Noise-Robust Voice Conversion Based on Spectral Mapping on Sparse Space

Ryoichi Takashima, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics, Kobe University, Japan
takashima@me.cs.scitec.kobe-u.ac.jp, aihara@me.cs.scitec.kobe-u.ac.jp
takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

This paper presents a voice conversion (VC) technique for noisy environments based on a sparse representation of speech. In our previous work, we discussed an exemplar-based VC technique for noisy environments. In that report, source exemplars and target exemplars are extracted from the parallel training data, having the same texts uttered by the source and target speakers. The input source signal is represented using the source exemplars and their weights. Then, the converted speech is constructed from the target exemplars and the weights related to the source exemplars. However, this exemplar-based approach needs to hold all training exemplars (frames) and it requires high computation times to obtain the weights of the source exemplars. In this paper, we propose a framework to train the basis matrices of source and target exemplars so that they have a common weight matrix. By using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method. The effectiveness of this method was confirmed by comparing its effectiveness, in speaker conversion experiments using noise-added speech data, with the effectiveness of an exemplar-based method and a conventional Gaussian mixture model (GMM)-based method.

Index Terms: voice conversion, sparse representation, non-negative matrix factorization, noise robustness

1. Introduction

Voice conversion (VC) is generally a technique for changing specific information in an input speech while maintaining the other information in the utterance, such as its linguistic information. One of the most popular applications using the VC technique is speaker conversion, where an utterance spoken by a source speaker is morphed so that it sounds as if it had been spoken by a specified target speaker. There have also been studies on various tasks, such as emotion conversion ([1, 2]), speaking assistance ([3, 4]), and so on, which make use of VC techniques.

Many statistical approaches to VC have been studied ([5, 6, 7]). Among these approaches, the GMM-based mapping approach [7] is widely used, and a number of improvements have been proposed. Toda et al. [8] introduced dynamic features and the global variance (GV) of the converted spectra over a time sequence. Helander et al. [9] proposed transforms based on partial least squares (PLS) in order to prevent the over-fitting problem of standard multivariate regression. There have also been approaches that do not require parallel data that make use of GMM adaptation techniques [10] or eigen-voice GMM (EV-GMM) ([11, 12]).

However, the effectiveness of these approaches was confirmed with clean speech data, and the utilization in noisy environments was not considered. The noise in the input signal is not only output with the converted signal, but may also degrade the conversion performance itself due to unexpected mapping of source features. Hence, a VC technique that takes into consideration the effect of noise is of interest.

Recently, approaches based on sparse representations have gained interest in a broad range of signal processing. In the field of speech processing, non-negative matrix factorization (NMF) [13] is a well-known approach for source separation and speech enhancement ([14, 15]). In these approaches, the observed signal is represented by a linear combination of a small number of atoms, such as the exemplar and basis of NMF. In some approaches for source separation, the atoms are grouped for each source, and the mixed signals are expressed with a sparse representation of these atoms. By using only the weights of the atoms related to the target signal, the target signal can be reconstructed. Gemmeke et al. [16] also proposes an exemplar-based method for noise robust speech recognition. In that method, the observed speech is decomposed into the speech atoms, noise atoms, and their weights. Then the weights of the speech atoms are used as phonetic scores instead of the likelihoods of hidden Markov models for speech recognition.

In our previous work [17], we discussed an exemplar-based VC technique for noisy environments. In that report, source exemplars and target exemplars are extracted from the parallel training data, having the same texts uttered by the source and target speakers. Also, the noise exemplars are extracted from the before- and after-utterance sections in an observed signal. For this reason, no training processes related to noise signals are required. The input source signal is expressed with a sparse representation of the source exemplars and noise exemplars. Only the weights related to the source exemplars are picked up, and the target signal is constructed from the target exemplars and the picked-up weights. This method showed better performances than the conventional GMM-based method in speaker conversion experiments using noise-added speech data. However, this exemplar-based approach needs to hold all training exemplars (frames) and it requires high computation times to obtain the weights of the source exemplars.

In this paper, we propose a framework to train the basis matrices of source and target exemplars so that they have a common weight matrix. The basis matrix of the source exemplars is trained using NMF, and then the weight matrix of the source exemplars is obtained. Next, the basis matrix of the target exemplars is trained using NMF, where the weight matrix is fixed to that obtained from the source exemplars. By using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method. The effectiveness of this method was confirmed by comparing its effectiveness, in speaker conversion experiments using clean speech data and noise-added speech data, with the effectiveness of an exemplar-based method and the conventional Gaussian mixture model (GMM)-based method.

