INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Phoneme-Discriminative Features for Dysarthric Speech Conversion

Ryo AIHARA[1]*, Tetsuya TAKIGUCHI, Yasuo ARIKI*

Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, Japan
aihara@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

We present in this paper a Voice Conversion (VC) method for a person with dysarthria resulting from athetoid cerebral palsy. VC is being widely researched in the field of speech processing because of increased interest in using such processing in applications such as personalized Text-To-Speech systems. A Gaussian Mixture Model (GMM)-based VC method has been widely researched and Partial Least Square (PLS)-based VC has been proposed to prevent the over-fitting problems associated with the GMM-based VC method. In this paper, we present phoneme-discriminative features, which are associated with PLS-based VC. Conventional VC methods do not consider the phonetic structure of spectral features although phonetic structures are important for speech analysis. Especially for dysarthric speech, their phonetic structures are difficult to discriminate and discriminative learning will improve the conversion accuracy. This paper employs discriminative manifold learning. Spectral features are projected into a subspace in which a near point with the same phoneme label is close to another and a near point with a different phoneme label is apart. Our proposed method was evaluated on dysarthric speaker conversion task which converts dysarthric voice into non-dysarthric speech.

Index Terms: Voice Conversion, Speech Synthesis, Partial Least Square, Assistive Technology, Manifold Learning

1. Introduction

Voice Conversion (VC) is a technique for converting specific information in speech while maintaining the other information in the utterance. One of the most popular VC applications is speaker conversion [1]. In speaker conversion, a source speaker's voice individuality is changed to a specified target speaker's so that the input utterance sounds as though a specified target speaker had spoken it. VC is also being used for Text-To-Speech (TTS) systems [2], spectrum restoring [3], bandwidth extension for audio [4] and more.

Assistive technology is one of the most important task of VC. Nakamura *et al.* [5] proposed GMM-based VC systems that reconstruct a speaker's individuality in electrolaryngeal speech and speech recorded by NAM microphones. This paper proposes a VC method for dysarthric speech resulting from the athetoid type of cerebral palsy. Cerebral palsy is a non-progressive disorder of movement, and most people with cerebral palsy are born with the athetoid type. We convert dysarthric speech into non-dysarthric speech without text information of input speech. Athetoid symptoms also restrict the movement of their arms and legs. Most people suffering from athetoid cerebral palsy cannot communicate by sign language or writing, so there is great need for voice systems for them.

[1]Present affiliation is Information Technology R&D Center, Mitsubishi Electric Corporation.

Rudzicz et al. [6] proposed speech adjustment method for dysarthric speech based on the observations from the database. In [7], we have proposed individuality-preserving VC system for dysarthric speech but this method only effective for the limited utterances.

There are many approaches for VC. Recent advances in deep learning for automatic speech recognition have introduced VC approaches using deep neural networks (DNN) [8, 9]. A non-statistical approach using non-negative matrix factorization (NMF) has also been attracting interest [10, 11, 12]. However, statistical approaches are still being widely researched because of their flexibility and good performance. Among these approaches, the Gaussian Mixture Model (GMM)-based mapping approach [1] is widely used. In this approach, the conversion function is interpreted as the expectation value of the target spectral envelope. The conversion parameters are evaluated using Minimum Mean-Square Error (MMSE) on a parallel training set. A number of improvements in this approach have been proposed. Toda *et al.* [13] introduced maximum-likelihood conversion and the Global Variance (GV) of the converted spectra over a time sequence. However, over-smoothing and over-fitting problems have been reported [14] in regard to these GMM-based approaches because of statistical averages and the large number of parameters. These problems degrade the quality of synthesized speech.

Helander *et al.* [14] proposed transforms based on Partial Least Squares (PLS), in order to prevent the over-fitting problem associated with standard multivariate regression. They also proposed Dynamic Kernel PLS (DKPLS)-based VC [15]. In DKPLS-based VC, source spectral features are projected to high-dimensional feature space by using kernel transformation and the transformed source features are regressed with target spectral features. Their approaches are evaluated with a small-number of parallel training data and they outperformed GMM-based VC However, it has not been evaluated on a situation involving the standard setting of parallel training data.

