methods is confirmed by several experimental evaluations assuming a speech-to-speech translation process among Japanese, English, and Chinese.

## 2. One-to-Many Eigenvoice Conversion

### 2.1. Eigenvoice GMM (EV-GMM)

The joint probability density function (p.d.f.) of the source and target feature vectors is modeled by the EV-GMM as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left([\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top; \boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}), \boldsymbol{\Sigma}_m^{(X,Y)}\right), \quad (1)$$

where the mean vector $\boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w})$ is written as

$$\boldsymbol{\mu}_m^{(X,Y)}(\boldsymbol{w}) = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)}(\boldsymbol{w}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{B}_m^{(Y)}\boldsymbol{w} + \boldsymbol{b}_m^{(Y)}(0) \end{bmatrix}. \quad (2)$$

In one-to-many EVC, the target mean vector of the $m^{\text{th}}$ mixture component is represented as a linear combination of a bias vector $\boldsymbol{b}_m^{(Y)}(0)$ and representative vectors $\boldsymbol{B}_m^{(Y)} = \left[\boldsymbol{b}_m^{(Y)}(1), \cdots, \boldsymbol{b}_m^{(Y)}(J)\right]$, where the number of representative vectors is $J$. The $J$-dimensional weight vector $\boldsymbol{w} = [w(1), \cdots, w(J)]^\top$ is adapted to an arbitrary target speaker while the parameter set of the EV-GMM $\boldsymbol{\lambda}^{(EV)}$ is tied over different target speakers.

### 2.2. Training

The tied parameter set of the EV-GMM is trained in advance using the multiple parallel data sets consisting of the single source speaker and many pre-stored target speakers. Let $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t^{(s)}$ be the feature vector of the source speaker and that of the $s^{\text{th}}$ pre-stored target speaker at frame $t$. Not only the tied parameter set $\boldsymbol{\lambda}^{(EV)}$ but also a set of the weight vectors $\boldsymbol{w}_{1:S} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_S\}$ adapted to individual pre-stored target speakers are optimized as follows:

$$\left\{\hat{\boldsymbol{\lambda}}^{(EV)}, \hat{\boldsymbol{w}}_{1:S}\right\} = \arg\max_{\left\{\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}_{1:S}\right\}} \prod_{s=1}^{S}\prod_{t=1}^{T_s} P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}_s\right). \quad (3)$$

To enable maximum a posteriori (MAP) estimation in the adaptation process, a prior p.d.f. of the weight vector is modeled by a Gaussian distribution as follows:

$$P\left(\boldsymbol{w}|\boldsymbol{\lambda}^{(w)}, \tau\right) = \mathcal{N}\left(\boldsymbol{w}; \boldsymbol{\mu}^{(w)}, \tau^{-1}\boldsymbol{\Sigma}^{(w)}\right), \quad (4)$$

where $\tau$ is a hyper-parameter. A model parameter set $\boldsymbol{\lambda}^{(w)}$ consisting of the mean vector $\boldsymbol{\mu}^{(w)}$ and the covariance matrix $\boldsymbol{\Sigma}^{(w)}$ is estimated using a set of the weight vectors optimized for individual pre-stored target speakers in Eq. (3).

### 2.3. Unsupervised adaptation

The EV-GMM is adapted to an arbitrary target speaker by estimating the optimum weight vector for given speech samples of the target speaker in a completely unsupervised manner, *i.e.*, using neither parallel data nor linguistic information. For a time sequence of the given target feature vectors $\boldsymbol{Y}'_1, \cdots, \boldsymbol{Y}'_T$, the MAP estimation of the weight vector is performed as follows:

$$\begin{aligned} \hat{\boldsymbol{w}} &= \arg\max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{Y}'_1, \cdots, \boldsymbol{Y}'_T, \boldsymbol{\lambda}) \\ &= \arg\max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{\lambda}^{(w)}, \tau) \prod_{t=1}^{T} P(\boldsymbol{Y}'_t|\boldsymbol{\lambda}^{(EV)}, \boldsymbol{w}). \end{aligned} \quad (5)$$

This adaptation process works reasonably well even if using only one or two utterances since the number of adaptive parameters (*i.e.*, the number of dimensions of the weight vector) is small enough.