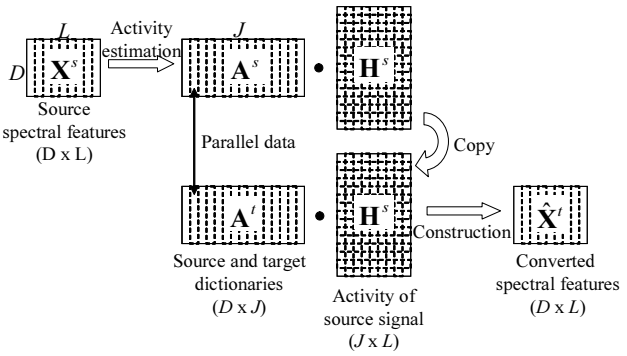

Figure 1: Voice conversion based on the sparse representation

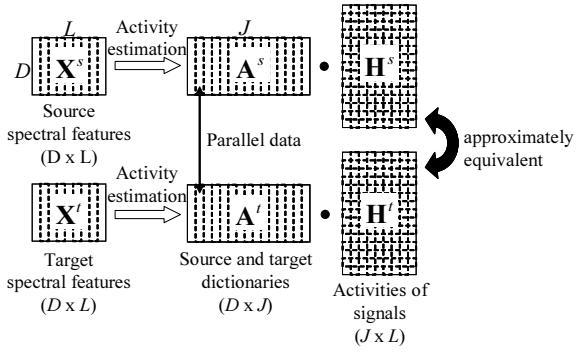

Figure 2: Assumption of the parallelism of source and target dictionaries

2. Voice Conversion Based on Sparse Representation

This section describes a VC method based on the sparse representation [17]. In the approaches based on sparse representations, the observed signal is represented by a linear combination of a small number of atoms.

$$\mathbf{x}_l \approx \sum_{j=1}^{J} \mathbf{a}_j h_{j,l} = \mathbf{A}\mathbf{h}_l \quad (1)$$

$\mathbf{x}_l$ is the l-th frame of the observation. $\mathbf{a}_j$ and $h_{j,l}$ are the j-th atom and the weight, respectively. $\mathbf{A} = [\mathbf{a}_1 \dots \mathbf{a}_J]$ and $\mathbf{h}_l = [h_{1,l} \dots h_{J,l}]^T$ are the collection of the atoms and the stack of weights. When the weight vector $\mathbf{h}_l$ is sparse, the observed signal can be represented by a linear combination of a small number of atoms that have non-zero weights. In this paper, the collection of atoms $\mathbf{A}$ and the weight vector $\mathbf{h}_l$ are called '*dictionary*' and '*activity*', respectively. For the frame sequence data $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_L]$, Eq. (1) is expressed as the inner product of two matrices.

$$\mathbf{X} \approx \mathbf{A}\mathbf{H} \quad (2)$$

$$\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_L], \quad \mathbf{H} = [\mathbf{h}_1 \dots \mathbf{h}_L] \quad (3)$$

L is the number of the frames.

Figure 1 shows the schema of the VC method based on the sparse representation. D, L, J are the numbers of dimensions, frames and atoms, respectively. In this method, the parallel dictionaries, which consist of source and target dictionaries having the same size, are used to map the source signal to the target one. The parallel dictionaries are structured from the parallel training data, which have the same texts uttered by the source and target speakers, and they are aligned using dynamic programming (DP) matching.

This method assumes that when the source signal and the target signal are expressed with sparse representations of the source dictionary and the target dictionary, respectively, then, the obtained activity matrices are approximately equivalent as shown in Figure 2. Based on this assumption, the activity of the source signal estimated with the source dictionary can be substituted for that of the target signal. Therefore, as shown in Figure 1, the input source signal is represented using the source dictionary and the activity. Then, the converted speech is constructed from the target dictionary and the activity related to the source dictionary.