In this paper, we propose a method that utilizes phoneme-discriminative features and their adaptation to PLS-based VC. In [15], spectral features were projected to a high-dimensional feature space; however, the phonetic structures of spectral features were not considered. Especially for dysarthric speaker conversion, which converts dysarthric speech to non-dysarthric speech, it is important to estimate phoneme-discriminative features because the phonetic structure of dysarthric speech fluctuates. We employed Locality Sensitive Discriminant Analysis (LSDA) [16] to estimate the phoneme-discriminative features. LSDA make it possible to discover the local geometrical structure of the data manifold thus dealing with a major disadvantage of Linear Discriminant Analysis (LDA). In this paper, spectral features are transformed into phoneme-discriminative features by using kernel transformation and LSDA. The within-class-graph matrix and the between-class-graph matrix are calculated based on the phoneme label of the training data to

http://dx.doi.org/10.21437/Interspeech.2017-664

model the local geometrical structure of the underlying manifold. The feature transformation matrix is estimated using the within-class and between-class graph Laplacians, where kernel transformation enables non-liner transformation and more effective learning. Phoneme-discriminative features are converted using PLS. Unlike GMM-based VC, PLS-based VC prevents the over-fitting problem, which might be caused by locality-sensitive learning.

Our proposed method was evaluated on both a non-dysarthric speaker conversion task and a dysarthric speaker conversion task. Objective and subjective evaluation were conducted and the evaluation reveals that our proposed method effectively improved the conversion quality of DKPLS-based VC in dysarthric speaker conversion task.

The rest of this paper is organized as follows: In Section 2, the summary of our algorithm is described. In Section 3, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. Phoneme-discriminative Feature Extraction and Dynamic PLS

2.1. Overview

Our proposed method consists of the training phase and the test phase. Fig. 1 shows the overview of the training phase of our proposed method. First, STRAIGHT analysis [17] is applied to the source and the target parallel utterances (training data), and then STRAIGHT spectra are extracted. Next, spectral features, which include mel-cepstra and delta features, are calculated from the STRAIGHT spectra. Then, k-means clustering is applied to the source spectral features and the estimated centroids are used as reference vectors in kernel transformation. Nest, LSDA is adopted to the kernel-transformed source spectral features, and feature transformation matrices are estimated. The estimated feature transformation matrices are adopted to kernel-transformed source spectral features. In order to consider the dynamic information, adjacent frames are concatenated as segmental features. The segmental source features and the target spectral features are aligned using DTW. (Alignment information is obtained from the source mel-cepstra and the target mel-cepstra.) Finally, the speaker transformation matrix is estimated by using PLS regression, which is estimated from the phoneme-discriminative source features and the target spectral features.

Fig. 2 shows the overview of the test phase of our proposed method. Mel-cepstra and their delta features are calculated from the STRAIGHT spectra and used as spectral features. The features are transformed using reference vectors and the kernel trick. The feature transformation matrix is adopted to the kernel-transformed feature, and the features are transformed into phoneme discriminative feature. Finally, the segmental features are constructed and the speaker transformation matrix is adopted to them.

2.2. Phoneme-discriminative Features

LSDA [16] is adopted to estimate the phoneme-discriminative feature. Because we want to discover the intrinsic geometry, spectral features are transformed into a high-dimensional feature space using kernel transformation. In this paper, Gaussian kernel is adopted to the source spectral feature $X_{d,i}, d = 1, ..., D, i = 1, ..., I$, and the reference vector $R_{d,n}, n = 1, ..., N$, where I, D, and N denote the number of frames of

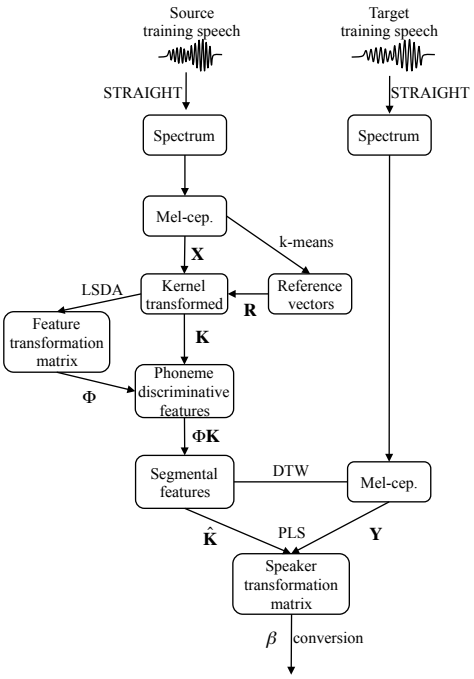

Figure 1: *Overview of the training phase of our proposed method.*

the source training data, the number of dimensions of the source spectral feature, and the number of frames reference vector, respectively. The kernel-transformed feature $k_{i,n}$ is calculated as follows:

