Voice Conversion Based on Non-negative Matrix Factorization and Its Application to Practical Tasks

Aihara, Ryo

(Degree)
博士（工学）

(Date of Degree)
2017-03-25

(Date of Publication)
2018-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第6935号

(URL)
https://hdl.handle.net/20.500.14094/D1006935

Doctoral Thesis

Voice Conversion Based on Non-negative Matrix Factorization and Its Application to Practical Tasks

非負値行列因子分解に基づく声質変換とその実用的課題への応用

Ryo AIHARA

Graduate School of System Informatics

Kobe University

A thesis submitted for PhD. degree

Jan. 2017

Doctoral Thesis

Voice Conversion Based on Non-negative Matrix Factorization and Its Application to Practical Tasks

Ryo AIHARA

The voice is one of the most natural communication tools for human beings and it provides not only linguistic information but also paralinguistic information, which include speaker information and emotion. Voice conversion (VC) is a technique for converting paralinguistic information in speech, while preserving the linguistic information in the utterance. The most popular VC application is speaker conversion however, VC is also being used for various tasks such as emotion conversion and assistive technology.

The Gaussian mixture model (GMM) is widely used for VC because of its flexibility and good performance, and a number of improvements have been proposed. 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) or maximum likelihood (ML) methods on a training set. In GMM-based approaches, low-dimensional spectral features (for example mel-cepstrum) are used in order to avoid the "curse of dimensionality".

Exemplar-based VC, which is a non-statistical approach, has attracted interest. For example, an exemplar-based VC using non-negative matrix factorization (NMF), in which linear transformation has been proposed. In this method, source and target exemplars are extracted from a parallel training set, in which the same texts are spoken by the source and target speakers. The input source signal is expressed with sparse representation of the source exemplars using NMF. By replacing the source speaker exemplar with that of the target speaker, the original speech spectrum is replaced with that of the target speaker.

We assume that the NMF-based VC method has an advantage over conventional VC methods as our approach results in a natural-sounding converted voice. Over-smoothing and over-fitting problems have been reported in statistical approaches because of statistical averaging and the large number of parameters involved. Because our approach is non-statistical one, it should avoid the over-fitting problem.

This paper proposes four advanced VC algorithms based on NMF that resolve each practical task. The four methods correspond to the following key words, noise-robustness, assistive technology, small-parallel corpus, and many-to-many VC.

The conventional VC method has been evaluated in a clean environment. Background noise, which is unavoidable factor in real environments, has been ignored in the VC tasks. The NMF-based VC method has noise-robustness and we propose a framework to train the basis matrices of the source and target exemplars so they have a common weight matrix. We call this method as "VC using sparse spectrum mapping". By using the basis matrices instead of the exemplars, the VC performs in less computation time than the conventional exemplar-based method.

Assistive technology is one of the most important tasks of a VC system. This paper focuses on an articulation disorder resulting from athetoid cerebral palsy. Some people with articulation disorders need a VC system that can improve the intelligibility of their disordered voice while preserving their voice intelligibility. In order to preserve the speaker's individuality, we used a combined dictionary that was constructed from the source speaker's vowels and target speaker's consonants. In this paper, in order to reduce the mismatching of phoneme alignment, we propose a phoneme-categorized sub-dictionary and a dictionary selection method using NMF.

The need to have a large amount of parallel data is great barrier to the practical application of VC. This paper presents a novel framework of an exemplar-based VC that only requires a small number of parallel exemplars. An adaptation matrix in an NMF framework is introduced to adapt the source dictionary to the target dictionary. This adaptation matrix is estimated using only a small-parallel speech corpus. We refer to this method as Affine-NMF, and its effectiveness is

confirmed by comparison with those of a conventional NMF-based method and a GMM-based method in noisy environments.

Many-to-many VC is a VC method for arbitrary speakers. Because NMF-based VC requires parallel training data from source and target speakers, the voices of arbitrary speakers cannot be converted in this framework. In this study, we propose the multiple non-negative matrix factorization (Multi-NMF) to allow the implementation of a many-to-many exemplar-based VC.

As explained above, this paper proposes new algorithms for four important practical tasks VC tasks. In addition, we expands the use of VC systems. The tasks are not only related to VC but also to speech recognition and other problems in the field of signal processing.

Contents

List of Figures

List of Tables

Declaration

I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other german or foreign examination board.

The related contents in this thesis have been previously published or submitted for publication by the author. A complete list of publications can be found on *pp.* xv-xxi.

Publication List

Journal Papers

1. Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi and Yasuo Ariki: “Noise-Robust Voice Conversion Based on Sparse Spectral Mapping Using Non-negative Matrix Factorization”, *IEICE Transactions on Information and Systems*, Vol.E97-D, No.6, pp.1411-1418, 2014.

2. Yuki Takashima, Yasuhiro Kakihara, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki, Nobuyuki Mitani, Kiyohiro Omori, Kaoru Nakazono: “Audio-Visual Speech Recognition Using Convolutive Bottleneck Networks for a Person with Severe Hearing Loss”, *IPSJ Transactions on Computer Vision and Applications*, Vol. 7, pp. 64-68, 2015.

3. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “Individuality-Preserving Voice Conversion for Articulation Disorders Using Phoneme-Categorized Exemplars”, *ACM Transactions on Accessible Computing (TACCESS)*, Vol. 6, No. 4, pp. 13:1-13:17, 2015.

4. Masaka Kenta, Aihara Ryo, Takiguchi Tetsuya, Ariki Yasuo: “Multimodal voice conversion based on non-negative matrix factorization”, *EURASIP Journal on Audio, Speech, and Music Processing*, 2015:24 DOI: 10.1186/s13636-015-0067-4m 2015.

5. Ryo Aihara, Takao Fujii, Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki: “Small-parallel exemplar-based voice conversion in noisy environments using affine non-negative matrix factorization”, *EURASIP Journal on Au-*

dio, Speech, and Music Processing, 2015:32 doi:10.1186/s13636-015-0075-4, 2015.

6. Ryo Aihara, Testuya Takiguchi, Yasuo 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.

International Conference Papers

1. Ryo AIHARA, Toru NAKASHIKA, Tetsuya TAKIGUCHI, Yasuo ARIKI: "VOICE CONVERSION BASED ON NON-NEGATIVE MATRIX FACTORIZATION USING PHONEME-CATEGORIZED DICTIONARY", *ICASSP 2014*, pp.7944-7948, 2014.

2. Kenta MASAKA, Ryo AIHARA, Tetsuya TAKIGUCHI, Yasuo ARIKI: "MULTIMODAL VOICE CONVERSION USING NON-NEGATIVE MATRIX FACTORIZATION IN NOISY ENVIRONMENTS", *ICASSP 2014* , pp.1561-1565, 2014.

3. Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki: "Individuality-preserving Voice Conversion for Articulation Disorders Using Dictionary Selective Non-negative Matrix Factorization", *SLPAT 2014, 5th Workshop on Speech and Language Processing for Assistive Technologies*, pp. 29-37, 2014.

4. E. Byambakhishig, K. Tanaka, R. Aihara, T. Nakashika, T. Takiguchi, Y. Ariki: "Error Correction of Automatic Speech Recognition Based on Normalized Web Distance", *Interspeech 2014*, pp.2852-2856, 2014.

5. Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki: "Multimodal Exemplar-based Voice Conversion using Lip Features in Noisy Environments", *Interspeech 2014*, pp.1159-1163, 2014.

6. Ryo AIHARA, Reina UEDA, Tetsuya TAKIGUCHI, Yasuo ARIKI: "Exemplar-based Emotional Voice Conversion Using Non-negative Matrix Factorization", *APSIPA 2014*, 4 pages, 2014.

7. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “ACTIVITY-MAPPING NON-NEGATIVE MATRIX FACTORIZATION FOR EXEMPLAR-BASED VOICE CONVERSION”, *ICASSP 2015*, pp. 4899-4903, 2015.

8. Ryo Aihara, Takao Fujii, Tetsuya Takiguchi, and Yasuo Ariki: “NOISE-ROBUST VOICE CONVERSION USING A SMALL PARALLEL DATA BASED ON NON-NEGATIVE MATRIX FACTORIZATION”, *The 23rd European Signal Processing Conference (EUSIPCO)*, pp.315-319, 2015.

9. Ryo Aihara, Testuya Takiguchi, and Yasuo Ariki: “Many-to-many Voice Conversion Based on Multiple Non-negative Matrix Factorization”, *INTERSPEECH 2015*, pp. 2749-2753, 2015.

10. Ryo AIHARA, Kenta MASAKA, Tetsuya TAKIGUCHI, Yasuo ARIKI: “LIP-TO-SPEECH SYNTHESIS USING LOCALITY-CONSTRAINT NON-NEGATIVE MATRIX FACTORIZATION”, *The First International Workshop on Machine Learning in Spoken Language Processing (MLSLP2015)*, 6 pages, 2015.

11. Reina Ueda, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki : “Individuality-Preserving Spectrum Modification for Articulation Disorders Using Phone Selective Synthesis”’, *SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies*, 6 pages, 2015.

12. Ryo Aihara, Testuya Takiguchi, and Yasuo Ariki: “MANY-TO-ONE VOICE CONVERSION USING EXEMPLAR-BASED SPARSE REPRESENTATION”, *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2015.

13. Ryo AIHARA, Testuya TAKIGUCHI, Yasuo ARIKI: “SEMI-NON-NEGATIVE MATRIX FACTORIZATION USING ALTERNATING DIRECTION METHOD OF MULTIPLIERS FOR VOICE CONVERSION”, *ICASSP 2016*, pp. 5170-5174, 2016.

14. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “Parallel Dictionary Learning for Voice Conversion Using Discriminative Graph-embedded Non-negative Matrix Factorization”, *Interspeech 2016*, pp. 292-296, 2016.

15. Yuki Takashima, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki, Nobuyuki Mitani, Kiyohiro Omori, Kaoru Nakazono: “Audio-Visual Speech Recognition Using Bimodal-Trained Bottleneck Features for a Person with Severe Hearing Loss”’, *Interspeech 2016*, pp.227-281, 2016.

16. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “Dysarthric Speech Modification Using Parallel Utterance Based on Non-negative Temporal Decomposition”, *SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies*, pp. 75-79, 2016.

Book

1. Ryo Aihara, Kenta Masaka, Tetsuya Takiguchi, and Yasuo Ariki: “Parallel Dictionary Learning for Multimodal Voice Conversion Using Matrix Factorization”, *Computer and Information Science*, edited by Roger Lee, Springer International Publishing, pp. 27-40, 2016.

Technical Reports

1. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “Individuality-preserving Voice Conversion for Articulation Disorders Using Sparse Dictionary Learning”, *IEICE Technical Report*, vol. 114, no. 91, SP2014-53, pp. 39-44, 2014. (in Japanese)

2. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “Many-to-one Voice Conversion using Multiple Non-negative Matrix Factorization”, *IEICE Technical Report*, vol. 114, no. 365, SP2014-126, pp. 75-80, 2014. (in Japanese)

3. Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki: “Multimodal Voice Conversion using Weighted Features in Noisy Environments”, *IEICE Technical Report*, vol. 114, no. 365, SP2014-126, pp. 87-92, 2014. (in Japanese)

4. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: "Exemplar-based Voice Conversion for Arbitrary Speakers", *IEICE Technical Report*, vol. 115, no. 253, pp. 1-6, 2015. (in Japanese)

5. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: "Parallel Dictionary Learning for Voice Conversion Using Alternating Direction of Multipliers", *IEICE Technical Report*, vol. 115, no. 346, SP2015-72, pp. 13-18, 2015. (in Japanese)

6. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: "Parallel Dictionary Learning for Voice Conversion Using Discriminative Graph-embedded Non-negative Matrix Factorization", *IEICE Technical Report*, vol. 116, no. 189, SP2016-38, pp. 59-64, 2016. (in Japanese)

Domestic Conference Papers

1. Byambakhishig Enkhbolor, Katsuyuki Tanaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki: "Error correction of automatic speech recognition based on Normalized Web Distance", *The 28th Annual Conference of the Japanese Society for Artificial Intelligence*, 301-1in, 2014. (in Japanese)

2. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: "Exemplar-based Voice Conversion Based on Activity-adaptive Non-negative Matrix Factorization", *Acoustical Society of Japan 2014 Autumn Meeting*, 1-7-16, pp.223-226, 2014. (in Japanese)

3. Takao Fujii, Ryo Aihara, Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki: "Voice Conversion based on NMF using Speaker Adaptation in Noisy Environments", *Acoustical Society of Japan 2014 Autumn Meeting*, 2-Q-36, pp. 345-348, 2014. (in Japanese)

4. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: "Many-to-one Voice Conversion Based on Multiple Non-negative Matrix Factorization", *Acoustical Society of Japan 2015 Spring Meeting*, 3-2-2, pp. 275-278, 2015. (in Japanese)

5. Takao Fujii, Ryo Aihara, Toru Nakashika, Tetsuya Takiguchi, Yasuo Ariki: “Voice Conversion using a Small Parallel Corpus based on Non-negative Matrix Factorization in Noisy Environments”, *Acoustical Society of Japan 2015 Spring Meeting*, 2-Q-39, pp. 393-396, 2015. (in Japanese)

6. Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki: “Speech Production from Lip Images based on Non-negative Matrix Factorization”, *Acoustical Society of Japan 2015 Spring Meeting*, 2-Q-38, pp. 389-392, 2015. (in Japanese)

7. Yuki Takashima, Yasuhiro Kakihara, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki, Nobuyuki Mitani, Kiyohiro Omori, Kaoru Nakazono: “Audio-Visual Speech Recognition Using Convolutive Bottleneck Networks for a Person with Severe Hearing Loss”, *MIRU 2015*, OS3-2, 2015. (in Japanese)

8. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “Many-to-many Voice Conversion Based on Multiple Non-negative Matrix Factorizations”, *Acoustical Society of Japan 2015 Autumn Meeting*, pp. 227-230, 2015. (in Japanese)

9. Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki: “Speech Generation from Lip Images based on Non-negative Matrix Factorization with Beta-divergence”, *Acoustical Society of Japan 2015 Autumn Meeting*, 1-Q-32, pp. 285-288, 2015. (in Japanese)

10. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “Parallel Dictionary Learning for NMF-based Voice Conversion Using Alternating Direction Method of Multipliers”, *Acoustical Society of Japan 2016 Spring Meeting*, 1-R-36, pp.325-328, 2016. (in Japanese)

11. Kenta Masaka, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki: “Multimodal Voice Conversion using Sparse-Parallel Training.”, *Acoustical Society of Japan 2016 Spring Meeting*, 1-R-35, pp. 321-325, 2016. (in Japanese)

12. Konjun I, Ryo Aihara, Tetsuya Takiguchi, Yasuo Ariki: “Voice Conversion using a Small Parallel Corpus based on NMF using ADMM in Noisy Environments”, *Acoustical Society of Japan 2016 Spring Meeting*, 1-R-38, 2016. (in Japanese)

13. Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki: “Parallel Dictionary Learning for Voice Conversion Using Discriminative Graph-embedded Non-negative Matrix Factorization”, *Acoustical Society of Japan 2016 Autumn Meeting*, 3-5-3, pp.155-158, 2016. (in Japanese)

Glossary

ASR	Automatic Speech Recognition
Affine-NMF	Affine Non-negative Matrix Factorization
ANN	Artificial Neural Network
CC	Cepstral Coefficients
DNN	Deep Neural Networks
DP	Dynamic Programing
DTW	Dynamic Time Warping
GMM	Gaussian Mixture Model
KL	Kullback-Leibler
LPC	Linear Prediction Coefficients
MAP	Maximum A Posteriori Probability
ML	Maximum Likelihood
Mel-CD	Mel-Cepstral Distortion
MFCC	Mel-Frequency Cepstral Coefficients
MLSA	Mel Log Spectral Approximation
Multi-NMF	Multiple Non-negative Matrix Factorization
MMSE	Minimum Mean-Square Error
NMF	Non-negative Matrix Factorization
NSD	Normalized Spectrum Distortion
PCA	Principal Component Analysis
STFT	Short-Time Fourier Transform
SDIR	Spectral Distortion Improvement Ratio
TTS	Text-To-Speech
VC	Voice Conversion

Chapter 1

Introduction

1.1 Background

The voice is one of the most natural communication tool for human beings and it provides not only linguistic information but also paralinguistic information, which includes speaker information and emotion. For decades researchers have tried to extract certain information from voice signals. However, it remains a difficult task to separate and extract certain information from voice signals because linguistic and paralinguistic information are closely connected each other.