### 2.4. Conversion

Based on the adapted EV-GMM, the source feature vectors are converted into the target feature vectors. The maximum likelihood estimation method considering dynamic features and global variance [7] is adopted in this paper.

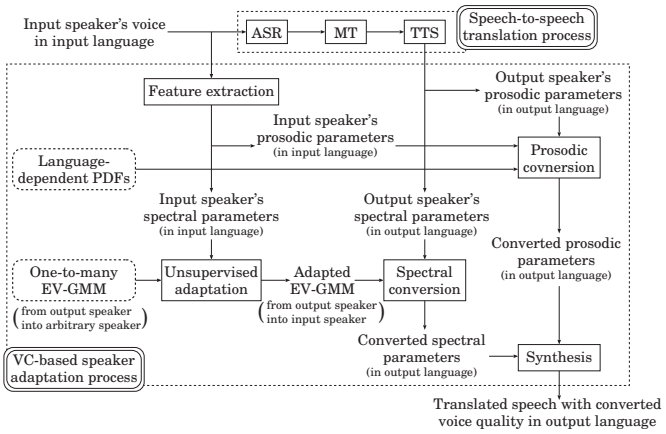


Figure 1: *Proposed speaker-adaptive speech synthesis framework for speech-to-speech translation system.*

## 3. Cross-Lingual Speech Synthesis Based on VC in Speech-to-Speech Translation

There are two main approaches to develop speaker-adaptive speech synthesis in speech-to-speech translation: one is to synthesize voices of the output language uttered by the input speaker as truly as possible (*e.g.*, presenting Japanese accented English if the input speaker's English is accented); and the other is to synthesize voices of the input speaker in the output language as fluently as native speakers. The use of bilingual data would be essential in the former approach but it is not always necessary in the latter approach. In this paper, we focus on the latter approach and propose a novel approach based on VC techniques without any bilingual data. A basic idea is to generate the input speaker's voices in the output language by properly mixing voices of various native speakers of the output language.

**Figure 1** shows the proposed framework. First the input speaker's voice is translated into a text in the output language by ASR and MT, and then speech parameters such as spectral and prosodic parameters are generated by TTS. After that, spectral parameters are converted with one-to-many EVC and prosodic parameters are converted based on language-dependent probability distribution functions (PDFs). This proposed framework has nice portability since it is straightforwardly applied to any speech-to-speech translation system. If the TTS system generates only speech waveforms, a speech analysis process is necessary to extract speech parameters from the generated output waveform. In this paper, HMM-based speech synthesis is used as the TTS system. Thanks to its parametric speech synthesis framework, speech parameters generated from the translated text are available to be used in one-to-many EVC without any speech analysis process.

### 3.1. Spectral Conversion Based on One-to-Many EVC

The output speaker of a speech-to-speech translation system is used as the source speaker and the input speaker of the system (*i.e.*, a user) is used as the target speaker to be adapted in one-to-many EVC. First one-to-many EV-GMM is adapted into the input speaker using only his/her voices input to the system. The spectral parameter sequence generated by the TTS is converted with the adapted EV-GMM so as to exhibit the input speaker's voice quality. An excitation parameter such as aperiodic components may also be converted in the same manner using another EV-GMM developed for such a parameter.

To train the EV-GMMs, it is necessary to use multiple parallel data sets consisting of the speaker modeled by the TTS

system as the single source speaker and a lot of other speakers as the pre-defined target speakers. However, it is laborious work to collect those data sets. To address this issue, we use the synthetic speech to create the parallel data. There exist speech data of a lot of speakers with transcriptions available, for instance, speech data used in acoustic model training for speech recognition. Because the speaker of the TTS system is used as the single source speaker in the proposed framework, it is straightforward to develop the parallel data by generating the single source speaker's voices corresponding to individual existing speakers' voices from their transcriptions.