$$k_{i,n} = exp(\frac{-||X_{d,i} - R_{d,n}||^2}{2\sigma^2}) \quad (1)$$

σ denotes the width of a Parzen window for the kernel where the selection of is not highly crucial [15]. Spectral features are transformed into a non-linear and a high-dimensional feature space.

In order to estimate the phonetic structure of kernel-transformed features, we calculate a within-class scatter graph and a between-class scatter graph. Adjacency matrices of a within-class scatter graph and a between-class scatter graph of training data are defined as follows:

$$\mathbf{A}^w_{ij} = \begin{cases} 1 & \left(\begin{array}{c} \mathbf{k}_i \in N_{k_w}(\mathbf{k}_i) \; or \; \mathbf{k}_j \in N_{k_w}(\mathbf{k}_j) \\ and \\ c_i = c_j \end{array} \right) \\ 0 & (otherwise) \end{cases} \quad (2)$$

$$\mathbf{A}^b_{ij} = \begin{cases} 1 & \left(\begin{array}{c} \mathbf{k}_i \in N_{k_b}(\mathbf{k}_i) \; or \; \mathbf{k}_j \in N_{k_b}(\mathbf{k}_j) \\ and \\ c_i \neq c_j \end{array} \right) \\ 0 & (otherwise) \end{cases} \quad (3)$$

where $N_{k_w}(\mathbf{k}_i)$ and $N_{k_b}(\mathbf{k}_i)$ denote the set of k_w nearest neighbors of $\mathbf{k}_i$ in the within-class scatter graph and k_b nearest neighbors of $\mathbf{k}_i$ in the between-class scatter graph. c_i and c_j denote the phoneme label of $\mathbf{k}_i$ and $\mathbf{k}_j$. Using adjacency matrices, graph Laplacians of between-class scatter are defined as follows:

$$\mathbf{L}^b = \mathbf{D}^b - \mathbf{A}^b \quad (4)$$

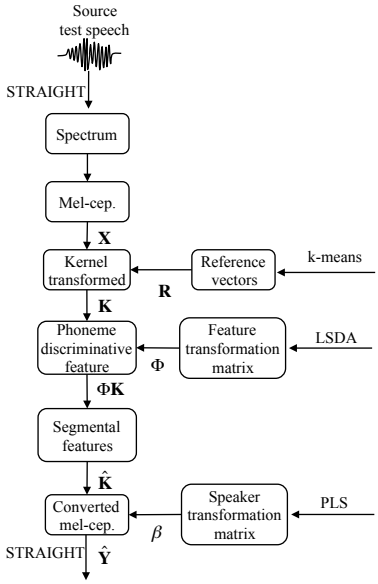

Figure 2: *Overview of the test phase of our proposed method.*

where $\mathbf{D}^b$ denotes the diagonal column (or row) sum of $\mathbf{A}^b$.

Based on LSDA, a phoneme-discriminative feature transformation matrix $\mathbf{\Phi}$ is estimated from the following optimization problem:

$$\begin{aligned} arg\ max\ & \mathbf{\Phi}^\mathsf{T}\mathbf{K}(\alpha\mathbf{L}^b + (1-\alpha)\mathbf{A}^w)\mathbf{K}^\mathsf{T}\mathbf{\Phi} \\ s.t.\ & \mathbf{\Phi}^\mathsf{T}\mathbf{K}\mathbf{D}^w\mathbf{K}^\mathsf{T}\mathbf{\Phi} = 1 \end{aligned} \tag{5}$$

where $\mathbf{D}^w$ and α ($0 \leq \alpha \leq 1$) denote the diagonal column (or row) sum of $\mathbf{A}^w$ and the weight for between-class graph, respectively. The transformation matrix $\mathbf{\Phi}$ that maximizes (5) is given by the maximum eigenvalue problem:

$$\mathbf{\Phi}^\mathsf{T}\mathbf{K}(\alpha\mathbf{L}^b + (1-\alpha)\mathbf{A}^w)\mathbf{K}^\mathsf{T}\mathbf{\Phi} = \lambda\mathbf{\Phi}^\mathsf{T}\mathbf{K}\mathbf{D}^w\mathbf{K}^\mathsf{T}\mathbf{\Phi} \tag{6}$$

where λ denotes eigenvalues.