Voice conversion (VC) is a technique for converting specific information in speech, while preserving the other information in the utterance. In general, speech includes linguistic and paralinguistic information. The most popular VC application is speaker conversion [1, 2, 3], which converts a source speaker's voice individuality to that of a specified target speaker, while preserving the linguistic information. VC has also been applied to emotion conversion [4, 5, 6, 7], which converts the emotional information in the input speech while preserving the speaker and linguistic information. Other uses of VC include assistive technology [8], text-to-speech (TTS) systems [9], spectrum restoring [10], and bandwidth extension for audio [11].

VC is challenging as there is no unique correct answer to its task. Every time a source speaker utters the same sentence, the observed spectrum is different. Furthermore, the perception of conversion quality is subjective. Therefore, listening tests must be used for evaluation of VC systems and these take a great deal of time. Some objective measures such as mel-cepstral distortion (Mel-CD), spectral distortion improvement ratio (SDIR) and normalized spectrum distortion (NSD) are used to complement the subjective evaluation [12].

It is considered that some indications of a speaker's identity are more apparent in paralinguistic information more than in linguistic information. The paralinguistic factors can be categorized into two types: sociological and physiological factors. Sociological factors, including the place of birth, social class, and age of the speaker, mainly affect prosodic features (pitch contour, duration, rhythm, etc.). On the other hand, physiological factors (the shape of the vocal tract) directly affect the spectral information and determine the individuality. It has been reported that the most important acoustic features for the identification of a speaker include third formant, fourth formant, fundamental frequency, and the closing phase of the glottal wave [13].

There are two approaches to VC methods. One is a combined approach of automatic speech recognition (ASR) and TTS. In this approach, an input utterance is recognized using ASR systems and estimated text is synthesized with the target speaker's voice using TTS systems. The other approach is a more direct one, which does not recognize the text information of input utterances. The former approach may be effective because of recent developments in ASR and TTS. However, the linguistic information of the output voice will be incorrect if the ASR system fails to recognize the input utterance. Moreover, a large amount of training data is required to develop ASR and TTS systems for arbitrary speakers. Therefore, this paper adopts the latter approach.

Most existing VC systems use the direct approach to deal with the conversion of spectral features, and this is used here. However, prosodic features, such as fundamental frequency (F_0), can also be seen as important factors in speaker identity. Helander *et al.* showed that when true prosody features are used, we can recognize a person who is familiar to us [14]. Nevertheless, we obtained good results from simple statistical mean and variance scaling methods for F_0 conversion. More advanced F_0 conversion methods can be found in the literature [15, 16, 17].

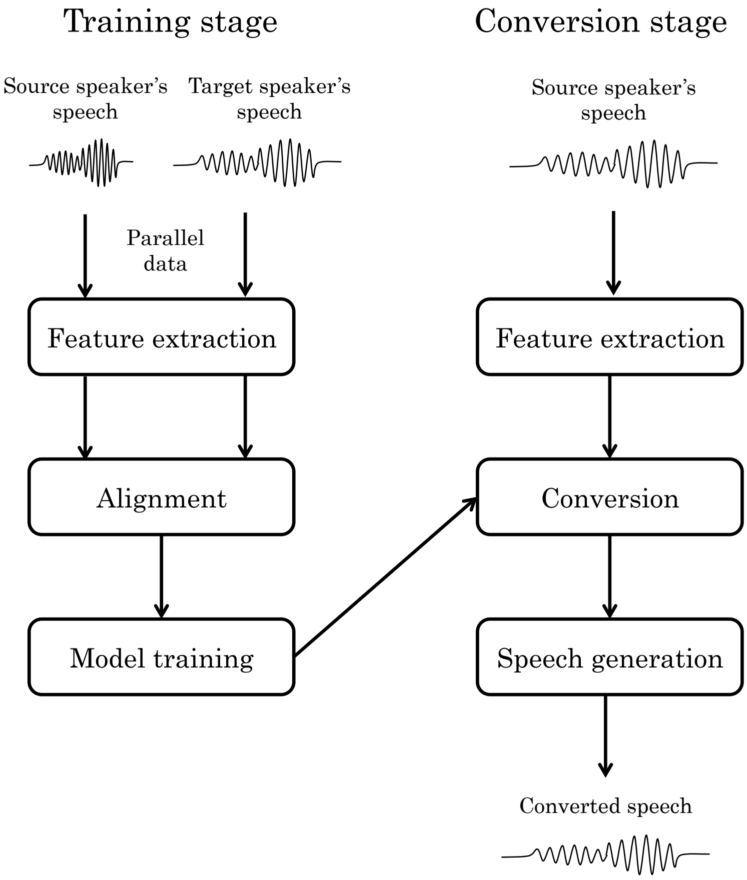


Figure 1.1: Flowchart of a typical VC system

1.2 Approaches

Most VC systems have a system flow as shown in Figure 1.1. The system can broadly be divided into two stages: training and conversion. Both stages begin with feature extraction. In this process, acoustic features, such as mel-frequency cepstral coefficients (MFCC), cepstral coefficients (CC), or linear prediction coefficients (LPC) are extracted from the speech signals. In the training stage, the extracted time-series features are aligned to adjust the time positions of the source speaker's features and target speakers' features (such aligned data is called parallel data). Therefore, typical VC systems require pairs of speech signals to be produced by the source and the target speakers uttering the same sentence. Dynamic time warping (DTW) [18] is often used for the alignment process. Any statistical or non-statistical models is trained using the obtained parallel data. In the conversion stage, the acoustic features are extracted from the source speaker's speech and fed to the model, resulting in acoustic features that are supposed to be those of the target speakers. Finally, the features are back-projected into a speech signal. In this way we obtain converted speech.

Various statistical approaches to VC have been studied over the past few decades and can be divided into two broad categories, those which employ linear transformation and those which employ non-linear transformation. The most widely researched example of the first approach is Gaussian mixture model (GMM)-based VC [3]. 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) or maximum likelihood (ML) methods on a training set [19]. An alternative approach is the artificial neural network (ANN), a classic method adopting a non-linear transformation function. Recent advances in deep learning for ASR have introduced VC approaches using deep neural networks (DNN) [20, 21, 22]. These approaches try to capture the non-linear relationship between the source and target spectra.

In recent years, exemplar-based VC, a non-statistical approach, has attracted interest [23, 24]. These approaches are based on non-negative matrix factorization (NMF) [25] and employs a linear transformation [26]. In this method, source and target exemplars are extracted from a parallel training set, in which the same texts are spoken by the source and target speakers. The input source signal is expressed with a sparse representation of the source exemplars using NMF. By replacing the source speaker exemplar with the that of the target speaker, the original speech spectrum is replaced with that of the target speaker. We assume that the NMF-based VC method has advantages over conventional VC methods. This approach results in a natural-sounding converted voice [27]. Over-smoothing and over-fitting problems have been reported [28] in statistical approaches because of statistical averaging and the large number of parameters involved.

1.3 Purpose of This Thesis

This paper proposes four different VC algorithms for each practical task. All our methods are based on NMF, which provides a natural-sounding converted voice. The four methods correspond to the following key words; noise-robustness, assistive technology, small-parallel corpus, and many-to-many VC. Fig. 1.2 shows the relationships of our proposed method.

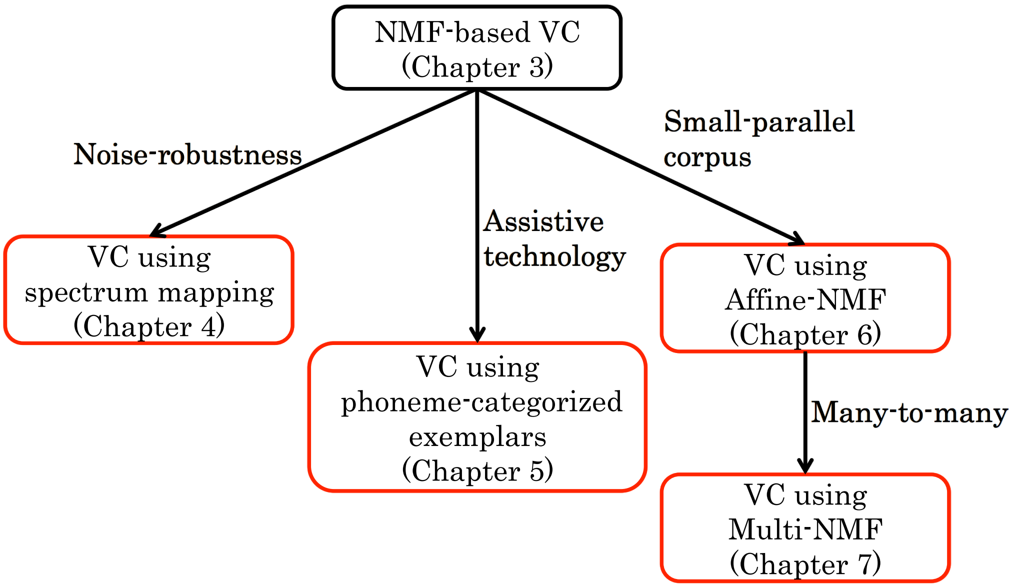


Figure 1.2: Flow of this thesis

1.3.1 Four Practical VC Tasks

1.3.1.1 Noise-robust VC

The conventional VC method is evaluated in a clean environment. Background noise, which is an unavoidable factor in real environments, has previously been ignored in the VC tasks. Because NMF has been used for source separation and a noise-robust ASR, the NMF-based VC method has noise robustness. In Chapter 4, we propose a framework to train the basis matrices of the source and target exemplars so they have a common weight matrix. We call this method "VC using sparse spectrum mapping". The basis matrix of the source exemplars is trained using NMF, then their weight matrix can be obtained. Next, the basis matrix of the target exemplars is trained using NMF, where the weight matrix is fixed to the one obtained from the source exemplars. By using the basis matrices instead of the exemplars, VC is performs in less computation time than the conventional exemplar-based method.

1.3.1.2 Assistive Technology for Articulation Disorders

In Chapter 5, we present a VC method for a person with an articulation disorder resulting from athetoid cerebral palsy. The movements of such speakers are limited by their athetoid symptoms, and their consonants are often unstable or unclear, which makes it difficult for them to communicate. In this paper, exemplar-based spectral conversion using NMF is applied to a voice with an articulation disorder. In order to preserve the speaker's individuality, we used a combined dictionary that was constructed from the source speaker's vowels and target speaker's consonants. However, this exemplar-based approach needs to hold all the training exemplars (frames), and may cause a mismatch of phonemes between the input signals and selected exemplars. In this paper, in order to reduce the mismatching of phoneme alignment, we propose a phoneme-categorized sub-dictionary and a dictionary selection method using NMF. The effectiveness of this method is confirmed by comparing it with those of conventional GMM-based and NMF-based methods.

1.3.1.3 VC Using Small-parallel Training Data

The need to have a large amount of parallel data is a great barrier to the practical application of VC. This paper presents a novel framework for exemplar-based VC that only requires a small number of parallel exemplars. In the framework of conventional GMM-based VC, some approaches that do not need parallel exemplars have been proposed. However, in the framework of exemplar-based VC in noisy environments, such a method has never been proposed. In Chapter 6, an adaptation matrix in an NMF framework is introduced to adapt the source dictionary to the target dictionary. This adaptation matrix is estimated using only a small-parallel speech corpus. We refer to this method as Affine-NMF, and its effectiveness is confirmed by comparing its with those of a conventional NMF-based method and a GMM-based method in noisy environments.

1.3.1.4 Many-to-many VC

A novel VC method for arbitrary speakers is proposed in Chapter 7. Because NMF-based VC requires parallel training data from source and target speakers, the voice of arbitrary speakers cannot be converted in this framework. In this study, we propose Multi-NMF to allow the implementation of many-to-many, exemplar-based VC. Our experimental results demonstrate that the conversion quality of the proposed method is close to that of conventional one-to-one VC, even though the proposed method requires neither the source nor the target speakers' spectra to be included in the training set.

1.3.2 Novelties of This Thesis

This paper proposes new algorithms for four practical VC tasks.

In "VC using sparse spectrum mapping" both noise-robustness and low-computational times are achieved. Background noise is problem not only in VC but also in other speech problem tasks for example ASR. Our proposed method will also effective, for such other tasks.

Assistive technology for a person with articulation disorders is proposed in Chapter 5. A phoneme-categorized sub-dictionary and a dictionary selection method enables individuality-preserving VC for articulation disorders.

By using VC based on Affine-NMF, conversion using small-parallel corpus is introduced. Affine-NMF enables speaker adaptation of the exemplars, which is used for exemplar-based VC. This adaptation technique will also effective in NMF-based speech separation.

Exemplar-based many-to-many VC is enabled by Multi-NMF. Experimental results show that the conversion quality of the proposed method is almost the same as that of conventional one-to-one VC, suggesting that it can be applied to voice quality control and noise-robust VC.

1.4 Outline

Starting in Chapter 2, we list some feature extraction methods related to VC and conventional GMM-based VC. In Chapter 3, basic NMF-based VC is described. In Chapter 4, noise-robust VC based on sparse spectral mapping is described. In Chapter 5, a VC method for articulation disorders is described. Chapter 6 describes a novel framework of NMF-based VC that only requires a small number of parallel exemplars. In Chapter 7, a many-to-many VC method using Multi-NMF is described. Finally, Chapter 8 concludes the thesis.

Chapter 2

Acoustic Features and Conventional VC methods

In this chapter, we review common acoustic features used in voice conversion and conventional VC methods. First, cepstrum analysis, which is the basis for the extraction of important information from a speech signal, is presented. Second, we discuss MFCC, which is a well-known feature in speech signal processing. Third, an analyzing speech synthesis method using vocoder is presented. Finally, conventional statistical VC methods are explained [21].

2.1 Acoustic Features

2.1.1 Cepstrum

The speech signal is generated from vibrations of the vocal cord. The vibration passes the vocal tract of the speaker and arrives at the listener's ears or a microphone. When we vibrate our vocal chords and change the shape of our mouth and vocal tracts, the sounds of various phonemes such as /a/ or /i/ can be generated. This speech generating process is modeled as a source-filter model (Figure 2.1). The filter associated with the vocal tract is called the formant and the fundamental frequency from the vibration is called the pitch. In speech signal processing, the information related to the formant that determines phonemes is reasonably important. Therefore, when the system recognizes a given speech, we obtain better results with the formant information than with the original spectrum.