### 3.2. Prosodic Conversion with Language-Dependent Probability Distribution Functions (PDFs)

In the proposed method, the prosodic parameters are globally converted as follows:

$$\hat{p}^{(y)} = \frac{\sigma^{(y)}}{\sigma^{(x)}}\left(p^{(x)} - \mu^{(x)}\right) + \mu^{(y)}, \quad (6)$$

where $p^{(x)}$ and $\hat{p}^{(y)}$ are a prosodic parameter of the source speaker (*i.e.*, the TTS output speaker) and that converted to the target speaker (*i.e.*, the input speaker of the translation system), respectively. Parameters of this conversion function include mean values of the prosodic parameters for the source and target speakers, $\mu^{(x)}$ and $\mu^{(y)}$, and standard deviation values for those speakers, $\sigma^{(x)}$ and $\sigma^{(y)}$. The parameters for the source speaker, $\mu^{(x)}$ and $\sigma^{(x)}$, are easily extracted in advance using a large number of synthetic voices from the TTS system or speech data used in voice building of the TTS system. The parameters for the target speaker, $\mu^{(y)}$ and $\sigma^{(y)}$, are extracted from utterances input to the translation system.

Although the above process assumes that the parameters for the target speaker are the same even if a language is different, this assumption does not always hold. Namely, the prosodic parameters would depend on not only individual speakers but also individual languages: *e.g.*, the standard deviation value of $F_0$ would be larger in a tonal language such as Chinese or Japanese than that in a non-tonal language such as English.

To consider the effect of each language on the prosodic parameters, we propose a prosodic conversion method based on language-dependent PDFs of those parameters. A speech corpus including a lot of speakers in the input language and that in the output language are separately used to extract language-dependent features of the prosodic parameters. First, the mean and standard deviation values, $\mu^{(y)}$ and $\sigma^{(y)}$, are calculated speaker by speaker. Then, the PDF of each parameter for each language is drawn using the calculated parameter values of all speakers in the same language as follows:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(x')dx', \quad (7)$$

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(y')dy', \quad (8)$$

where $x$ and $X$ show a speaker-dependent parameter value ($\mu^{(y)}$ or $\sigma^{(y)}$) and its random variable in the input language, respectively, and $y$ and $Y$ show those in the output language, respectively. The p.d.f.s are given by $f_X(x)$ for the input language and $f_Y(y)$ for the output language. In conversion, first we extract the mean and standard deviation values of each prosodic parameter from the given input speaker's voice. Under an assumption that the following equation holds,

$$P(Y \leq y) = P(X \leq x), \quad (9)$$

those values for the input language are converted to those for the output language as follows:

$$\hat{y} = F_Y^{(-1)}(F_X(x)). \quad (10)$$

Finally, the prosodic parameters generated from the TTS system are globally converted using the conversion function by Eq. (6) with the parameter values converted in Eq. (10) as $\mu^{(y)}$ and $\sigma^{(y)}$. In this paper, we use log-scaled $F_0$ as the prosodic parameter and its mean and standard deviation values are converted using language-dependent PDFs.[1]

In the proposed method, we need to use speech data including a lot of speakers in each language but we don't have to use bilingual data. It is easy to find those speech data available rather than to develop bilingual data. However, the resulting PDF is strongly affected by the number of available samples (*i.e.*, the number of speakers) in each language. To alleviate the overfitting effect, the p.d.f. in each language is modeled by the following constrained GMM,

$$f_X(x) = \sum_{m=1}^{M} \frac{1}{M}\mathcal{N}\left(x; \mu_m - \mu^{(X)}, \sigma_m^2\right), \quad (11)$$

where $M$ is the number of mixture components, $\mu_m$ and $\sigma_m$ are mean and standard deviation values of the $m^{\text{th}}$ mixture component, respectively, which are tied over different languages, and $\mu^{(X)}$ is a language-dependent bias tied over different mixture components. Using this GMM for modeling the p.d.f. in each language, the conversion process by Eq. (10) is simplified as

$$\hat{y} = x - \mu^{(X)} + \mu^{(Y)}, \quad (12)$$

where $\mu^{(X)}$ and $\mu^{(Y)}$ are bias terms for the input language and the output language.