2.3. Dynamic Partial Least Square Regressions

Before estimating PLS, segmental features $\hat{\mathbf{K}}$ are constructed from the phonetic discriminative source features $\mathbf{\Phi}\mathbf{K}$.

$$\hat{\mathbf{K}}_t = [\mathbf{\Phi}\mathbf{K}_{t-1}^\mathsf{T}\mathbf{\Phi}\mathbf{K}_t^\mathsf{T}\mathbf{\Phi}\mathbf{K}_{t+1}^\mathsf{T}]^\mathsf{T} \tag{7}$$

DPLS is estimated from the segmental features and the target spectral features.

In PLS, the source segmental vector $\hat{\mathbf{K}}_t$ and the target spectral vector $\mathbf{Y}_t$ are represented by a linear transformation of a speaker-independent latent variable vector $\mathbf{h}_t$ as follows:

$$\hat{\mathbf{K}}_t = \mathbf{Q}\mathbf{h}_t + \mathbf{e}_t^x \tag{8}$$

$$\mathbf{Y}_t = \mathbf{P}\mathbf{h}_t + \mathbf{e}_t^y \tag{9}$$

where $\mathbf{Q}$ and $\mathbf{P}$ denote the speaker specific matrix. $\mathbf{e}_t^x$ and $\mathbf{e}_t^y$ denote residual terms. Solving $\mathbf{Q}$ and $\mathbf{P}$, the speaker-transformation matrix β is estimated based-on SIMPLS algorithm [18].

In the test phase, segmental features $\hat{\mathbf{K}}$ are constructed from the phonetic discriminative source features of test data. The converted spectral features $\hat{\mathbf{y}}$ is obtained as follows:

$$\hat{\mathbf{Y}}_t = \beta\hat{\mathbf{K}}_t \tag{10}$$

3. Experiments

3.1. Experimental Conditions

The proposed method was evaluated in a non-dysathric speaker conversion task and dysarthric speaker conversion task using clean speech data. In non-dysathric speaker conversion, two males and two females were used from the ATR Japanese speech database [19] and male-to-male conversion (M101→M102), female-to-female conversion (F101→F102), male-to-female conversion (M101→F101), and female-to-male (F101→M101) conversion were conducted. Fifty parallel sentences are used for training and the other 50 sentences were used for testing.

For dysarthric speaker conversion, one Japanese male with dysarthric speech resulting from athetoid cerebral palsy was stored as the source speaker. The target male speaker is chosen from the ATR Japanese speech database. Fifty parallel sentences are used for training and the other 50 sentences are used for testing.

The sampling rate was 16 kHz. Each sample was analyzed by STRAIGHT [17], and F0, spectral envelope, aperiodic components were extracted. Mel-cepstral features, which are used as spectral features, were calculated from the STRAIGHT spectral envelope, and Δ features were added to them. For non-dysarthric speaker conversion, the energy of mel-cepstrum were not used, and the number of spectral features is 48. For dysarthric speaker conversion, the energy of the mel-cepstrum was used, and the number of spectral features was 50.

We compared the following methods for spectral conversion.

- **ML-GMM-D**: A joint-density GMM with diagonal covariance matrices is modeled on spectral features and converted by trajectory estimation [13].
- **ML-GMM-F**: A joint-density GMM with full covariance matrices is modeled on spectral features and converted by trajectory estimation [13].
- **DKPLS**: PLS is modeled on kernel-transformed spectral features [15].
- **PDKPLS** (proposed method): PLS is modeled on phoneme-discriminative features.

For the GMM-based method, the number of Gaussians was chosen from the set 1, 2, 4, 8, 16, 32, 64, 128, 256. For PLS-based methods, the number of latent components was chosen from the same set. For our proposed method, α was chosen from the set 0.1, 0.5, 1.0.