One well-known technique for extracting formants is cepstrum analysis [29]. The cepstrum is obtained as follows: 1) execute Fourier transform to the given speech, 2) take absolute and logarithm, and 3) execute inverse Fourier transform. Letting $G(\omega)$ and $H(\omega)$ be the spectrum of the vocal cord and the tract, respectively, the spectrum of the speech $S(\omega)$ is represented as

$$S(\omega) = G(\omega) \cdot H(\omega). \tag{2.1}$$

When we apply a logarithm and inverse Fourier transform to Eq. (2.1), we obtain

$$\log |S(\omega)| = \log |G(\omega)| + \log |H(\omega)| \tag{2.2}$$

and

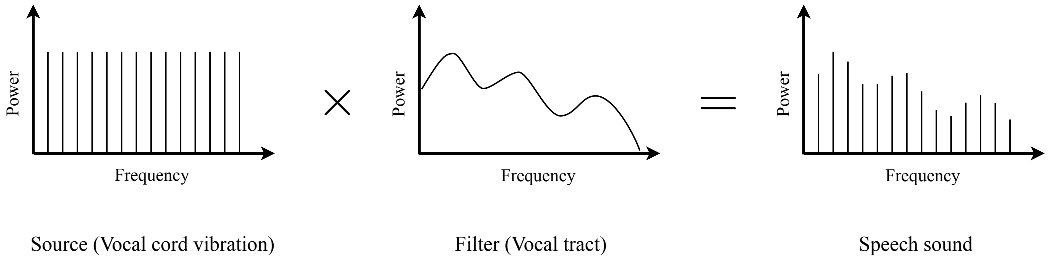
$$\begin{aligned} S_{cep}(d) &= DFT^{-1}\{\log |S(\omega)|\} && (2.3)\\ &= DFT^{-1}\{\log |G(\omega)|\} + DFT^{-1}\{\log |H(\omega)|\}, && (2.4) \end{aligned}$$

where $S_{cep}(d)$ indicates the d-th cepstrum (d is quefrency axis).

When we regard each spectrum in Figure 2.1 as a signal, we notice that the vocal cord $G(\omega)$ signal changes rapidly, and the vocal tract $H(\omega)$ signal changes slowly by contrast. By applying an inverse Fourier transform to these signals, $DFT^{-1}\{\log |G(\omega)|\}$ appears in high quefrency and $DFT^{-1}\{\log |H(\omega)|\}$ appears in low quefrency. Furthermore, as Eq. (2.4) shows, a speech signal is represented as the sum of the vocal cord and vocal tract information. Therefore, the vocal tract information $DFT^{-1}\{\log |H(\omega)|\}$ can be easily obtained by simple subtraction. In other words, the formant information is extracted by liftering (low-pass filtering in quefrency) as shown in Figure 2.2.

2.1.2 MFCC

The cepstrum feature introduced in the previous section was obtained as a linear log spectrum and is called the linear frequency cepstral coefficient; LFCC). On



Source (Vocal cord vibration) Filter (Vocal tract) Speech sound

Figure 2.1: Source filter model: speech sound can be generated by multiplying spectra of the vocal cord vibration (source) and spectra of the vocal tract (filter).

the other hand, there is another formant extraction method, MFCC [30] is extracted from the transformation on the mel-scale, which approximates the human auditory scale sensitive to pitch.

When we, human beings, hear something, auditory sensitivity becomes poorer as pitch increases. That means that the frequency resolution is very low at high frequencies and high at low frequencies. The relationship is not linear but non-linear, as approximated by

$$f' = 1127.01048 \log(1 + \frac{f}{700}), \tag{2.5}$$

where f' is the mel frequency. Typically, we use filter-bank representation instead of using the coefficients themselves. In the MFCC approach, mel-filter-banks as shown in Figure 2.3 are used. Each filter is triangular, and outputs the sum value of the multiplication. The power spectrum of the mel-scale frequency $M(i)$ is obtained by

$$M(i) = \sum_{f=f'_{i-1}}^{f'_{i+1}} W_i(f) \cdot |X(f)|^2. \tag{2.6}$$

From Eq. (2.6), the MFCC can be calculated as follows:

$$M_{cep}(d) = DFT^{-1}\{\log M(i)\}. \tag{2.7}$$

MFCC contains some useful information on representing speech using a small numbers of dimensions, therefore most works in speech signal processing, specifically in speech recognition, deal with these features. However, as the MFCC is based on a filter-bank calculation, it is, in general, difficult to reconstruct

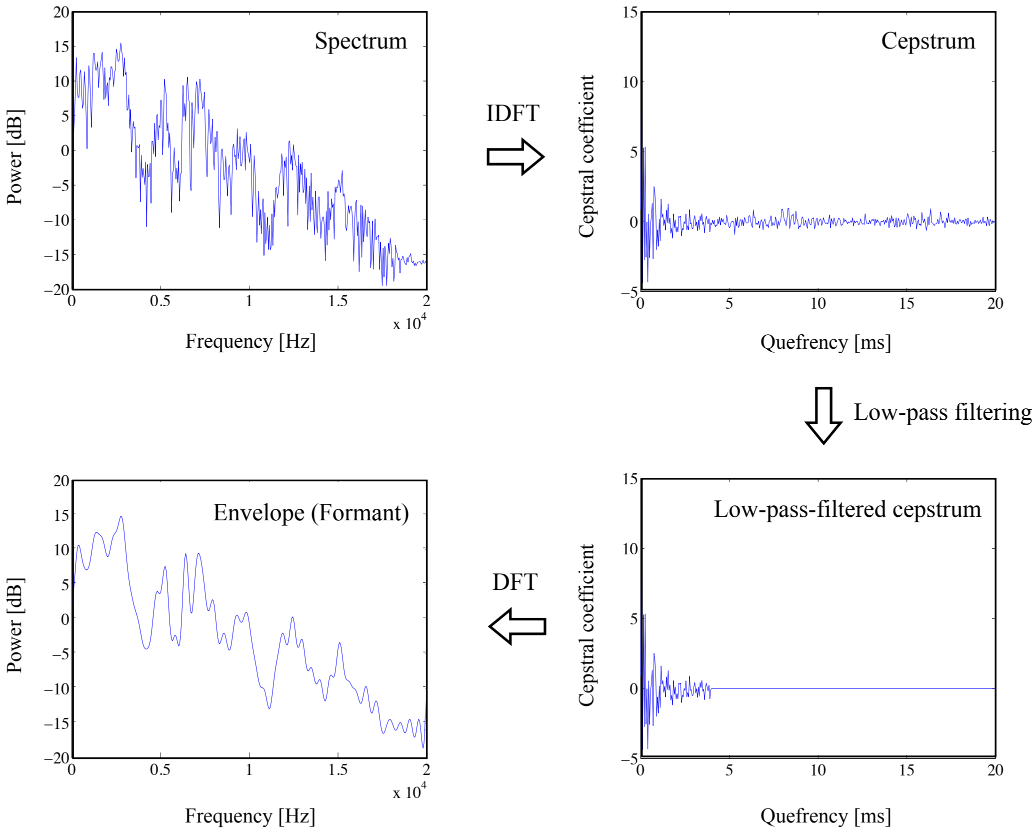


Figure 2.2: Formant extraction using cepstrum analysis. This example uses a speech signal where "a" is uttered.

the speech signal from the obtained MFCC because the values in a filter-bank are summed up with the weights. Milner and Shao proposed an approximation approach for speech reconstruction from MFCC using a source-filter model [31]. Other approaches such as using a z-transform [32, 33] and using a mel-log spectral approximation (MLSA) filter [34] have also been proposed.

2.1.3 Vocoder

Vocoder is an analysis and synthesis method for speech signal processing. This paper uses STRAIGHT [35, 36], which is a vocoder provided by Kawahara that analyzes, modifies, and synthesizes speech, written as MATLAB codes. The tool extracts three components from a speech signal: spectrum parameters, fundamen-

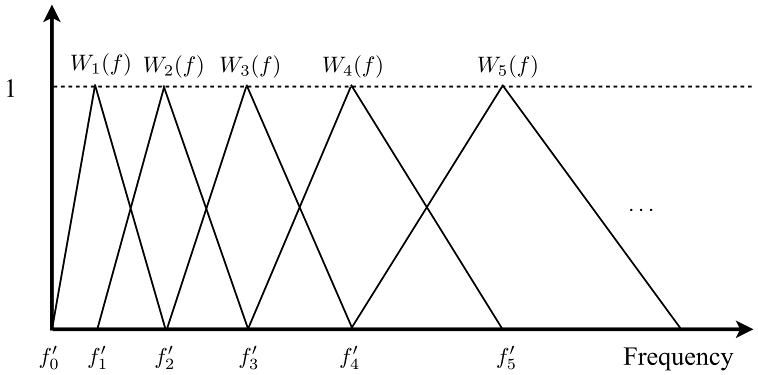


Figure 2.3: Mel-scale filter banks

tal frequencies, and aperiodic parameters. The analysis and synthesis quality is fairly high, so many researchers in speech signal processing use it. Figure 2.4 depicts the common way of using STRAIGHT to modify a speech signal. The spectrum features extracted using STRAIGHT resemble formant information in cepstrum analysis. Therefore, in voice conversion, we usually modify the spectrum features and the F_0 to the desired ones; we do not change the aperiodic features at the synthesis stage. The spectrum features are also referred to as the STRAIGHT spectrum, STRAIGHT parameters, or just "the spectrum". In the rest of this thesis, we refer to spectrum parameters extracted from a speech signal using STRAIGHT as the spectrum.

We can also extract MFCC features from the STRAIGHT spectrum. As discussed in the previous section, the (approximate) MFCC can further be back-projected into the spectrum space. The raw spectrum obtained using STRAIGHT tends to be high-dimensional. If we focus on modifying acoustic features in MFCC space after executing STRAIGHT, we can reduce the dimensional sizes of the features of interest, which leads to high performance in the estimation of the model, and obtain high-quality speech sounds.

In [37], Kawahara proposed advanced version of STRIAGHT, which is named as "TANDEM-STRAIGHT". In [38], Morise proposed another vocoder, WORLD which reduces computational cost compared with STRAIGHT. Ahocoder [39] is also used in some VC and TTS application.

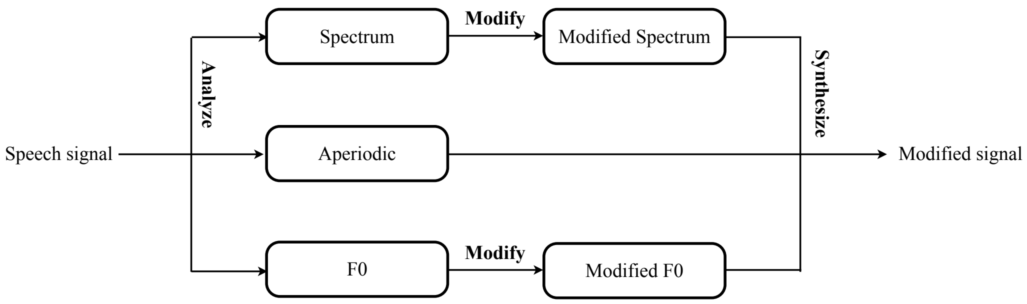


Figure 2.4: Modifying a speech signal using STRAIGHT.

2.2 Conventional VC Methods

As discussed in section 2.1, in order to estimate the models, most existing VC approaches need sets of speech signals where the same sentences are uttered by a source and a target speaker. Furthermore, the training data must be aligned at the frame level. The aligned data is called parallel data, and is often obtained using DTW or dynamic programming (DP) [3, 18, 40]. Such feature vectors are used for training each model. Let us refer to the D-dimensional feature vectors in each frame of the source and target speakers as $\mathbf{x}$ and $\mathbf{y}$, respectively. Assuming the parallel data includes T frames, the training data consists of the source speaker's set $\mathbf{X} \ni \{\mathbf{x}_t\}_{t=1}^{T}$ and the target speaker's set $\mathbf{Y} \ni \{\mathbf{y}_t\}_{t=1}^{T}$.

2.2.1 GMM-based VC

2.2.1.1 Gaussian Mixture Model (GMM)

A Gaussian mixture model (GMM) is a statistical probabilistic model for representing observed data that can be categorized into sub-components. Each component is represented as a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ with parameters of a D-dimensional mean vector $\boldsymbol{\mu}$ and a variance matrix $\boldsymbol{\Sigma}$, defined as

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right), \tag{2.8}$$

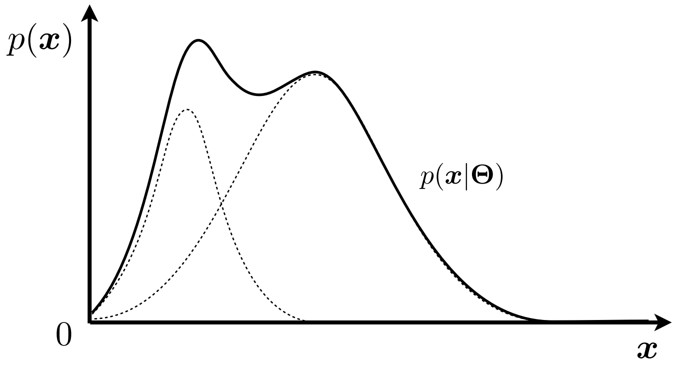


Figure 2.5: Example of a Gaussian mixture model ($M = 2$).

where $\mathbf{x}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ have the elements as follows:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1D}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{D1}^2 & \cdots & \cdots & \sigma_{DD}^2 \end{bmatrix}, \tag{2.9}$$

and $(\cdot)^{\mathsf{T}}$ and $|\cdot|$ indicate the transpose and determinant of the matrix, respectively.

GMM represents the overall distribution of the data using a weighted sum of the components. The probability density function (pdf) of GMM is defined as

$$p(x|\boldsymbol{\Theta}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{2.10}$$

where M indicates the number of mixtures. $\boldsymbol{\Theta}$ is a set of parameters of GMM, which contains α_m, $\boldsymbol{\mu}_m$, and $\boldsymbol{\Sigma}_m$ for all m.

Figure 2.5 shows an example of a one-dimensional GMM that has two components, depicted by a solid line. The GMM was obtained from the weighted sum of the two Gaussian distributions, depicted by dotted lines.