This VC method can be combined with an NMF-based noise reduction method. Then, the noise dictionary is extracted from the before- and after-utterance sections in an observed signal, and the noise dictionary is concatenated with the source dictionary. The noisy source signal is expressed with a sparse representation of the source dictionary and noise dictionary. Only the weights related to the source dictionary are picked up, and the target signal is constructed from the target dictionary and the picked-up weights.

However, this exemplar-based approach defines the parallel dictionary with the parallel training data themselves. Hence, this method needs to hold all training exemplars (frames) and it requires high computation times to obtain the weights of the source exemplars. In conventional NMF-based noise reduction methods, the dictionary $\mathbf{A}$ is not defined with the training exemplars, but with much fewer bases. These bases are trained using the NMF in advance. However, when the basis matrices of source exemplars and target exemplars are trained using the NMF independently, the parallelism of the source and target dictionaries shown in Figure 2 is lost.

Therefore, in this paper, we propose a framework to train the basis matrices of source and target exemplars so that they have a common weight matrix. By using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method.

3. Proposed Method

3.1. Training of the Parallel Basis Matrices

This section describes the framework to train the basis matrices of source and target exemplars. We optimize the source basis matrix $\mathbf{A}^s$ and target basis matrix $\mathbf{A}^t$ so that when the source signal and target signal are expressed with sparse representations of $\mathbf{A}^s$ and $\mathbf{A}^t$, respectively, the obtained activity matrices are equivalent, as shown in Figure 2.

Table 1 shows the algorithm of the training of the parallel basis matrices. At first, for the training source data (exemplars) $\mathbf{X}^s$, the basis matrix $\mathbf{A}^s$ and the activity matrix $\mathbf{H}^s$ are optimized using the NMF with the sparse constraint [16]. In the framework of the NMF with the sparse constraint, it minimizes the following cost function:

$$d(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}^s) + ||(\lambda \mathbf{1}^{(1\times L)}) .* \mathbf{H}^s||_1 \quad s.t. \quad \mathbf{A}^s,\ \mathbf{H}^s \geq 0. \quad (4)$$

Here, $.*$ and $\mathbf{1}$ are an element-wise multiplication and an all-one vector, respectively. The first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}^s$ and $\mathbf{A}^s\mathbf{H}^s$. The second term is the sparse constraint with the L1-norm regularization term that

Table 1: Algorithm of the training of the parallel basis matrices

Training of source basis matrix $\mathbf{A}^s$
• Set source training exemplars to $\mathbf{X}^s$
• Optimize $\mathbf{A}^s$ and $\mathbf{H}^s$ by Eq. (5) and (6)
Training of target basis matrix $\mathbf{A}^t$
• Set target training exemplars to $\mathbf{X}^t$
• Fix the activity matrix to $\mathbf{H}^s$, and optimize $\mathbf{A}^t$ by Eq. (8)

causes $\mathbf{H}^s$ to be sparse. λ is the weight of the sparse constraint. $\mathbf{A}^s$ and $\mathbf{H}^s$ minimizing (4) are estimated iteratively applying the following update rules:

$$\mathbf{A}^s_{n+1} = \mathbf{A}^s_n . * (\mathbf{H}^s_n(\mathbf{X}^s ./ \mathbf{A}^s_n\mathbf{H}^s_n)^T ./ \mathbf{H}^s_n\mathbf{1}^{(1\times D)})^T \quad (5)$$

$$\mathbf{H}^s_{n+1} = \mathbf{H}^s_n . * (\mathbf{A}^{s^T}_n (\mathbf{X}^s ./ (\mathbf{A}^s_n\mathbf{H}^s_n))) ./ (\mathbf{A}^{s^T}_n \mathbf{1}^{(J\times L)} + \lambda\mathbf{1}^{(1\times L)}) \quad (6)$$

where $./$ and $\mathbf{1}$ are an element-wise division and an all-one matrix, respectively.