F0 information is converted using a conventional linear regression based on the mean and standard deviation [13]. Aperiodic components were synthesized without any conversion.

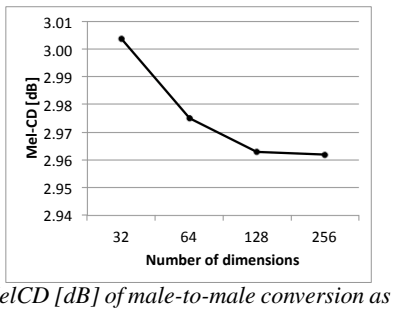

Figure 3: *MelCD [dB] of male-to-male conversion as a function of number of dimensions of phoneme discriminative feature.*

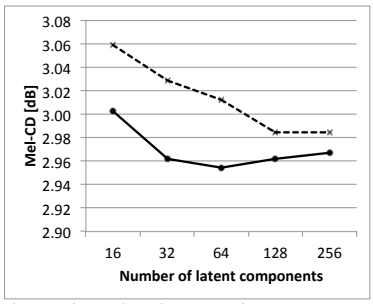

Figure 4: *MelCD [dB] of male-to-male conversion as a function of latent components of PLS. Dashed line denotes DKPLS and solid line denotes prosed PDKPLS.*

3.2. Objective Evaluations

Objective tests were carried out using Mel-cepstrum distortion (MelCD) [dB] as follows:

$$MelCD = (10/\log 10)\sqrt{2\sum_{d}^{24}(mc_d^{conv} - mc_d^{tar})^2} \quad (11)$$

where mc_d^{conv} and mc_d^{tar} denote the d-th dimension of the converted and target mel-cepstra.

Fig. 3 shows the MelCD as a function of the number of dimensions of the proposed phoneme-discriminative feature. The figure shows that using too small of a number of dimension leads to worse results.

Fig. 4 shows the MelCD as a function of the number of latent components. The figure shows that our proposed PDKPLS works well when the number of the latent components is small. This result shows that our proposed phoneme-discriminative feature effectively represents the phonetic feature space.

Table 1 shows the MelCD of non-dysarthric speech conversion. The bold result shows the best result. Considering the results depicted in the table, the difference between ML-GMM-D and ML-GMM-F is not significant. DKPLS obtained a better result than the GMM-based method in female-to-female conversion and male-to-female conversion. Our proposed method outperformed DKPLS in all conversion pairs and obtained the best score in female-to-female and female-to-male conversion.

Fig. 5 shows the MelCD of dysathric speaker conversion. The target speaker is chosen from the database, which obtained the smallest MelCD between the source dysarthric speaker. The results show that our proposed method significantly outperformed the other method. Compared to non-dysarthric speech conversion, the effectiveness of our proposed method is significant in dysarthric speech conversion. We assume thit is because the phoneme structure of dysarthric speech is fluctuats more compared to non-dysarthric speech, and LSDA effectively reduce the fluctuation of the phoneme structure.

Table 1: *MelCD of non-dysarthric speech conversion [dB]*

	M-to-M	F-to-F	M-to-F	F-to-M
Source	3.96	3.69	4.17	4.17
ML-GMM-D	2.96	2.84	2.88	2.86
ML-GMM-F	**2.95**	2.84	2.88	**2.82**
DKPLS	2.98	2.84	2.87	2.84
PDKLPS	**2.95**	**2.81**	**2.84**	**2.82**

3.3. Subjective Evaluations

The subjective evaluation was conducted on "speech quality" and "similarity to the target speaker (individuality)" for the task of dysarthric speech conversion. For the subjective evaluation, 25 sentences for each conversion pair were evaluated

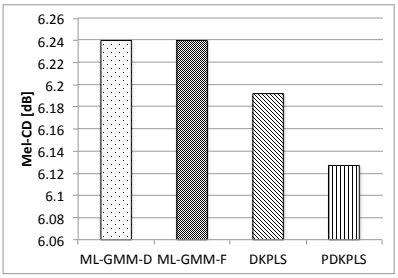

Figure 5: *MelCD [dB] of dysarthric speech conversion.*

by 8 Japanese speakers. For the evaluation of speech quality, we performed a Mean Opinion Score (DMOS) test [20]. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). For the similarity evaluation, a XAB test was carried out. In the XAB test, each subject listened to the voice of the target speaker. Then the subject listened to the voice converted by the two methods and selected which sample sounded most similar to the target speaker's voice.