GMM parameters can be estimated using the expectation maximization (EM) algorithm [41]. The algorithm repeats E-step (expectation) and M-step (maximization) by turns. First, all parameters are randomly initialized. In the E-step, we calculate a Q-function (expectation of log-likelihood) defined as

$$\begin{aligned} Q(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}) &= E[\log p(\mathbf{x}, m|\hat{\boldsymbol{\Theta}})]_{p(m|\mathbf{x},\boldsymbol{\Theta})} \\ &= \sum_{m=1}^{M} p(m|\mathbf{x}, \boldsymbol{\Theta}) \log p(\mathbf{x}, m|\hat{\boldsymbol{\Theta}}), \end{aligned} \tag{2.11}$$

where

$$\begin{aligned} p(\mathbf{x}, m|\hat{\boldsymbol{\Theta}}) &= \prod_{n=1}^{N} p(\mathbf{x}_n, m|\hat{\boldsymbol{\Theta}}) \\ &= \prod_{n=1}^{N} \hat{\alpha}_m \mathcal{N}(\mathbf{x}_n; \hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m). \end{aligned} \tag{2.12}$$

Therefore, Eq. (2.11) becomes

$$\begin{aligned} Q(\hat{\boldsymbol{\Theta}}|\boldsymbol{\Theta}) = &\sum_k \sum_n p(m = k|\mathbf{x}_n, \boldsymbol{\Theta}) \log \hat{\alpha}_k \\ &+ \sum_k \sum_n p(m = k|\mathbf{x}_n, \boldsymbol{\Theta}) \log \mathcal{N}(\mathbf{x}_n; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k). \end{aligned} \tag{2.13}$$

In the M-step, update rules for each parameter are derived to maximize the Q-function (Eq. (2.13)). The derived update rules are

$$\hat{\alpha}_k = \frac{1}{N} \sum_{n=1}^{N} p(k|\mathbf{x}_n, \boldsymbol{\Theta}), \tag{2.14}$$

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_n p(k|\mathbf{x}_n, \boldsymbol{\Theta})\mathbf{x}_n}{\sum_n p(k|\mathbf{x}_n, \boldsymbol{\Theta})}, \tag{2.15}$$

$$\hat{\boldsymbol{\Sigma}}_k = \frac{\sum_n p(k|\mathbf{x}_n, \boldsymbol{\Theta})(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_n - \hat{\boldsymbol{\mu}}_k)^T}{\sum_n p(k|\mathbf{x}_n, \boldsymbol{\Theta})}, \tag{2.16}$$

where $p(k|\mathbf{x}_n, \boldsymbol{\Theta})$ is the probability that $\mathbf{x}_n$ is sampled from the k-th component, which is calculated by

$$p(k|\mathbf{x}_n, \boldsymbol{\Theta}) = \frac{\alpha_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \alpha_k \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}. \tag{2.17}$$

2.2.1.2 Conversion Based on Maximum Likelihood Estimation

When it comes to voice conversion based on GMM, we modeled a joint probability of the source and the target speakers using GMM. Therefore, this model is called the joint density GMM (JD-GMM). In the training stage of the JD-GMM, we use a joint vector $\mathbf{Z}$ that concatenates the source speakers vector $\mathbf{X} = [\mathbf{x}^\top \Delta\mathbf{x}^\top]^\top$ and target speakers vector $\mathbf{Y} = [\mathbf{y}^\top \Delta\mathbf{y}^\top]^\top$. (i.e., $\mathbf{Z} = [\mathbf{X}^\top \mathbf{Y}^\top]^\top$). The probability $p(\mathbf{Z})$ is modeled using GMM as follows:

$$p(\mathbf{Z}|\mathbf{\Theta}^{(z)}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\mathbf{Z}; \boldsymbol{\mu}_m^{(z)}, \mathbf{\Sigma}_m^{(z)}), \tag{2.18}$$

where $\boldsymbol{\mu}_m^{(z)}$ and $\mathbf{\Sigma}_m^{(z)}$ consist of

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \quad \mathbf{\Sigma}_m^{(z)} = \begin{bmatrix} \mathbf{\Sigma}_m^{(xx)} & \mathbf{\Sigma}_m^{(xy)} \\ \mathbf{\Sigma}_m^{(yx)} & \mathbf{\Sigma}_m^{(yy)} \end{bmatrix}. \tag{2.19}$$

The parameters $\boldsymbol{\mu}_m^{(x)}$ and $\mathbf{\Sigma}_m^{(xx)}$, and the parameters $\boldsymbol{\mu}_m^{(y)}$ and $\mathbf{\Sigma}_m^{(yy)}$ correspond to the source speaker's and target speaker's Gaussian distributions, respectively. The parameter $\mathbf{\Sigma}_m^{(xy)} (= \mathbf{\Sigma}_m^{(yx)\top})$ indicates the covariance matrix between the observed data $\mathbf{X}$ and $\mathbf{Y}$. In voice conversion, we usually use a diagonal matrix for $\mathbf{\Sigma}_m^{(xx)}$, $\mathbf{\Sigma}_m^{(xy)}$, and $\mathbf{\Sigma}_m^{(xx)}$ because full-covariance matrices involve the estimation of a lot of parameters.

At the conversion stage (assuming that the parameters $\mathbf{\Theta}^{(z)}$ have already been estimated), we consider the probability of $\mathbf{Y}$ given an input $\mathbf{X}$. That is

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta}^{(z)}) &= \sum_{all\mathbf{m}} p(\mathbf{m}|\mathbf{X}, \mathbf{\Theta}^{(z)}) p(\mathbf{Y}|\mathbf{X}, \mathbf{m}, \mathbf{\Theta}^{(z)}) \\ &= \prod_{t=1}^{T} \sum_{m=1}^{M} p(m|\mathbf{X}_t, \mathbf{\Theta}^{(z)}) p(\mathbf{Y}_t|\mathbf{X}_t, m, \mathbf{\Theta}^{(z)}) \end{aligned} \tag{2.20}$$

where $\mathbf{m} = \{m_1, m_2, \cdots, m_T\}$ is a mixture component sequence. The probabilities on the right side in Eq. (2.20) are represented as

$$p(m|\mathbf{X}_t, \mathbf{\Theta}^{(z)}) = \frac{\alpha_m \mathcal{N}(\mathbf{X}_t; \boldsymbol{\mu}_m^{(x)}, \mathbf{\Sigma}_m^{(xx)})}{\sum_{n=1}^{M} \alpha_n \mathcal{N}(\mathbf{X}_t; \boldsymbol{\mu}_n^{(x)}, \mathbf{\Sigma}_n^{(xx)})} \tag{2.21}$$

$$p(\mathbf{Y}_t|\mathbf{X}_t, m, \mathbf{\Theta}^{(z)}) = \mathcal{N}(\mathbf{Y}_t; \mathbf{E}_{m,t}^{(y|x)}, \mathbf{D}_m^{(y|x)}) \tag{2.22}$$

$$\mathbf{E}_{m,t}^{(y|x)} = \boldsymbol{\mu}_m^{(y)} + \mathbf{\Sigma}_m^{(yx)} (\mathbf{\Sigma}_m^{(xx)})^{-1} (\mathbf{X}_t - \boldsymbol{\mu}_m^{(x)}) \tag{2.23}$$

$$\mathbf{D}_m^{(y|x)} = \mathbf{\Sigma}_m^{(yy)} - \mathbf{\Sigma}_m^{(yx)} (\mathbf{\Sigma}_m^{(xx)})^{-1} \mathbf{\Sigma}_m^{(xy)}. \tag{2.24}$$

A time sequence of the converted feature vector $\hat{\mathbf{y}}$ is determined as follows:

$$\hat{\mathbf{y}} = \arg\max P(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta}^{(z)}). \tag{2.25}$$

Eq. (2.25) is performed under the linear conversion between the static feature vectors $\mathbf{y}$ and the static and dynamic feature vectors $\mathbf{Y}$

$$\mathbf{Y} = \mathbf{W}\mathbf{y} \tag{2.26}$$

where $\mathbf{W}$ is a transformation matrix [42].

Eq. (2.20) is approximated with a single mixture component sequence as follows:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta}^{(z)}) \simeq p(\hat{\mathbf{m}}|\mathbf{X}, \mathbf{\Theta}^{(z)}) p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \mathbf{\Theta}^{(z)}). \tag{2.27}$$

$\hat{\mathbf{m}}$ denotes the suboptimum mixture component sequence, which is determined as follows

$$\hat{\mathbf{m}} = \arg\max P(\mathbf{m}|\mathbf{X}, \mathbf{\Theta}^{(z)}). \tag{2.28}$$

The logarithm of the likelihood function is written as

$$\log p(\mathbf{Y}|\mathbf{X}, \hat{\mathbf{m}}, \mathbf{\Theta}^{(z)}) = -\frac{1}{2}\mathbf{Y}^{\mathsf{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{Y} + \mathbf{Y}^{\mathsf{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)} + cons \tag{2.29}$$

where

$$\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)} = [\mathbf{E}_{\hat{m}_1,1}^{(y|x)}, \mathbf{E}_{\hat{m}_2,2}^{(y|x)}, \cdots, \mathbf{E}_{\hat{m}_T,T}^{(y|x)}] \tag{2.30}$$

$$\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)} = \mathrm{diag}[\mathbf{D}_{\hat{m}_1,1}^{(y|x)}, \mathbf{D}_{\hat{m}_2,2}^{(y|x)}, \cdots, \mathbf{D}_{\hat{m}_T,T}^{(y|x)}]. \tag{2.31}$$

we can estimate the most probable $\hat{\mathbf{y}}$ as follows:

$$\hat{\mathbf{y}} = (\mathbf{W}^{\mathsf{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{W})^{-1}\mathbf{W}^{\mathsf{T}}\mathbf{D}_{\hat{\mathbf{m}}}^{(y|x)^{-1}}\mathbf{E}_{\hat{\mathbf{m}}}^{(y|x)}. \tag{2.32}$$

We can also maximize the logarithm of the likelihood function of Eq. (2.20) by employing the EM algorithm, however, there is little difference in the conversion accuracy between it and the suboptimum mixture component sequence [19].

2.2.1.3 Problems

It is often reported that the GMM-based VC includes over-smoothing problems and over-fitting. The over-smoothing arises because the parameters of multiple Gaussian components are estimated by averaging observations with similar context descriptions. As a result, the outputs distribute near the modes of each component. Over-fitting problems come from this complexity. If we supply more Gaussian components, the model is over-fitted to the training data.

Some methods to tackle the over-smoothing problems have been proposed. Toda *et. al.* proposed a Global Variance (GV) of converted spectra over a time sequence [19]. A modulation spectrum is used for advanced methods of GV in [43].

2.2.2 PLS-based VC

Partial least squares (PLS) regression is a statistical method that combines PCA and multivariate regression. The observed variable is generated by a small number of latent variables, which explains most of the variations in the target.

In PLS-based VC, a source speaker's vector $\mathbf{X}$ is generated by a small number of latent variables, which explains most of the variations in the target speaker's vector $\mathbf{Y}$. $\mathbf{X}_t$ and $\mathbf{Y}_t$ are represented by a linear transformation of the speaker-independent latent variable vector $\mathbf{h}_t$ as follows:

$$\mathbf{X}_t = \mathbf{Q}\mathbf{h}_t + \mathbf{e}_t^x \tag{2.33}$$

$$\mathbf{Y}_t = \mathbf{P}\mathbf{h}_t + \mathbf{e}_t^y \tag{2.34}$$

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 the SIMPLS algorithm [44].

In the test phase, segmental features $\mathbf{X}$ are constructed from the phonetic discriminative source features of test data. The converted spectral feature $\hat{\mathbf{Y}}$ is obtained as follows [28]:

$$\hat{\mathbf{Y}}_t = \beta \mathbf{X}_t \tag{2.35}$$

Helander *et. al.* proposed dynamic kernel PLS (DKPLS)-based VC in [45]. In this, source spectral features are projected to high-dimensional feature space using kernel transformation, and the transformed source features are regressed with target spectral features. Their approaches was evaluated using a small number of parallel training data and it outperformed GMM-based VC However, it has not been evaluated in a situation involving a standard setting of parallel training data.

Chapter 3

Voice Conversion Based on Non-negative Matrix Factorization

3.1 Non-negative Matrix Factorization

NMF is a sparse representation method with non-negativity constraints. In the sparse representation approach, the observed signal is represented by a linear combination of a small number of bases:

$$\mathbf{v}_l \approx \textstyle\sum_{j=1}^{J} \mathbf{w}_j h_{jl} = \mathbf{W}\mathbf{h}_l \tag{3.1}$$

where $\mathbf{v}_l$ represents the l-th frame of the observation, $\mathbf{w}_j$ and h_{jl} represent the j-th basis and the weight, respectively, and $\mathbf{W} = [\mathbf{w}_1 \dots \mathbf{w}_J]$ and $\mathbf{h}_l = [h_{1l} \dots h_{Jl}]^\mathsf{T}$ are the collection of the bases and weights. In this paper, the collection of the bases are called the "dictionary" and weights are called the "activity".

Approaches using NMF are divided into two categories: supervised NMF and unsupervised NMF. Supervised NMF estimates only activity from observation: the dictionary must be provided. In contrast, unsupervised NMF generates both a dictionary and an activity from observation. Exemplar-based VC, is a supervised technique.

Lee *et al.* proposed an NMF algorithm [25]. The cost function used was the Euclidean distance and the Kullback-Leibler (KL) divergence. NMF using other divergence functions have been proposed [46]. This paper adopts KL-NMF, which is widely used in speech signal processing.

A number of sparseness criteria have been proposed, among which L1 norm regularization has been most widely used [47]. We also use it in our NMF-based VC. A quantitative definition of sparseness was proposed in [48], however this is not a closed-form. Group sparsity was introduced in [49].

In almost all approaches using NMF, a maximization-minimization algorithm [25] is used for optimization, and we adopt it in our proposed Multi-NMF. In recent research, Newton's method has been employed, through other optimization methods have also been proposed [50]. Virtanen *et al.* [51] used an active-set Newton method to reduce the computational time of supervised-NMF.

NMF has been applied to a range of speech processing tasks. Gemmeke *et al.* [47, 52] proposed noise robust automatic speech recognition using supervised NMF. Unsupervised NMF has been used in single channel speech separation [53, 54] and music transcription [55, 56], among other application. NMF has formed the basis for Non-negative Tensor Factorization (NTF) and applied to the analysis of brain data [57] and source separation [58]. Our proposed Multi-NMF is an extension of conventional NMF, and one of the simplest forms of NTF.

3.2 Exemplar-based VC

3.2.1 Basic Idea

In the exemplar-based approach, the observed signal is represented by a linear combination of a small number of bases. Each basis represents an exemplar of the spectrum, calculated from the training data. The observed spectra can therefore be represented by a linear combination of a small number of bases with non-zero weights:

$$\mathbf{V} \approx \mathbf{W}\mathbf{H} \tag{3.2}$$

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_L], \quad \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_L] \tag{3.3}$$

where $\mathbf{V}, \mathbf{W}, \mathbf{H}$, and L represent the observed spectra, the collection of bases, the collection of weights, and the number of frames of the observed spectra, respectively. The collection of bases and weights are referred to hereafter as the dictionary and the activity, respectively. In this exemplar-based VC, we use supervised NMF [25]. The dictionary is fixed and NMF is used only to estimate activity.

Fig. 3.1 shows the basic approach of our exemplar-based VC, where D, L, and J represent the numbers of dimensions, frames, and bases, respectively. Our method requires two parallel dictionaries. $\mathbf{W}^s$ represents a source dictionary that consists of the source speaker exemplars, and $\mathbf{W}^t$ represents a target dictionary that consists of the target speaker exemplars. The two dictionaries contain the same set of words, and are aligned using dynamic time warping (DTW), just as in conventional GMM-based VC. Hence, these dictionaries have the same number of bases.

A matrix of input source spectra $\mathbf{V}^s$ is decomposed into the source dictionary $\mathbf{W}^s$ and the activity matrix $\mathbf{H}^s$. This method assumes that the obtained activity matrices will be approximately equivalent, when the source signal and the target signal (the same words but spoken by different speakers) are expressed with sparse representations of the source dictionary and the target dictionary, respectively. Fig. 3.2 shows the activity matrices estimated from the parallel dictionaries. As can be seen, the shapes of these two activity matrices are similar. This phenomenon shows that parallel activity matrices capture speaker-independent information when they are estimated from parallel dictionaries. Therefore, a matrix of target spectra $\hat{\mathbf{V}}^t$ can be constructed using the target dictionary $\mathbf{W}^t$ and the activity matrix of the source signal $\mathbf{H}^s$, as shown in Fig. 3.1.