Next, using the activity matrix $\mathbf{H}^s$ obtained by Eq. (6), the target basis matrix $\mathbf{A}^t$ of the training target exemplars $\mathbf{X}^t$ is optimized. Then, $\mathbf{A}^t$ is optimized so that the activity matrix is equivalent to $\mathbf{H}^s$, i.e. $\mathbf{A}^t$ is optimized to minimize the following cost function:

$$d(\mathbf{X}^t, \mathbf{A}^t\mathbf{H}^s) \quad s.t. \quad \mathbf{A}^t \geq 0. \quad (7)$$

In this optimization, the activity matrix is fixed to $\mathbf{H}^s$, and only $\mathbf{A}^t$ is updated by the following update rule:

$$\mathbf{A}^t_{n+1} = \mathbf{A}^t_n . * (\mathbf{H}^s(\mathbf{X}^t ./ \mathbf{A}^t_n\mathbf{H}^s)^T ./ \mathbf{H}^s\mathbf{1}^{(1\times D)})^T. \quad (8)$$

3.2. Voice Conversion of Noisy Source Signal

3.2.1. Estimation of Activity from Noisy Source Signal

From the before- and after-utterance sections in the observed (noisy) signal, the exemplars (frames) of the noise are extracted, and the noise dictionary is structured from the noise exemplars for each utterance. For this reason, no training processes related to noise signals are required. In the approach based on the sparse representation, the spectrum of the noisy source signal at frame l is approximately expressed by a non-negative linear combination of the source dictionary, noise dictionary, and their activities.

$$\begin{aligned}\mathbf{x}_l &= \mathbf{x}^s_l + \mathbf{x}^n_l \\ &\approx \sum_{j=1}^{J} \mathbf{a}^s_j h^s_{j,l} + \sum_{k=1}^{K} \mathbf{a}^n_k h^n_{k,l} \\ &= [\mathbf{A}^s\mathbf{A}^n] \begin{bmatrix}\mathbf{h}^s_l \\ \mathbf{h}^n_l\end{bmatrix} \quad s.t. \quad \mathbf{h}^s_l, \mathbf{h}^n_l \geq 0 \\ &= \mathbf{A}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \end{aligned} \quad (9)$$

$\mathbf{x}^s_l$ and $\mathbf{x}^n_l$ are the magnitude spectra of the source signal and the noise, respectively. $\mathbf{A}^s$, $\mathbf{A}^n$, $\mathbf{h}^s_l$ and $\mathbf{h}^n_l$ are the source dictionary (basis matrix) trained by Eq. (5), noise dictionary (exemplars), and their activities at frame l, respectively. Given the spectrogram, (9) can be written as follows:

$$\begin{aligned}\mathbf{X} &\approx [\mathbf{A}^s\mathbf{A}^n] \begin{bmatrix}\mathbf{H}^s \\ \mathbf{H}^n\end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0 \\ &= \mathbf{A}\mathbf{H} \quad s.t. \quad \mathbf{H} \geq 0. \end{aligned} \quad (10)$$

In order to consider only the shape of the spectrum, $\mathbf{X}$, $\mathbf{A}^s$ and $\mathbf{A}^n$ are first normalized for each frame, basis or exemplar so that the sum of the magnitudes over frequency bins equals unity.

$$\begin{aligned}\mathbf{M} &= \mathbf{1}^{(D\times D)}\mathbf{X} \\ \mathbf{X} &\leftarrow \mathbf{X} ./ \mathbf{M} \\ \mathbf{A} &\leftarrow \mathbf{A} ./ (\mathbf{1}^{(D\times D)}\mathbf{A}) \end{aligned} \quad (11)$$

The joint matrix $\mathbf{H}$ is estimated based on NMF with the sparse constraint that minimizes the following cost function:

$$d(\mathbf{X}, \mathbf{A}\mathbf{H}) + ||(\lambda\mathbf{1}^{(1\times L)}) . * \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \quad (12)$$

The weights of the sparsity constraints can be defined for each basis and exemplar by defining $\lambda^T = [\lambda_1 \ldots \lambda_J \ldots \lambda_{J+K}]$. In this paper, the weights for source bases $[\lambda_1 \ldots \lambda_J]$ were set to 0.15, and those for noise exemplars $[\lambda_{J+1} \ldots \lambda_{J+K}]$ were set to 0. $\mathbf{H}$ minimizing (12) is estimated iteratively applying the following update rule:

$$\mathbf{H}_{n+1} = \mathbf{H}_n . * (\mathbf{A}^T(\mathbf{X} ./ (\mathbf{A}\mathbf{H}))) ./ (\mathbf{1}^{((J+K)\times L)} + \lambda\mathbf{1}^{(1\times L)}). \quad (13)$$