The left side of Fig. 6 shows the results of a speech-quality test on dysarthric speaker conversion. Our proposed method obtained a better score than ML-GMM-D. The difference between the two methods is significant for the p-value test, $p = 0.03 < 0.05$. The right side of Fig. 6 shows the results of a speaker similarity test on dysarthric speaker conversion. The difference between the two methods is significant for the p-value test, $p = 0.04 < 0.05$. These results did not contradicted to the results of objective evaluation and show the effectiveness of our proposed method.

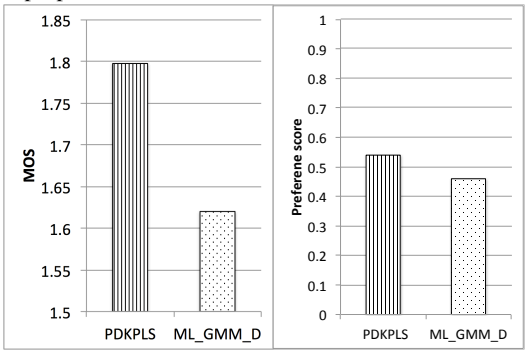

Figure 6: *Results of subjective evaluation for the task of dysarthric speaker conversion. Left: MOS test on speech quality. Right: XAB test on similarity.*

4. Conclusions

We proposed phoneme-discriminative features associated with PLS-based VC. In the case of dysarthric speaker conversion, the phoneme structure of input speech fluctuates compared to non-dysarthric speech, and a discriminative structure is needed. In order to model the local geometric structure of a phoneme spectrum, LSDA was employed using phoneme labels. The phoneme-discriminative feature space is modeled using PLS regressions, which enables us to avoid an over-fitting problem. Experimental results show that our proposed method makes it possible to obtain higher speech quality compared to conventional GMM-based VC, especially for the task of dysarthric speaker conversion.

In this study, there was only one subject person, so in future experiments, we will increase the number of subjects and further examine the effectiveness of our method.

5. Acknowledgements

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2).

6. References

[1] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *in Proc. ICASSP, vol. 1, pp. 285-288*, 1998.

[3] K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "A mel-cepstral analysis technique restoring high frequency components from low-sampling-rate speech," *in Proc. Interspeech*, pp. 2494–2498, 2014.

[4] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine, "GMM-based bandwidth extension using sub-band basis spectrum model," *in Proc. Interspeech*, pp. 2489–2493, 2014.

[5] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[6] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech and Language*, vol. 27, no. 6, pp. 1163–1177, 2014.

[7] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "A preliminary demonstration of exemplar-based voice conversion for articulation disorders using an individuality-preserving dictionary," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014:5, doi:10.1186/1687-4722-2014-5, 2014.

[8] C. Ling-Hui, L. Zhen-Hua, S. Yan, and D. Li-Rong, "Joint spectral distribution modeling using restricted boltzmann machines for voice conversion," *in Proc. Interspeech*, pp. 3052—3056, 2013.

[9] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using rnn pre-trained by recurrent temporal restricted boltzmann machines," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 3, pp. 580–587, 2015.

[10] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," *in Proc. SLT*, pp. 313–317, 2012.

[11] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 10, pp. 1506–1521, 2014.

[12] R. Aihara, T. Takiguchi, and Y. Ariki, "Multiple non-negative matrix factorization for many-to-many voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1175–1184, 2016.

[13] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, 2007.

[14] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech, Lang. Process., vol. 18, Issue:5, pp. 912-921*, 2010.

[15] E. Helander, H. Silén, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 3, pp. 806–817, 2012.

[16] D. Cai, X. He, K. Zhou, J. Han, and H. Bao, "Locality sensitive discriminant analysis," *in Pro. the 20th international joint conference on Artifical intelligence*, pp. 708–713, 2007.

[17] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[18] S. de Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics Intell. Lab. Syst*, vol. 18, no. 3, pp. 251–263, 1993.

[19] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, pp. 357–363, 1990.

[20] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," *ITU-T Recommendation P.800*, 2003.