3.2.2 Dictionary Construction

Fig. 3.3 illustrates the process for constructing parallel dictionaries. First, we construct the parallel dictionaries of the source and target speakers. To do so, parallel spectra are extracted from the parallel words of the source and target speakers. Using Dynamic Time Warping (DTW), these spectra are then aligned so that they have the same number of frames. Then, the source and target dictionaries are obtained by lining up the parallel spectra, which are used as parallel training data in GMM-based VC [19]. Therefore, the dictionary consists

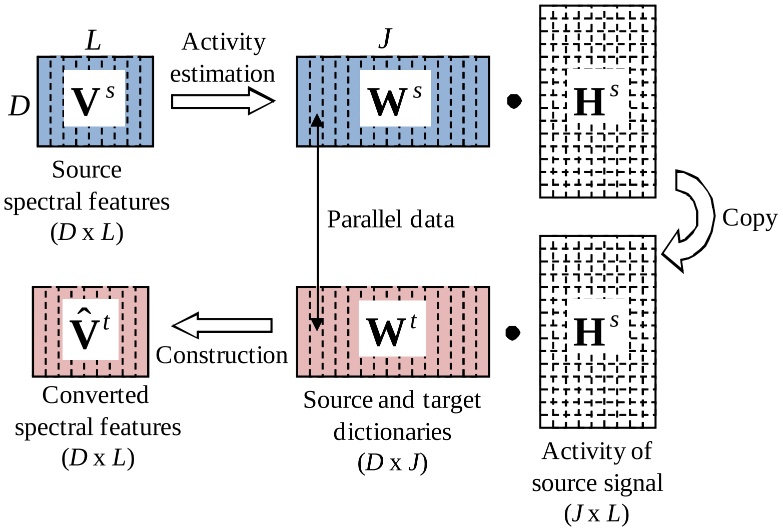

L
D
V^s
Activity estimation
J
W^s
H^s
Source spectral features (D x L)
Parallel data
Copy
V̂^t
Construction
W^t
H^s
Converted spectral features (D x L)
Source and target dictionaries (D x J)
Activity of source signal (J x L)

Figure 3.1: One-to-one VC using NMF

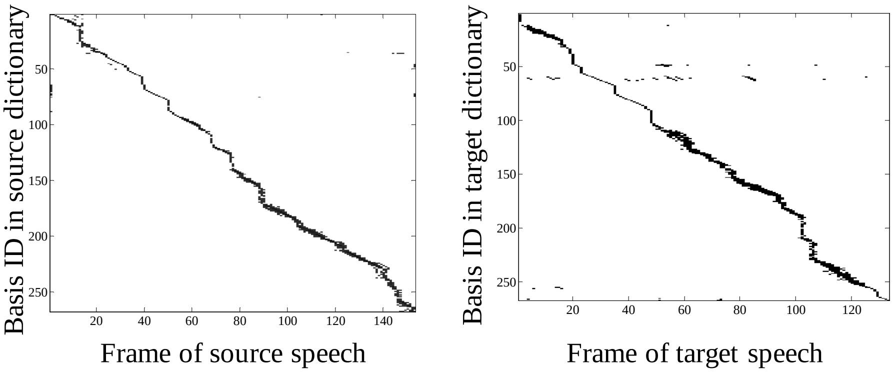


Figure 3.2: Activity matrices for parallel utterances.

of short-time spectra obtained from all training speech data using Short-Time Fourier Transform (STFT), where one spectrum corresponds to one basis of the dictionary. Using this method, unlike GMM-based VC, no dictionary training procedure is required.

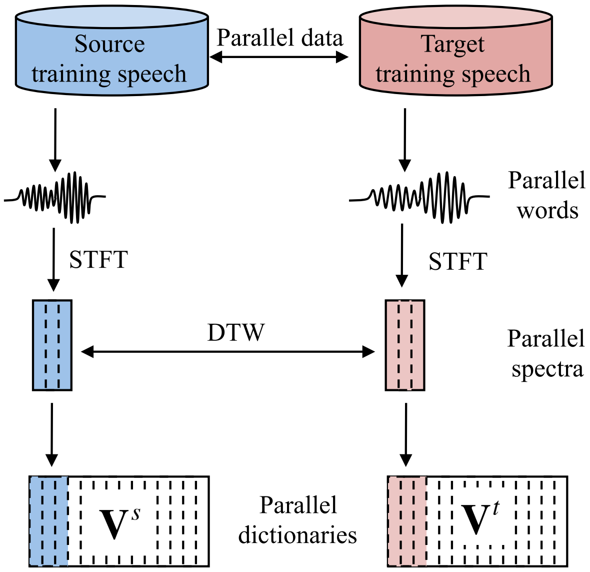


Figure 3.3: Construction of parallel dictionaries

3.2.3 Activity Estimation

$\mathbf{W}^s$ and $\mathbf{W}^t$ are fixed, and the source speaker's activity matrix $\mathbf{H}^s$ is estimated using NMF. The cost function of NMF is defined as follows:

$$d(\mathbf{V}^s, \mathbf{W}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \quad s.t. \quad \mathbf{H} \geq 0. \tag{3.4}$$

In (3.4), the first term is the KL divergence between $\mathbf{V}^s$ and $\mathbf{W}^s\mathbf{H}^s$ and the second term is the sparse constraint with the L1-norm regularization term that causes the activity matrix to be sparse. λ represents a weighting of sparse constraints. This function is minimized by iteratively updating the following equation [47]:

$$\begin{aligned}\mathbf{H}^s \quad \leftarrow \quad & \mathbf{H}^s.*(\mathbf{W}^{s\mathsf{T}}(\mathbf{V}^s./(\mathbf{W}^s\mathbf{H}^s))) \\ & ./(\mathbf{W}^{s\mathsf{T}}\mathbf{1}^{D\times L} + \lambda\mathbf{1}^{J\times L})\end{aligned} \tag{3.5}$$

where $.*$, $./$, and $\mathbf{1}$ denote element-wise multiplication, element-wise division, and an all-ones matrix, respectively. The derivation of (3.5) is shown in Appendix 8.

The estimated source activity $\mathbf{H}^s$ is multiplied to the target dictionary $\mathbf{W}^t$, and the target spectra $\hat{\mathbf{V}}^t$ are constructed.

$$\hat{\mathbf{V}}^t = \mathbf{W}^t\mathbf{H}^s \tag{3.6}$$

Chapter 4

Noise-Robust Voice Conversion Based on Sparse Spectral Mapping

The related publications for this chapter are [59].

4.1 The Motivation and Related Work

4.1.1 Motivation

Background noise is an unavoidable factor in speech processing. In the task of automatic speech recognition (ASR), one problem is that the recognition performance remarkably decreases under noisy environments, and this creates a significant problem in regard to the development of a practical use of ASR.

The problem of background noise is also troublesome in the task of VC. 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. In this capter, we propose noise-robust VC based on sparse representation.

NMF [25] is a popular approach for source separation or speech enhancement [53, 54]. In some approaches for NMF-based source separation, the exemplars, which are called “bases”, are grouped for each source, and the mixed signals are expressed with a sparse representation of these bases. By using only the weights of the atoms related to the target signal, the target signal can be reconstructed. Gemmeke *et al.* [47] also propose an exemplar-based method for noise-robust speech recognition using NMF. Noise-robust ASR using NMF has also been proposed in [60].

In our previous work, we proposed an exemplar-based method for noise-robust Voice Conversion (VC) using NMF. VC is a technique for converting a speaker’s voice individuality while maintaining phonetic information in the utterance. In [23], we evaluated the conventional statistical VC method in a noisy environment and revealed that noise in the input signal is not only output with the converted signal, but also tends to degrade the conversion performance itself due to unexpected mapping of source features. In NMF-based VC, noise exemplars are extracted from before- and after- utterance sections and input noisy signals are decomposed into a linear combination of noise and speaker’s clean exemplars. For this reason, no training processes related to noise signals are required. Only the weights related to the source exemplars are taken, and the target signal is constructed from the target exemplars and the 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.

Therefore, this chapter propose a novel noise robust VC which requires less computational times compared to the conventional VC method. 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 that of an exemplar-based method and the conventional Gaussian mixture model (GMM)-based method.

4.2 Proposed Method

4.2.1 Dictionary Construction for a Noisy Environment

In the preceding section, both dictionaries (source and target) consisted of the same spectral envelope features (STRAIGHT spectrum) for simplicity in explaining the proposed method. Indeed, the use of these features worked without any problems in a preliminary experiment using clean speech data. However, when it came to constructing a noise dictionary, STRAIGHT analysis could not express the noise spectrum well since STRAIGHT itself is an analysis and synthesis method for speech data. In order to express the noisy source speech with a sparse representation of source and noise dictionaries, a simple magnitude spectrum calculated using short-time Fourier transform (STFT) is used to construct the source and noise dictionaries.

Parallel dictionaries are constructed from clean speech data. For the target training speech, a STRAIGHT spectrum is used to extract its dictionary. Mel-cepstral coefficients are estimated from the STRAIGHT spectrum and used for DTW. For the source training speech, on the other hand, the STRAIGHT spectrum is converted into mel-cepstral coefficients and only used for DTW in order to align the temporal fluctuation, and the magnitude spectrum is used to extract its dictionary. When an input source signal is converted, the source signal is also applied to STFT and STRAIGHT analysis. The magnitude spectrum is used to extract the noise dictionary and to estimate the activity. The STRAIGHT spectrum, F0 and aperiodic components are used to synthesize the converted signal.

4.2.2 Training of the Parallel Basis Matrices

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 the sparse representations of $\mathbf{A}^s$ and $\mathbf{A}^t$, respectively, the obtained activity matrices are equivalent, as shown in Fig. 3.2.

Table 4.1 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 NMF with the sparse constraint [47]. In the framework of NMF with the sparse constraint, it minimizes the following cost function:

$$\begin{aligned} d(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}^s) + ||(\lambda_{train}\mathbf{1}^{(1\times L)}).*\mathbf{H}^s||_1 \\ s.t. \quad \mathbf{A}^s,\ \mathbf{H}^s \geq 0. \end{aligned} \tag{4.1}$$

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 causes $\mathbf{H}^s$ to be sparse. λ_{train} is the weight of the sparse constraint. $\mathbf{A}^s$ and $\mathbf{H}^s$ minimizing (4.1) are estimated iteratively applying the following update rules:

$$\mathbf{A}^s_{n+1} = \mathbf{A}^s_n.*(\mathbf{H}^s_n(\mathbf{X}^s./\mathbf{A}^s_n\mathbf{H}^s_n)^{\mathsf{T}}./(\mathbf{H}^s_n\mathbf{1}^{(L\times D)}))^{\mathsf{T}} \tag{4.2}$$

$$\begin{aligned} \mathbf{H}^s_{n+1} &= \mathbf{H}^s_n.*(\mathbf{A}^{s^{\mathsf{T}}}_n(\mathbf{X}^s./(\mathbf{A}^s_n\mathbf{H}^s_n))) \\ &\quad ./(\mathbf{A}^{s^{\mathsf{T}}}_n\mathbf{1}^{(J\times L)} + \lambda_{train}\mathbf{1}^{(1\times L)}) \end{aligned} \tag{4.3}$$

where ./ is an element-wise division.

Next, using the activity matrix $\mathbf{H}^s$ obtained by (4.3), 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:

$$d(\mathbf{X}^t, \mathbf{A}^t\mathbf{H}^s) \quad s.t. \quad \mathbf{A}^t \geq 0. \tag{4.4}$$

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

$$\mathbf{A}^t_{n+1} = \mathbf{A}^t_n.*(\mathbf{H}^s(\mathbf{X}^t./\mathbf{A}^t_n\mathbf{H}^s)^{\mathsf{T}}./(\mathbf{H}^s\mathbf{1}^{(L\times D)}))^{\mathsf{T}}. \tag{4.5}$$

$\mathbf{A}^t_n$ and $\mathbf{A}^t_{n+1}$ represent the target basis matrices of the n-th and the $(n+1)$-th iteration, respectively.