## 4. Experimental Evaluations

### 4.1. Experimental Conditions

Experimental evaluations on cross-lingual speech synthesis were conducted assuming the speech-to-speech translation among Japanese, English, and Chinese. One female speaker was used in each language as the output speaker of each TTS system. In training of one-to-many EV-GMM of spectral parameters for each language, 100 speakers (50 male and 50 female) were used as the pre-defined target speakers. The number of mixture components and the number of representative vectors of each EV-GMM were set to 128 and 99, respectively. In training of PDFs of prosodic parameters for each language, 326 speakers (163 male and 163 female) in Japanese, 200 speakers (100 male and 100 female) in English, and 540 speakers (270 male and 270 female) in Chinese were used. To minimize the effect of different speaking styles on the prosodic parameters, these speakers were selected from speech corpora of travel conversation. In conversion, 4 speakers (2 male and 2 female) in each language were used as the input speaker (*i.e.*, the target speaker to be adapted) not included in the training data. Only 2 sentences for each speaker were used in adaptation and 40 sentences for each speaker were used in evaluation.

As a spectral parameter, the $0^{\text{th}}$ through $24^{\text{th}}$ mel-cepstral coefficients were used. As a prosodic parameter, log-scaled $F_0$ was used in the global conversion and its mean and standard deviation values were used as parameters converted with

[1] We also tried converting a duration parameter but we did not find any significant improvements in naturalness of the converted speech.

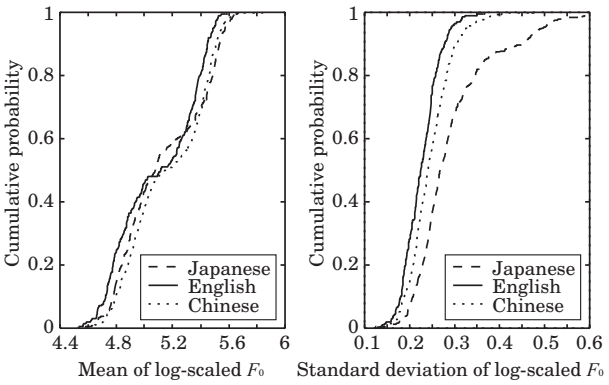


Figure 2: *Language-dependent PDFs of prosodic parameters.*

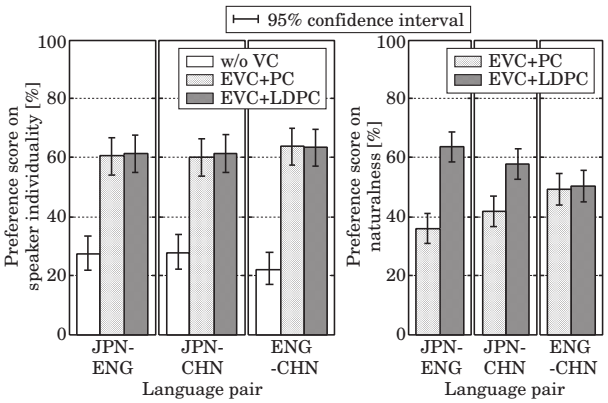


Figure 3: *Results of subjective evaluations.*

language-dependent PDFs. STRAIGHT [13] was used as a speech analysis/synthesis method. The shift length was 5 ms.

Preference tests (XAB tests) of conversion accuracy for speaker individuality and naturalness were conducted separately. In the preference test of conversion accuracy for speaker individuality, 1) the output voice without VC (w/o VC), 2) the output voice converted with one-to-many EVC and global prosodic conversion without considering language-dependent differences (EVC+PC), and 3) the output voice converted with one-to-many EVC and global prosodic conversion using language-dependent PDFs (EVC+LDPC) were compared with each other. In the preference test of naturalness, the latter two methods (EVC+PC and EVC+LDPC) were compared with each other. After vocoded speech of the input speaker (in the input language) was presented as a reference, a pair of the output voices (in the output language) by different two methods was presented to listeners. In the first preference test, the listeners evaluated which voice sounded more similar to the reference in terms of speaker individuality. In the other preference test, the listeners evaluated which voice sounded more natural as the output language voice. These tests were performed separately for each output language by the listeners whose native languages were the same as the output language. The number of listeners for each language was 10.