3.2.2. Target Speech Construction

From the estimated joint matrix $\mathbf{H}$, the activity of source signal $\mathbf{H}^s$ is extracted, and by using the activity and the target dictionary, the converted spectral features are constructed. Then, the target dictionary is also normalized for each basis in the same way the source dictionary was.

$$\mathbf{A}^t \leftarrow \mathbf{A}^t ./ (\mathbf{1}^{(D\times D)}\mathbf{A}^t) \quad (14)$$

$\mathbf{A}^t$ is the target dictionary (basis matrix) trained by Eq. (8). Next, the normalized target spectral feature is constructed, and the magnitudes of the source signal calculated in (11) are applied to the normalized target spectral feature.

$$\hat{\mathbf{X}}^t = (\mathbf{A}^t\mathbf{H}^s) . * \mathbf{M} \quad (15)$$

In this paper, the input source feature is expressed using the magnitude spectrum calculated by STFT because the magnitude spectrum is compatible with the NMF-based noise reduction. On the other hand, the converted spectral feature is expressed as a STRAIGHT spectrum [18] that is compatible with the speech synthesis. The target speech is synthesized using a STRAIGHT synthesizer. Then, F0 information is converted using a conventional linear regression based on the mean and standard deviation.

4. Experiments

4.1. Experimental Conditions

The proposed VC technique was evaluated by comparing it with an exemplar-based method [17] and a conventional GMM-based method [7] in a speaker conversion task using clean speech data and noise-added speech data. The source speaker and target speaker were one male and one female speaker, whose speech is stored in the ATR Japanese speech database, respectively. The sampling rate was 8 kHz.

Two hundred sixteen words of clean speech were used to construct parallel dictionaries in the methods based on the sparse representation and used to train the GMM in GMM-based method. In the exemplar-based method, the number of

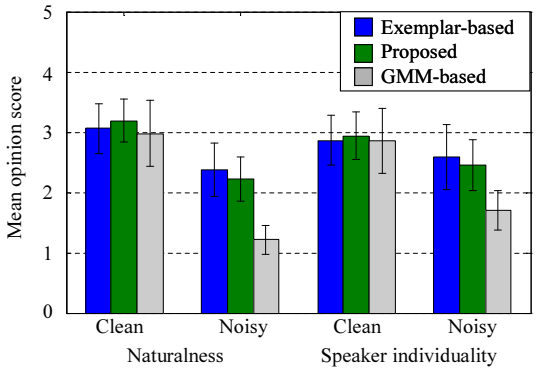

Figure 3: Mean opinion scores (MOS) for each method

exemplars of source and target dictionaries was 58,426. Then, in our proposed method, 1,000 bases were trained from the exemplars for each dictionary. Twenty-five sentences of clean speech or noisy speech were used to evaluate. The noisy speech was created by adding a noise signal recorded in a restaurant (taken from the CENSREC-1-C database) to the clean speech sentences. The SNR was 15 dB. The noise dictionary is extracted from the before- and after-utterance section in the evaluation sentence. The average number of exemplars in the noise dictionary for one sentence was 110.

In the methods based on the sparse representation, a 257-dimensional magnitude spectrum was used as the feature vectors for input signal, source dictionary and noise dictionary, and a 513-dimensional STRAIGHT spectrum was used for the target dictionary. The number of iterations used to estimate the activity was 500. In the GMM-based method, the 1^{st} through 40^{th} linear-cepstral coefficients obtained from the STRAIGHT spectrum were used as the feature vectors. The number of mixtures was 64.

4.2. Experimental Results

We performed an opinion test on the naturalness and speaker individuality of the converted speech. In the opinion test, the opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The tests were carried out with 7 subjects. For the evaluation of naturalness, each subject listened to the converted speech and evaluated how natural the sample sounded. For the evaluation of speaker individuality, each subject listened to the target speech. Then the subject listened to the converted speech and evaluated how similar the converted speech and the target one.

Figure 3 shows the mean opinion scores (MOS) for each method. The error bars show 95% confidence intervals. As shown in this figure, when clean speech data was used, the performances of the three methods were not so different in both evaluation criteria. However, when noisy speech data was used, the performances of GMM-based method degraded considerably especially in naturalness. This might be because the noise caused unexpected mapping in the GMM-based method, and the speech was converted with a lack of naturalness. On the other hand, the degradations of the performances of the VC methods based on the sparse representation were less than those of GMM-based method. The performances of the proposed method were slightly lower than that of the exemplar-based method when noisy speech data was used. However, for obtaining the activity matrix, the computation time of the proposed method (about 30 seconds for 1 sentence on Intel Core i7 2.80 GHz personal computer) was about 30 times faster than that of the exemplar-based method (about 910 seconds).