Table 4.1: 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 (4.2) and (4.3)
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 (4.5)
```

4.2.3 Voice Conversion of Noisy Source Signal

4.2.3.1 Estimation of Activity from Noisy Source Signal

Figure 4.1 shows the conversion framework of our method. The exemplars (frames) of the noise are extracted from the before- and after-utterance sections in the observed (noisy) signal, 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 clean source dictionary, noise dictionary, and their activities.

$$\begin{aligned}
\mathbf{x}_l &= \mathbf{x}_l^s + \mathbf{x}_l^n \\
&\approx \sum_{j=1}^{J} \mathbf{a}_j^s h_{j,l}^s + \sum_{k=1}^{K} \mathbf{a}_k^n h_{k,l}^n \\
&= [\hat{\mathbf{A}}^s \mathbf{A}^n] \begin{bmatrix} \mathbf{h}_l^s \\ \mathbf{h}_l^n \end{bmatrix} \quad s.t. \quad \mathbf{h}_l^s, \mathbf{h}_l^n \geq 0 \\
&= \mathbf{A}\mathbf{h}_l \quad s.t. \quad \mathbf{h}_l \geq 0 \qquad (4.6)
\end{aligned}$$

$\mathbf{x}_l^s$ and $\mathbf{x}_l^n$ are the magnitude spectra of the source signal and the noise, respectively. $\hat{\mathbf{A}}^s$, $\mathbf{A}^n$, $\mathbf{h}_l^s$ and $\mathbf{h}_l^n$ are the source dictionary (basis matrix) trained by (4.2), noise dictionary (exemplars), and their activities at frame l, respectively.

Given the spectrogram, (4.6) can be written as follows:

$$\begin{aligned}\mathbf{X} &\approx [\hat{\mathbf{A}}^s\mathbf{A}^n]\begin{bmatrix}\mathbf{H}^s\\ \mathbf{H}^n\end{bmatrix} \quad s.t. \quad \mathbf{H}^s, \mathbf{H}^n \geq 0\\ &= \mathbf{AH} \quad s.t. \quad \mathbf{H} \geq 0. \end{aligned} \tag{4.7}$$

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

$$\begin{aligned}\mathbf{M} &= \mathbf{1}^{(D\times D)}\mathbf{X}\\ \bar{\mathbf{X}} &\leftarrow \mathbf{X}./\mathbf{M}\\ \bar{\mathbf{A}} &\leftarrow \mathbf{A}./(\mathbf{1}^{(D\times D)}\mathbf{A})\end{aligned} \tag{4.8}$$

$\bar{\mathbf{X}}$ and $\bar{\mathbf{A}}$ are normalized $\mathbf{A}$ and $\mathbf{X}$, respectively. The joint matrix $\mathbf{H}$ is estimated based on NMF with the sparse constraint that minimizes the following cost function:

$$d(\bar{\mathbf{X}}, \bar{\mathbf{A}}\mathbf{H}) + ||(\lambda_{conv}\mathbf{1}^{(1\times L)}). * \mathbf{H}||_1 \quad s.t. \quad \mathbf{H} \geq 0. \tag{4.9}$$

The weights of the sparsity constraints can be defined for each basis and exemplar by defining $\lambda_{conv}{}^{\mathsf{T}} = [\lambda_1 \dots \lambda_J \dots \lambda_{J+K}]$. In this paper, the weights for source bases $[\lambda_1 \dots \lambda_J]$ were set to 0.15, and those for noise exemplars $[\lambda_{J+1} \dots \lambda_{J+K}]$ were set to 0. $\mathbf{H}$ minimizing (4.9) is estimated iteratively applying the following update rule:

$$\begin{aligned}\mathbf{H}_{n+1} &= \mathbf{H}_n. * (\bar{\mathbf{A}}^{\mathsf{T}}(\bar{\mathbf{X}}./(\bar{\mathbf{A}}\mathbf{H}_n)))\\ &\quad ./(\mathbf{1}^{((J+K)\times L)} + \lambda_{conv}\mathbf{1}^{(1\times L)}).\end{aligned} \tag{4.10}$$

$\mathbf{H}_n$ and $\mathbf{H}_{n+1}$ represent the activity matrices of the n-th and the $(n+1)$-th iteration, respectively.

4.2.3.2 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.

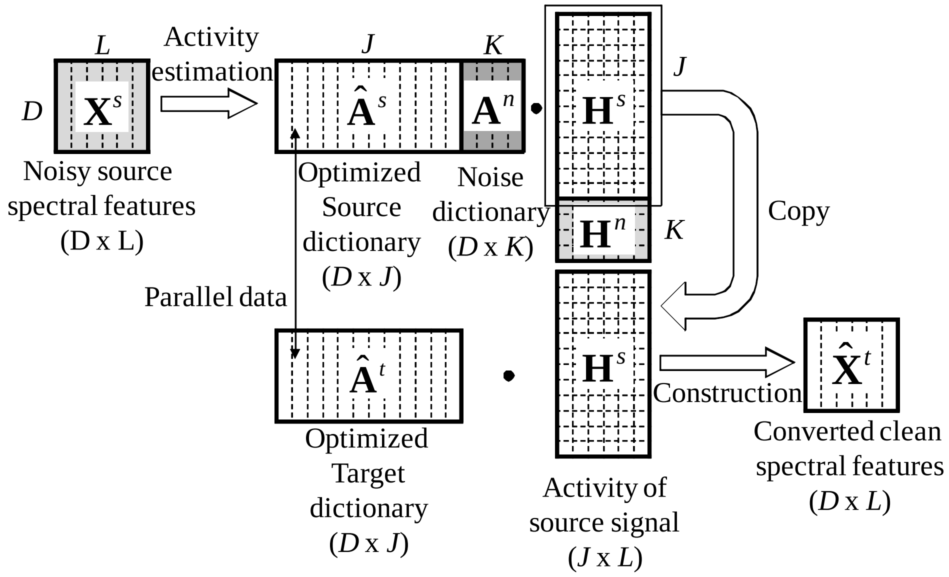


Figure 4.1: Proposed noise-robust voice conversion

Then, the target dictionary is also normalized for each basis in the same way the source dictionary was.

$$\overline{\hat{\mathbf{A}}^t} \leftarrow \hat{\mathbf{A}}^t./(\mathbf{1}^{(D\times D)}\hat{\mathbf{A}}^t) \tag{4.11}$$

$\hat{\mathbf{A}}^t$ is the target dictionary (basis matrix) trained by (4.5) and $\overline{\hat{\mathbf{A}}^t}$ is the normalized target dictionary of $\hat{\mathbf{A}}^t$. Next, the normalized target spectral feature is constructed, and the magnitudes of the source signal calculated in (4.8) are applied to the normalized target spectral feature.

$$\hat{\mathbf{X}}^t = (\overline{\hat{\mathbf{A}}^t}\mathbf{H}^s). * \mathbf{M} \tag{4.12}$$

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 as follows:

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)} \tag{4.13}$$

where x_t, $\hat{y}_t$, $\mu^{(x)}$, $\mu^{(y)}$, $\sigma^{(x)}$, and $\sigma^{(y)}$ are a log F0 of the source speaker and the converted F0 at frame t, mean of the source and target speaker's log F0, standard deviation of the source and target speaker's log F0, respectively. Mean and standard deviation are calculated from training data of the source and target speaker.

4.3 Experiments

4.3.1 Experimental Conditions

The proposed VC technique was evaluated by comparing it with an exemplar-based method [23] and a conventional GMM-based method [3] 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 [61], respectively. The sampling rate was 8 kHz.

A total of 216 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 the GMM-based method. In the exemplar-based method, the number of exemplars of the source and target dictionaries was 58,426. Then, in our proposed method, several bases were trained from the exemplars for each dictionary. Twenty-five sentences of clean speech or noisy speech were used in the evaluation. The noisy speech was created by adding a noise signal recorded in a restaurant (taken from the CENSREC-1-C database [62]) to the clean speech sentences. The SNR was 15 dB. The noise dictionary is extracted from the before- and after-utterance sections in the evaluation sentence. The average number of exemplars in the noise dictionary for one sentence was 110.

In the methods based on sparse representation, a 257-dimensional magnitude spectrum was used as the feature vectors for the 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 1st through 40th linear-cepstral coefficients obtained from the STRAIGHT spectrum were used as the feature vectors. The number of mixtures was 64.

4.3.2 Experimental Results

Table 4.2 shows the spectral distortion improvement ratio (SDIR) [dB] and the computation time of the conversion method (1 sentence on Intel Core i7 2.80 GHz personal computer) for noisy input source signal. In our proposed method, 1,000, 2,500 and 5,000 bases were trained from the exemplars for each dictionary. In the exemplar-based method, all 58,426 exemplars and 1,000 exemplars (which are chosen randomly) are used. The SDIR is defined as follows.

$$\mathrm{SDIR[dB]} = 10 \log_{10} \frac{\sum_d |\mathbf{X}^t(d) - \mathbf{X}^s(d)|^2}{\sum_d |\mathbf{X}^t(d) - \hat{\mathbf{X}}^t(d)|^2} \tag{4.14}$$

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 improvement for the methods based on the sparse representation was higher than the GMM-based method regardless of the number of the trained bases. In our proposed method, the case of 1,000 bases shows the best distortion improvement. The distortion improvement of the proposed method was slightly lower than that of the exemplar-based method which uses all 58,426 exemplars. However, compared to the exemplar-based method (which uses 1,000 exemplars) our proposed method obtained higher distortion improvement. Moreover, for obtaining the activity matrix, the computation time of the proposed method (which uses 1,000 exemplars) was about 30 times faster than that of the exemplar-based method, which uses all 58,426 exemplars. The computation time is reduced as the number of the bases is reduced.

We performed a mean opinion score (MOS) test [63] 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 were.

Figure 7.4 shows the mean opinion scores (MOS) for each method. The error bars show 95% confidence intervals. As shown in this figure, when clean

Table 4.2: Spectral distortion improvement ratio (SDIR) [dB] for noisy speech

	SDIR [dB]	time [s]
Proposed (1,000 bases)	5.14	30
Proposed (2,500 bases)	4.68	75
Proposed (5,000 bases)	4.38	137
Exemplar-based (58,426 exemplars)	5.23	910
Exemplar-based (1,000 exemplars)	4.91	30
GMM-based (64 mixtures)	4.11	1

speech data was used, the performances of the three methods were not so different under both evaluation criteria. However, when noisy speech data was used, the performances of the 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 the GMM-based method. The performances of the proposed method were slightly lower than those of the exemplar-based method when noisy speech data was used.

4.4 Chapter 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 NMF. Then, 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. 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

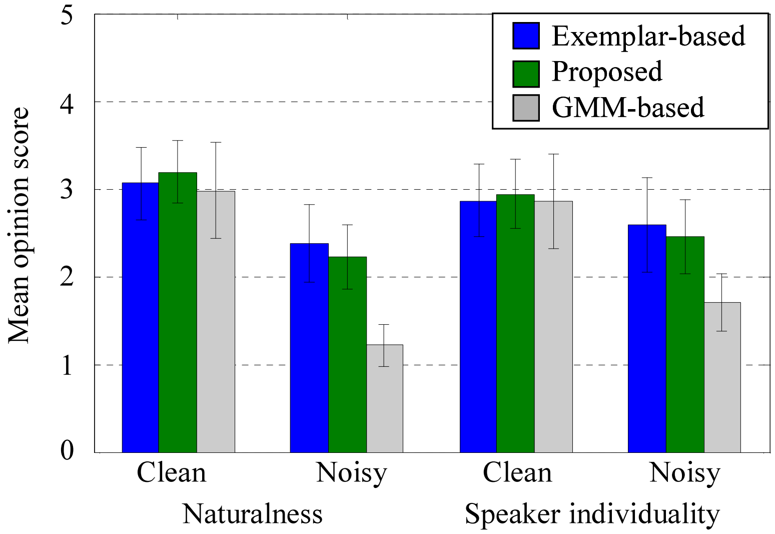


Figure 4.2: Mean opinion scores (MOS) for each method

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 the GMM-based method when evaluating noisy speech. The performances of the proposed method were slightly lower than those 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 the GMM-based method. While our proposed method took about 30 seconds to convert the speech features for 1 sentence, the GMM-based method took 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.

In this paper, the source and target dictionaries are estimated separately. We can estimate the source and target dictionaries simultaneously by using the joint vector of the source and the target features just as is done in the conventional GMM-based VC. However, the performance of that method, which was evaluated experimentally, is worse than our proposed method. We will also try to improve the performance of that method. In [64] we have proposed multimodal-VC and we try to improve our proposed method by applying multimodal features.

In [65], exemplar-based VC using temporal information is proposed. 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.

Chapter 5

Voice Conversion for Articulation Disorders Using Phoneme-categorized Exemplars

The related publications for this chapter are [27, 66, 67, 68].

5.1 The Motivation and Related Work

5.1.1 Motivation

An articulation disorder is a speech impediment in which a person has difficulty producing speech sounds or phonemes correctly. Articulation disorders are classified into three categories. Organic articulation disorders are mainly caused by hearing loss or mouth injuries; motor speech disorders are caused by motor paralysis; and functional articulation disorders are any articulation disorders that do not fit well in either of the other two categories.

In this study, we focused on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. Cerebral palsy is a non-progressive disorder of movement, and most cerebral palsy sufferers are born with the athetoid type. About two babies in 1,000 are born with cerebral palsy [69]. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. Three general times are given for the onset of the disorder:

before birth, at the time of delivery, and after birth. Cerebral palsy is classified into the following types: 1) spastic, 2) athetoid, 3) ataxic, 4) atonic, 5) rigid, and a mixture of these types [70].

Athetoid symptoms develop in about 10-15% of cerebral palsy sufferers [69]. In the case of a person with this type of articulation disorder, his/her movements are sometimes more unstable than usual. That means their utterances (especially their consonants) are often unstable or unclear due to the athetoid symptoms. 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.

In [71], we proposed robust feature extraction based on PCA (Principal Component Analysis) with more stable utterance data instead of DCT. In [72], we used multiple acoustic frames (MAF) as an acoustic dynamic feature to improve the recognition rate of a person with an articulation disorder, especially in speech recognition using dynamic features only. In spite of these efforts, the recognition rate for articulation disorders is still lower than that of physically unimpaired persons. The recognition rate using a speaker-independent model which is trained by well-ordered speech, is 3.5% [71]. This recognition rate suggests that for people who have not communicated with a person with athetoid cerebral palsy, it will be very hard for them to understand what that person is trying to say.