### 4.2. Experimental Results

The PDF of each parameter is shown in **Figure 2**. It can be observed that the PDFs of the $F_0$ mean value are similar to each other among different languages but the PDFs of the $F_0$ standard deviation values are quite different especially between Japanese and the other two languages. In the preference tests, these PDFs were modeled by the constrained GMMs. The number of mixture components was set to 2 for the $F_0$ mean value and set to 1 for the $F_0$ standard deviation value.

**Figure 3** shows preference scores on conversion accuracy for speaker individuality and those on the naturalness. The one-to-many EVC effectively generates synthetic speech of which voice quality is similar to the input speaker over all language pairs. Furthermore, the language-dependent prosodic conversion yields significant improvements in naturalness of the converted speech in the language pairs of which PDFs of $F_0$ standard deviations are quite different from each other (*i.e.*, Japanese-English and Japanese-Chinese).

## 5. Conclusions

In this paper, we have proposed novel voice conversion techniques to control voice quality of translated speech in speech-to-speech translation. In the proposed techniques, spectral parameters are converted with one-to-many eigenvoice conversion and prosodic parameters are globally converted considering differences of their probability distribution functions between different languages. Experimental results have demonstrated that the proposed techniques are effective for developing speaker-adaptive speech synthesis in speech-to-speech translation.

**Acknowledgment:** This research was supported in part by MEXT Grant-in-Aid for Young Scientists (A) and MIC SCOPE.

## 6. References

[1] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto. The ATR multilingual speech-to-speech translation system. *IEEE Trans. ASLP*, vol.14, no.2, pp.365–376, 2006.

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

[3] S. King, K. Tokuda, H. Zen, and J. Yamagishi. Unsupervised adaptation for HMM-based speech synthesis, *Proc. INTERSPEECH*, pp.1869–1872, Brisbane, Australia, 2008.

[4] Y.-N. Chen, Y. Jiao, Y. Qian, and F.K. Soong. State mapping for cross-language speaker adaptation in TTS. *Proc. of ICASSP*, pp.4273–4276, 2009.

[5] M. Gibson and W. Byrne. Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction. *IEEE Trans. ASLP*, vol.19, no.4, pp.895–904, 2011.

[6] Y. Stylianou, O. Cappé, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. SAP*, vol.6, no.2, pp.131–142, 1998.

[7] T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. ASLP*, vol.15, no.8, pp.2222–2235, 2007.

[8] M. Abe, K. Shikano, and H. Kuwabara. Statistical analysis of bilingual speaker's speech for cross-language voice conversion. *J. Acoust. Soc. Am.*, vol.90, no.1, pp.76–82, 1991.

[9] M. Mashimo, T. Toda, H. Kawanami. K. Shikano, and N. Campbell. Cross-language voice conversion evaluation using bilingual databases. *IPSJ Journal*, vol.43, no.7, pp.2177–2185, July 2002.

[10] D. Erro, A. Moreno, and A. Bonafonte. INCA algorithm for training voice conversion systems from nonparallel corpora. *IEEE Trans. ASLP*, vol.18, no.5, pp.944–953, 2010.

[11] T. Toda, Y. Ohtani, and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. *Proc. ICASSP*, pp.1249–1252, Hawaii, USA, Apr. 2007.

[12] M. Charlier, Y. Ohtani, T. Toda, A. Moinet, and T. Dutoit. Cross-language voice conversion based on eigenvoices. *Proc. INTERSPEECH*, pp.1635–1638, Brighton, UK, Sep. 2009.

[13] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based $F_0$ extraction: possible role of a repetitive structure in sounds. *Speech Communication*, vol.27, no.3–4, pp.187–207, 1999.