Table 2: Spectral distortion improvement ratio (SDIR) [dB] for noisy speech

	Exemplar-based	Proposed	GMM-based
SDIR [dB]	3.8	3.7	3.2

Table 2 shows the spectral distortion improvement ratio (SDIR) [dB] for noisy input source signal. The SDIR is defined as follows.

$$\mathrm{SDIR[dB]} = 10\log_{10}\frac{\sum_d |\mathbf{X}^t(d) - \hat{\mathbf{X}}^t(d)|^2}{\sum_d |\mathbf{X}^t(d) - \mathbf{X}^s(d)|^2} \quad (16)$$

Here, $\mathbf{X}^s$, $\mathbf{X}^t$ and $\hat{\mathbf{X}}^t$ are normalized so that the sum of the magnitudes over frequency bins equals unity. As shown in this table, the distortion improvements of the methods based on the sparse representation were higher than GMM-based method. The distortion improvements of the proposed method was slightly lower than that of the exemplar-based method.

5. Conclusions

In this paper, we discussed a noise-robust VC technique based on sparse representation. We proposed a framework to train the basis matrices of source and target exemplars so that they have a common activity matrix. The basis matrix of the source exemplars is trained using the NMF. Then, the basis matrix of the target exemplars is trained using the NMF, where the weight matrix is fixed to that obtained from the source exemplars. By using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method. When a noisy input signal is converted to the target signal, the noise exemplars are extracted from the before- and after-utterance sections in an observed signal. The noisy signal is expressed with a sparse representation of the source basis matrix and noise exemplars. The target signal is constructed from the target basis matrix and the activity matrix related to the source basis matrix.

In comparison experiments between the proposed method, an exemplar-based method and a conventional GMM-based method, the proposed method showed better performances than GMM-based method when evaluating noisy speech. The performances of the proposed method were slightly lower than that of the exemplar-based method when noisy speech data was used. But for obtaining the activity matrix, the computation time of the proposed method was about 30 times faster than that of the exemplar-based method.

However, the proposed method still requires higher computation times than that of GMM-based method. While our proposed method took about 30 seconds for 1 sentence to convert speech features, the GMM-based method spent about 1 second to do this. In future work, we will investigate the optimal number of bases and evaluate the performances under other noise conditions. We will also try to introduce dynamic information, such as segment features. In addition, this method has a limitation in that it can be applied to only one-to-one voice conversation because it requires parallel speech data having the same

texts uttered by the source and target speakers. Hence, we will investigate a method that does not use parallel data. Future work will also include efforts to study other noise conditions, such as a low-SNR condition, and apply this method to other VC applications.

6. References

[1] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," in *Proc. INTERSPEECH*, 2003, pp. 2401–2404.

[2] C. Veaux and X. Robet, "Intonation conversion from neutral to expressive speech," in *Proc. INTERSPEECH*, 2011, pp. 2765–2768.

[3] 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.

[4] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models," *IEICE Trans. Information and Systems*, vol. E93-D, no. 9, pp. 2472–2482, 2010.

[5] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988, pp. 655–658.

[6] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Communication*, vol. 11, no. 2–3, pp. 175–187, 1992.

[7] 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.

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

[9] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Trans. Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[10] C. H. Lee and C. H. Wu, "Map-based adaptation for speech conversion using adaptation data selection and non-parallel training," in *Proc. INTERSPEECH*, 2006, pp. 2254–2257.

[11] T. Toda, Y. Ohtani, and K. Shikano, "Eigenvoice conversion based on gaussian mixture model," in *Proc. INTERSPEECH*, 2006, pp. 2446–2449.

[12] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose, "One-to-many voice conversion based on tensor representation of speaker space," in *Proc. INTERSPEECH*, 2011, pp. 653–656.

[13] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Neural Information Processing System*, 2001, pp. 556–562.

[14] T. Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[15] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. INTERSPEECH*, 2006, pp. 2614–2617.

[16] J. F. Gemmeke, T. Viratnen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

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

[18] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.