In this chapter, we propose a voice conversion (VC) method for articulation disorders. VC is a hands-free system that converts one's voice to another's voice. In recent years, people with an articulation disorder may use slideshows and a previously synthesized voice when they give a lecture. However, because their movement is restricted by their athetoid symptoms, it is hard for them to make slides or synthesize their voice in advance. People with articulation disorders desire a VC system that converts their voice into a clear voice that preserves their voice's individuality. However, a speech conversion method for people with articulation disorders resulting from athetoid cerebral palsy has not been successfully developed.

In [66], we proposed individuality-preserving VC for articulation disorders. In our VC, 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 expressed with a sparse representation of the source exemplars using NMF. By replacing a source speaker's exemplar with a target speaker's exemplar, the original speech spectrum is replaced with the target speaker's spectrum. People with articulation disorders wish to communicate by their own voice if they can; therefore, we proposed a combined-dictionary that consists of a source speaker's vowels and target speaker's well-ordered consonants. In the voice of a person with an articulation disorder, their consonants are often unstable and that makes their voices unclear. Their vowels are relatively stable compared to their consonants. Hence, by replacing the articulation-disordered basis of consonants only, a voice with an articulation disorder is converted into a non-disordered voice that preserves the individuality of the speaker's voice.

In this chapter, we propose advanced individuality-preserving VC using NMF. In order to avoid a mixture of the source and target spectra in a converted phoneme, we applied a phoneme-categorized dictionary and a dictionary selection method to our VC using NMF. In conventional NMF-based VC, the number of dictionary frames becomes large because the dictionary holds all the training exemplar frames. Therefore, it may cause phoneme mismatching between input signals and selected exemplars, and some frames of converted spectra are mixed with the source and target spectra. In this paper, a training exemplar is divided into a phoneme-categorized sub-dictionary, and an input signal is converted by using the selected sub-dictionary. The effectiveness of this method was confirmed by comparing it with the conventional NMF-based method and the conventional GMM-based method.

5.1.2 Related Work

In recent years, a number of assistive technologies using information processing have been proposed. For example, sign language recognition using image recognition technology [73, 74, 75], text reading systems from natural scene images [76, 77, 78], and the design of wearable speech synthesizers [79].

In the field of assistive technology, Nakamura et al. [8, 80] proposed GMM-based VC systems that reconstruct a speaker's individuality in electrolaryngeal speech and speech recorded by NAM microphones. These systems are effective for electrolaryngeal speech and speech recorded by NAM microphones however, because these statistical approaches are mainly proposed for speaker conversion, the target speaker's individuality will be changed to the source speaker's individuality. People with articulation disorders wish to communicate by their own voice if they can and there is a needs for individuality-preserving VC.

Text-to-speech synthesis (TTS) is similar voice application to VC. Veaux et al. [81] used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting from Amyotrophic Lateral Sclerosis (ALS). Yamagishi et al. [82] proposed a project which is named "Voice Banking and reconstruction". In that project, various types of voices are collected and they proposed TTS for ALS using that database. The difference between TTS and VC is that TTS needs text input to synthesize speech however VC does not need text input. In the case of people with articulation disorders resulting from athetoid cerebral palsy, it is difficult for them to input text because of their athetoid symptoms.

Maier et. al. [83] proposed automatic speech recognition systems for the evaluation of speech disorders resulting from head and neck cancer. ASR-based home appliance control system for articulation disorders has also been proposed [84].

5.2 Proposed Method

5.2.1 Phoneme-categorized Dictionary

Fig. 5.1 shows how to construct the sub-dictionary. $\mathbf{A}^s$ and $\mathbf{A}^t$ imply the source and target dictionary which hold all the bases from training data. These dictionaries are divided into K dictionaries. In this paper, the dictionaries are divided into 10 categories according to the Japanese phoneme categories shown in Table 5.1.

In order to select the sub-dictionary, a "categorizing-dictionary", which consists of the representative vector from each sub-dictionary, is constructed. The representative vectors for each phoneme category consist of the mean vectors of the Gaussian Mixture Model (GMM).

$$p(\mathbf{x}_n^{(k)}) = \sum_{m=1}^{M_k} \alpha_m^{(k)} N(\mathbf{x}_n^{(k)}, \boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)}) \tag{5.1}$$

M_k, $\alpha_m^{(k)}$, $\boldsymbol{\mu}_m^{(k)}$ and $\boldsymbol{\Sigma}_m^{(k)}$ represent the number of the Gaussian mixture, the weights of mixture, mean and variance of the m-th mixture of the Gaussian, in the k-th sub-dictionary, respectively. Each parameter is estimated by using an EM algorithm.

The basis of the categorizing-dictionary, which corresponds to the k-th sub-dictionary $\boldsymbol{\Phi}_k^s$, is represented using the estimated phoneme GMM as follows:

$$\boldsymbol{\theta}_k = [\boldsymbol{\mu}_1^{(k)}, \dots, \boldsymbol{\mu}_{M_k}^{(k)}] \tag{5.2}$$

$$\boldsymbol{\Phi}_k^s = [\mathbf{x}_1^{(k)}, \dots, \mathbf{x}_{N_k}^{(k)}] \tag{5.3}$$

N_k represents the number of frames of the k-th sub-dictionary. The categorizing-dictionary $\boldsymbol{\Theta}$ is given as follows:

$$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K] \tag{5.4}$$

5.2.2 Dictionary Selection and Voice Conversion

Fig. 5.2 shows the flow of the dictionary selection and VC. The input spectral features $\mathbf{X}^s$ are represented by a linear combination of bases from the categorizing-dictionary $\boldsymbol{\Theta}$. The weights of the bases are represented as activities $\mathbf{H}_{\boldsymbol{\Theta}}^s$.

$$\mathbf{X}^s \approx \boldsymbol{\Theta}\mathbf{H}_{\boldsymbol{\Theta}}^s \quad s.t. \quad \mathbf{H}_{\boldsymbol{\Theta}}^s \geq 0 \tag{5.5}$$

$$\mathbf{X}^s = [\mathbf{x}_1^s, \dots, \mathbf{x}_L^s] \tag{5.6}$$

$$\mathbf{H}_{\boldsymbol{\Theta}}^s = [\mathbf{h}_{\boldsymbol{\Theta}1}^s, \dots, \mathbf{h}_{\boldsymbol{\Theta}L}^s] \tag{5.7}$$

$$\mathbf{h}_{\boldsymbol{\Theta}l}^s = [\mathbf{h}_{\boldsymbol{\theta}_1 l}^s, \dots, \mathbf{h}_{\boldsymbol{\theta}_K l}^s]^T \tag{5.8}$$

$$\mathbf{h}_{\boldsymbol{\theta}_k l}^s = [h_{\boldsymbol{\theta}_1 l}^s, \dots, h_{\boldsymbol{\theta}_{M_k} l}^s]^T \tag{5.9}$$

Then, the l-th frame of input feature $\mathbf{x}_l^s$ is represented by a linear combination of bases from the sub-dictionary of the source speaker. The sub-dictionary $\boldsymbol{\Phi}_{\hat{k}}^s$,

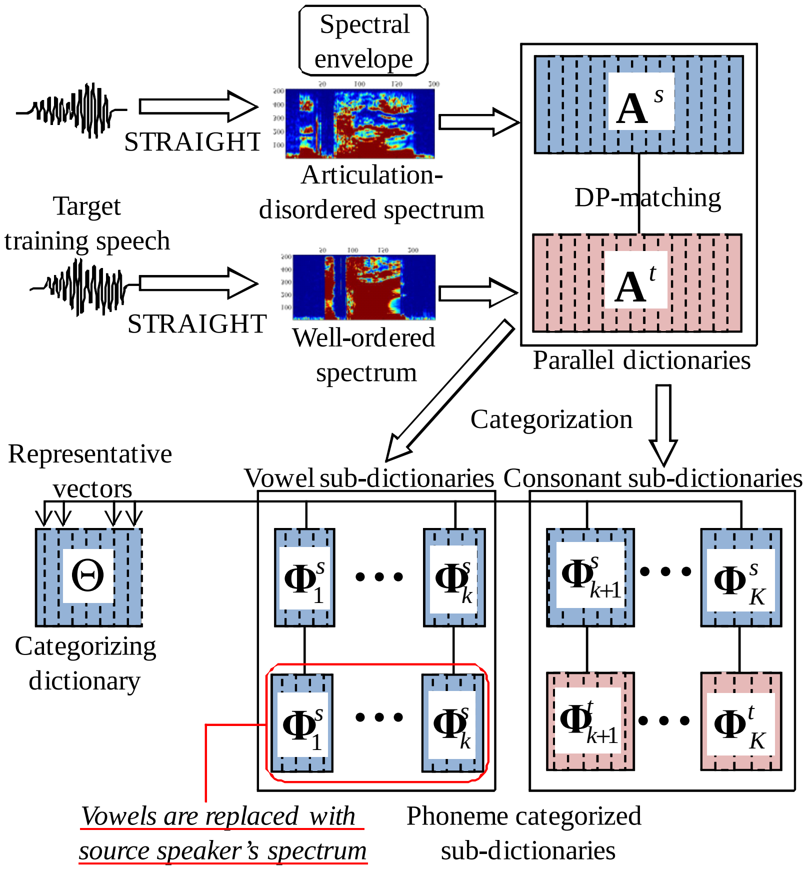


Figure 5.1: Making a sub-dictionary

which corresponds to $\mathbf{x}_l$, is selected as follows:

$$\begin{aligned} \hat{k} &= \arg\max_k \mathbf{1}^{1\times M_k}\mathbf{h}^s_{\boldsymbol{\theta}_k l} \\ &= \arg\max_k \sum_{m=1}^{M_k} h^s_{\boldsymbol{\theta}_m l} && (5.10) \\ \mathbf{x}_l &= \mathbf{\Phi}^s_{\hat{k}}\mathbf{h}_{\hat{k},l} && (5.11) \end{aligned}$$

The activity $\mathbf{h}_{l,\hat{k}}$ in Eq. (5.11) is estimated from the selected source speaker sub-dictionary.

If the selected sub-dictionary $\mathbf{\Phi}^s_{\hat{k}}$ is related to consonants, the l-th frame of the converted spectral feature $\hat{\mathbf{y}}_l$ is constructed by using the activity and the sub-dictionary of the target speaker $\mathbf{\Phi}^t_{\hat{k}}$.

$$\hat{\mathbf{y}}_l = \mathbf{\Phi}^t_{\hat{k}}\mathbf{h}_{\hat{k},l} \tag{5.12}$$

On the other hand, if the selected sub-dictionary $\mathbf{\Phi}^s_{\hat{k}}$ is related to vowels, the l-th frame of the converted spectral feature $\hat{\mathbf{y}}_l$ is constructed by using the activity and the sub-dictionary of the source speaker $\mathbf{\Phi}^s_{\hat{k}}$.

$$\hat{\mathbf{y}}_l = \mathbf{\Phi}^s_{\hat{k}}\mathbf{h}_{\hat{k},l} \tag{5.13}$$

5.3 Experiments

5.3.1 Experimental Conditions

The proposed VC technique was evaluated by comparing it with the conventional NMF-based method [66] (referred as the “sample-based method” in this paper) and the conventional GMM-based method [3] using clean speech data. We recorded 432 utterances (216 words, repeating each two times) included in the ATR Japanese speech database [61]. The speech signals were sampled at 12 kHz and windowed with a 25-msec Hamming window every 10 msec. A physically unimpaired Japanese male in the ATR Japanese speech database, was chosen as a target speaker.

Table 5.1: Japanese phoneme categories

	Category	phoneme
vowels	a	a
	e	e
	i	i
	o	o
	u	u
consonants	plosives	Q, b, d, dy, g, gy, k, ky, p, t
	fricatives	ch, f, h, hy, j, s, sh, ts, z
	nasals	m, my ny, N
	semivowels	w,y
	liquid	r, ry

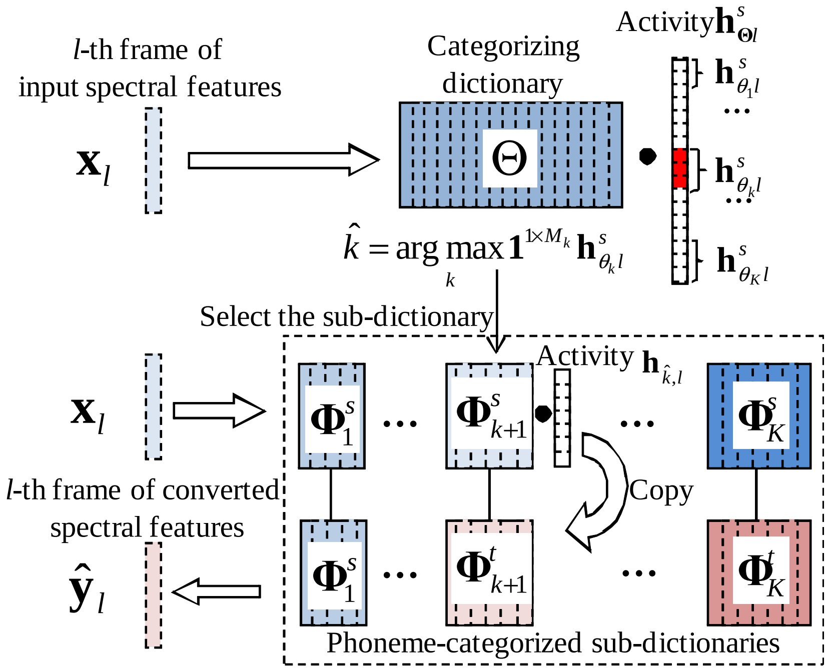

Activity $\mathbf{h}^{s}_{\Theta l}$
l-th frame of
input spectral features
Categorizing
dictionary
$\mathbf{x}_l$
Θ
$\mathbf{h}^{s}_{\theta_1 l}$
$\mathbf{h}^{s}_{\theta_k l}$
$\mathbf{h}^{s}_{\theta_K l}$
$\hat{k} = \arg\max_k \mathbf{1}^{1\times M_k} \mathbf{h}^{s}_{\theta_k l}$
Select the sub-dictionary
Activity $\mathbf{h}_{\hat{k},l}$
$\mathbf{x}_l$
$\mathbf{\Phi}^{s}_{1}$
$\mathbf{\Phi}^{s}_{k+1}$
$\mathbf{\Phi}^{s}_{K}$
Copy
l-th frame of converted
spectral features
$\hat{\mathbf{y}}_l$
$\mathbf{\Phi}^{s}_{1}$
$\mathbf{\Phi}^{t}_{k+1}$
$\mathbf{\Phi}^{t}_{K}$
Phoneme-categorized sub-dictionaries

Figure 5.2: NMF-based voice conversion using categorized dictionary

In the proposed and sample-based methods, the number of dimensions of the spectral feature is 2,565. It consists of 513 dimensional STRAIGHT spectrum [35] and its consecutive frames (the 2 frames coming before and the 2 frames coming after). The Gaussian mixture, which is used to construct categorizing-dictionary, is 1/500 of the number of bases of each sub-dictionary. The number of iterations for estimating the activity in proposed and sample-based methods was 300. In the conventional GMM-based method, MFCC+ΔMFCC+$\Delta\Delta$MFCC is used as a spectral feature. Its number of dimensions is 74. The number of Gaussian mixture is set to 64, which is experimentally selected.

In this paper, F0 information is converted using a conventional linear regression based on the mean and standard deviation [19]. The other information such as aperiodic components is synthesized without any conversion.

5.3.2 Objective Evaluations

5.3.2.1 Weights of Activities

Table 5.2 shows the rates of activities in each category using a combined-dictionary. The rows of the table show the target categories of each frame and the columns show the estimated categories. The total classification rate for the correct category is 0.22. This result indicates that in case where a combined-dictionary was used, most of the input spectra frames are decomposed into the bases from incorrect phoneme categories.

5.3.2.2 Phoneme Classification

We conducted phoneme classification using a categorizing-dictionary. Input spectra are categorized by using Eq. (5.11).

Table 5.3 shows the classification rates using the categorizing-dictionary, in which the number of bases is fixed. The second row of the table (Exemplar) shows the classification rates using all training data bases.

Table 5.4 shows the classification rates using the categorizing-dictionary, in which numbers of bases is changed. These tables show the effectiveness of the categorizing dictionary. As can be seen in Table 5.4, changing the number of

Tar. Categ.	a	e	i	o	u	plos.	fric.	nas.	semi.	liq.
a	**0.18**	0.05	0.07	0.07	0.08	0.30	0.14	0.07	0.03	0.02
e	0.05	**0.15**	0.14	0.06	0.08	0.23	0.14	0.07	0.02	0.04
i	0.07	0.08	**0.30**	0.05	0.09	0.20	0.09	0.07	0.04	0.03
o	0.05	0.02	0.04	**0.29**	0.10	0.25	0.15	0.06	0.02	0.02
u	0.05	0.04	0.08	0.08	**0.15**	0.26	0.19	0.09	0.03	0.03
plos.	0.05	0.04	0.07	0.08	0.10	**0.34**	0.15	0.09	0.02	0.03
fric.	0.07	0.07	0.08	0.11	0.08	0.25	**023**	0.06	0.01	0.01
nas.	0.05	0.04	0.09	0.10	0.13	0.26	0.13	**0.15**	0.02	00.0
semi.	0.13	0.04	0.06	0.21	0.12	0.20	0.09	0.09	**0.04**	0.02
liq.	0.07	0.08	0.15	0.07	0.13	0.20	0.14	0.07	0.03	**0.15**

Table 5.2: Confusion matrix of activity estimation

Category	Exemplar		Dictionary A		Dictionary B	
	#basis	rate	#basis	rate	#basis	rate
a	8400	0.49	1	0.22	3	0.07
e	5301	0.34	1	0.60	3	0.52
i	5487	0.59	1	0.43	3	0.30
o	10648	0.66	1	0.73	3	0.73
u	8289	0.25	1	0.12	3	0.48
plosives	9917	0.54	1	0.10	3	0.27
fricatives	8343	0.44	1	0.62	3	0.45
nasals	6435	0.19	1	0.78	3	0.72
semivowels	1325	0.00	1	0.04	3	0.27
liquid	1790	0.10	1	0.33	3	0.30
total	65935	0.41	9	0.43	27	0.42

Table 5.3: Classification rates of each dictionary

bases in the categorizing-dictionary is effective. In these experiments, the most effective dictionary was Dictionary D.

Category	Dictionary C		Dictionary D		Dictionary E	
	#basis	rate	#basis	rate	#basis	rate
a	4	0.02	8	0.05	16	0.02
e	3	0.67	5	0.78	10	0.48
i	3	0.37	5	0.37	10	0.43
o	5	0.68	10	0.71	25	0.76
u	4	0.44	8	0.52	10	0.63
plosives	4	0.54	9	0.67	10	0.69
fricatives	4	0.53	8	0.53	10	0.53
nasals	3	0.39	6	0.23	10	0.16
semivowels	3	0.25	3	0.23	3	0.13
liquid	3	0.31	3	0.29	3	0.02
total	36	0.44	65	0.47	107	0.44

Table 5.4: Classification rates of categorizing dictionary

Table 5.5 shows the confusion matrix of the phoneme classification using Dictionary D. Category “a”, “nasals”, “semivowels” and “liquid” are not well classified. The other categories had a greater than 50% classification rate.

5.3.2.3 Discussion

Phoneme category classification rates using all the training data bases was 0.22. This result suggests that, without sub-dictionary selection, a frame of input data is decomposed into many bases from different phoneme categories. We can avoid such a situation by using phoneme-categorized sub-dictionary selection because the input data frame is decomposed into bases from the selected sub-dictionary.

The effectiveness of using a categorizing dictionary was confirmed in our experiments. Sub-dictionaries were correctly selected in most categories. However, some categories were not well-classified. For example, the category “nasals” tended to be classified as “fricatives”. This may be because it is difficult for a person with an articulation disorder to move their tongue. In the case of “semivow-

Tar. Categ.	a	e	i	o	u	plos.	fric.	nas.	semi.	liq.
a	**0.05**	0.13	0.07	0.11	0.02	0.31	0.15	0.02	0.17	0.01
e	0	**0.78**	0.03	0.04	0.02	0.09	0.00	0.00	0.02	0.03
i	0	0.25	**0.37**	0.08	0.07	0.11	0.01	0.01	0.08	0.04
o	0	0.02	0.01	**0.71**	0.07	0.11	0.00	0.01	0.02	0.00
u	0	0.05	0.02	0.05	**0.52**	0.27	0.02	0.02	0.02	0.05
plos.	0.00	0.02	0.03	0.02	0.04	**0.67**	0.00	0.13	0.03	0.03
fric.	0	0.02	0.06	0.03	0.02	0.29	**0.53**	0.01	0.01	0.03
nas.	0	0.03	0.04	0.04	0.14	0.46	0.00	**0.23**	0.03	0.03
semi.	0	0.08	0.07	0.26	0.11	0.21	0.00	0.00	**0.24**	0.02
liq.	0	0.12	0.02	0.09	0.09	0.25	0.04	0.04	0.06	**0.29**

Table 5.5: Confusion matrix of dictionary D

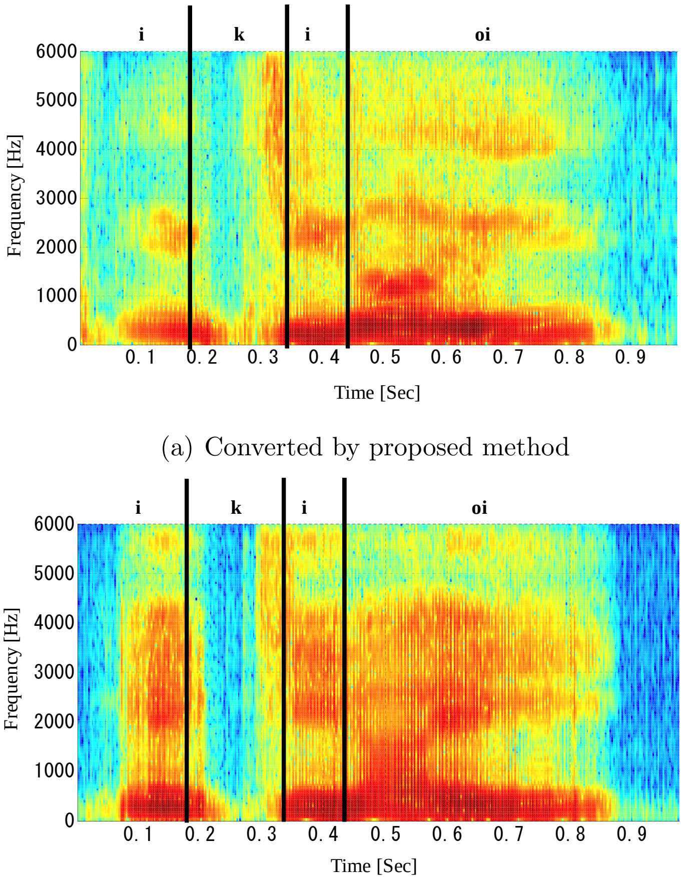
els" and "liquid", misclassification occurred because the number of training exemplars is too small to categorize the utterances correctly. The classification rate of "a" is significantly low compared to other categories. Compared to other vowel phonemes, "a" is easy to articulate. Because of this, the vague articulation of a person with an articulation disorder may become similar to "a".

5.3.3 Subjective Evaluations

5.3.3.1 Listening Tests

We conducted subjective evaluation on 3 topics. A total of 10 Japanese speakers took part in the test using headphones. For the "listening intelligibility" evaluation, we performed a MOS (Mean Opinion Score) test [63]. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Twenty-two words, which are difficult for a person with an articulation disorder to utter, were evaluated. The subjects were asked about the listening intelligibility in the articulation-disordered voice, the voice converted by our proposed method, and the GMM-based converted voice.

On the "similarity" evaluation, the XAB test was carried out. In the XAB test, each subject listened to the articulation disordered voice. Then the subject listened to the voice converted by the two methods and selected which sample sounded most similar to the articulation disordered voice. On the “naturalness” evaluation, a paired comparison test was carried out, where each subject listened to pairs of speech converted by the two methods and selected which sample sounded more natural.



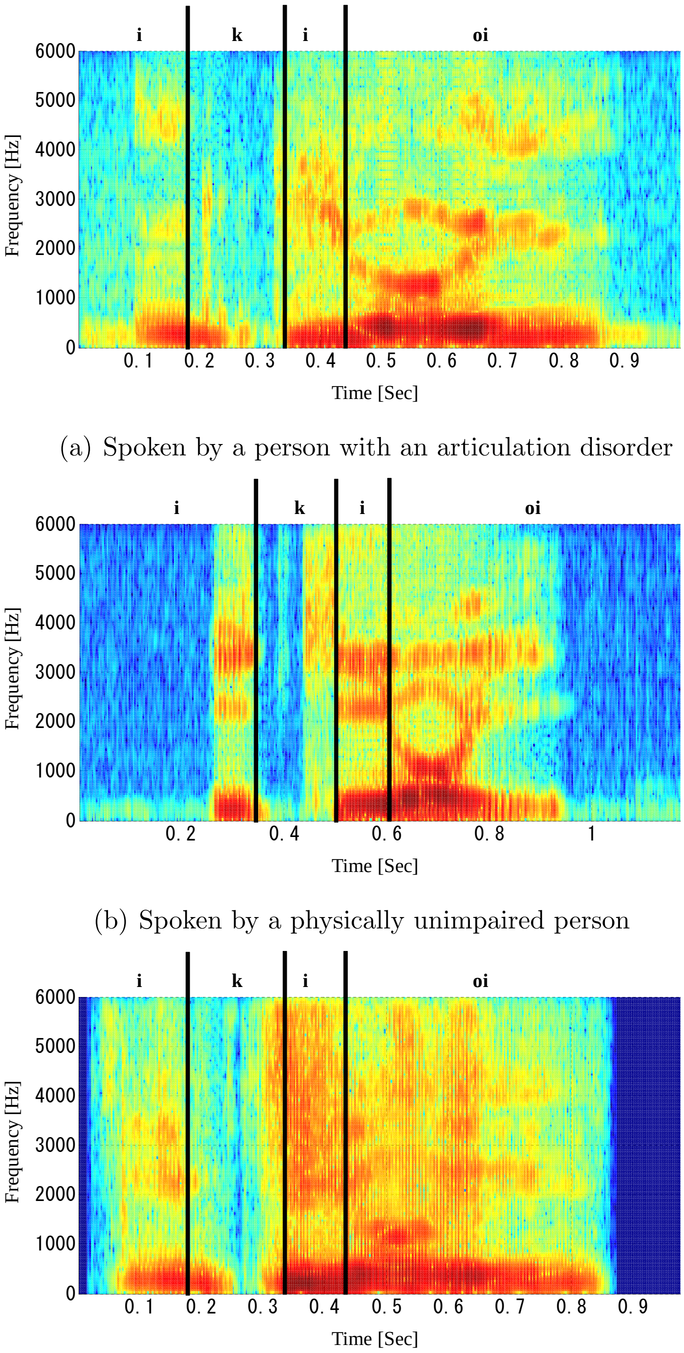
(a) Converted by proposed method

(b) Converted by GMM-based VC

Figure 5.3: Examples of converted spectrogram //i k i oi

5.3.3.2 Results and Discussion

Fig. 5.3(a) and 5.3(b) show examples of converted spectrogram using our proposed method and the conventional GMM-based method, respectively. In Fig. 5.3(a), there are less misconversions in vowel part compared to Fig. 5.4(c). Moreover, by



(a) Spoken by a person with an articulation disorder

(b) Spoken by a physically unimpaired person

(c) Converted by NMF-based VC

Figure 5.4: Examples of spectrogram //i k i oi

using GMM-based conversion, the area labeled “oi” becomes unclear compared to NMF-based conversion.

Fig. 5.5 shows the results of the MOS test for listening intelligibility. The error bars show a 95% confidence score. Our proposed VC can improve the listening intelligibility and clarity of consonants. On the other hand, GMM-based conversion can improve the clarity of consonants, but it deteriorates the listening intelligibility. This is because GMM-based conversion has the effect of noise resulting from measurement error. Proposed VC also has the effect of noise, but it is less than that created by GMM-based conversion.

Fig. 5.6 shows the results of the XAB test on the similarity to the source speaker and naturalness of the converted voice. The error bars show a 95% confidence score. Our proposed VC obtained a higher score than Sample-based and GMM-based conversion on similarity. Fig. 5.7 shows the preference score on the naturalness. The error bars show a 95% confidence score. Our proposed VC also obtained a higher score than Sample-based and GMM-based conversion on naturalness.

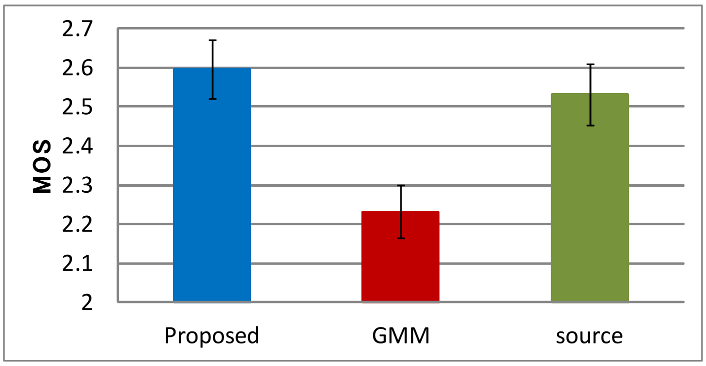


Figure 5.5: Results of MOS test on listening intelligibility

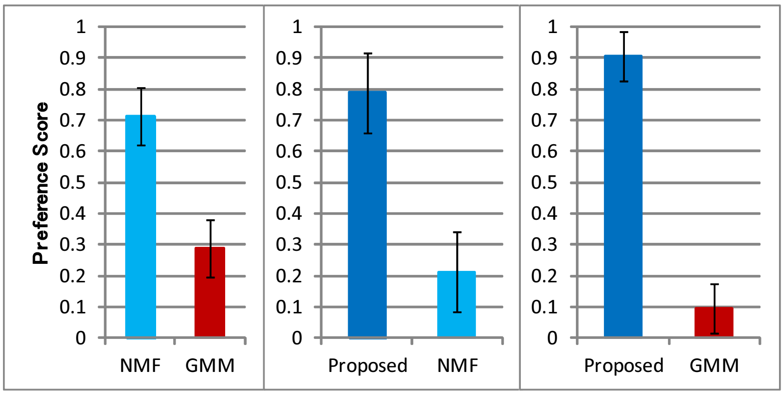


Figure 5.6: Preference scores for the individuality

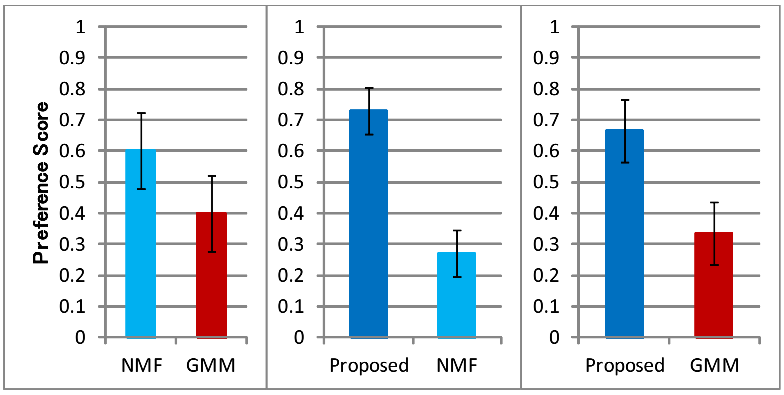


Figure 5.7: Preference scores for the naturalness

5.4 Chapter Conclusions

We proposed a spectral conversion method based on NMF for a voice with an articulation disorder. Our proposed method introduced a dictionary-selection method to conventional NMF-based VC. The results of our experiments demonstrated that our VC method can improve the listening intelligibility of words uttered by a person with an articulation disorder. Moreover, compared to conventional GMM-based VC and conventional NMF-based VC, our proposed VC can preserve the individuality of the source speaker's voice and the naturalness of the voice.

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. In addition, the structure of a phoneme-categorized sub-dictionary is not completely suitable for a person with an articulation disorder. We will examine other phoneme categories that are more suitable to a person with an articulation disorder.

Regarding computational times, the proposed method requires higher computation times than a GMM-based method. In [26], we proposed a novel NMF-based VC method that has lower computation times than exemplar-based VC. By using this method, we can reduce the computational time of VC for articulation disorders.

Future work will also include efforts to study the co-articulation effect between phonemes. In order to preserve voice individuality, our VC method retains their vowel. Solving this problem in the vowel portion will be a focus of our future work.

Chapter 6

Small-parallel Exemplar-based Voice Conversion Using Affine-NMF

The related publications for this chapter are [85, 86].

6.1 The Motivation and Related Work

6.1.1 Motivation

The need to have a large amount of parallel data is a large hurdle for the practical use of VC. For example, the conventional GMM-based VC method requires 50 parallel sentences or 216 parallel words between the source and the target speakers. In recent years, some statistical approaches that do not require parallel training data have been proposed [87, 88, 89, 90]. In this chapter, we propose noise-robust VC using a small parallel corpus based on an NMF-based speaker adaptation technique.

In [91], adaptation of speaker-specific bases in NMF for single channel speech-music separation has been presented. In this framework, speaker-specific bases are adapted to the other speaker using an affine matrix. We call this method Affine NMF (A-NMF) and apply it to VC. In VC, the source dictionary is constructed using sufficient source speaker data, and it is adapted using a small amount of parallel data (about ten words only) in order to obtain the target dictionary, where a linear regression transformation matrix (affine matrix) is trained based on NMF.

The contributions of this paper are summarized in two points. First, we have decreased the total amount of parallel training data required for NMF-based VC. Conventional NMF-based VC requires 216-word parallel data for dictionary construction. However, experimental results using our proposed approach, which requires only a small amount of parallel data, demonstrate a the conversion quality that is almost the same as that of conventional NMF-based VC. The second contribution is that there needs to be no concern about differences between parallel dictionaries.

6.1.2 Related Work

Statistical VCs require a large-volume parallel corpus of the source and target speakers. In this paper, “parallel” means that the texts of the corpus of the source speaker are the same as those of the target speaker. This can be a difficult requirement to meet in practice. Within the GMM-based framework, several approaches have been proposed that relax this constraint. Non-parallel training based on maximum a posteriori probability (MAP) was proposed in [87]. Mouchtaris *et al.* [88] proposed non-parallel training for GMM-based VC using ML constrained parameter adaptation. These methods require utterances by both source and target speakers, however the texts used do not have to be parallel.

6.2 Proposed Method Using Affine-NMF

In the framework of conventional NMF-based VC which is described in chapter 3, a large-parallel corpus of source and target speakers is needed for dictionary construction. In this section, we propose target dictionary estimation from a small-parallel corpus only.

Fig. 6.1 shows the estimation procedure of our proposed method. $\mathbf{X}^s$ and $\mathbf{X}^t$ show a small amount of parallel data between source and target speakers. In the Activity Estimation stage, a source spectral exemplar matrix $\mathbf{X}^s$ is decomposed into a linear combination of bases from the source dictionary $\mathbf{A}^s$. The source dictionary consists of the source speaker's exemplars. It is constructed the same way the dictionary constructed when using conventional NMF-based VC, as explained in chapter 3.

In the Dictionary Adaptation stage, speaker adaptation is conducted in order to obtain a target dictionary from a source dictionary using a small amount of (parallel) target speech data. The adaptation is performed using a linear regression transformation matrix based on an NMF framework. Given the transformation matrix, $\mathbf{W}$, the target feature vector at the l-th frame is obtained as follows:

$$\mathbf{x}_l^t \approx \mathbf{W}\mathbf{A}^s\mathbf{h}_l^s \tag{6.1}$$

where $\mathbf{A}^s$ is the source dictionary and $\mathbf{h}_l^s$ is the activity vector of the source signal at the l-th frame.

In order to find the transformation matrix, an NMF framework, which minimizes the KL divergence between $\mathbf{X}^t$ and $\mathbf{W}\mathbf{A}^s\mathbf{H}^s$ is used.

$$\mathbf{W} = \underset{\mathbf{W}}{\operatorname{argmin}}\, d(\mathbf{X}^t, \mathbf{W}\mathbf{A}^s\mathbf{H}^s) \quad s.t. \quad \mathbf{W} \geq 0. \tag{6.2}$$

The transformation matrix, $\mathbf{W}$, is estimated using $\mathbf{A}^s$, $\mathbf{H}^s$ and a small amount of the parallel target speech data, $\mathbf{X}^t$, as follows:

$$\begin{aligned}\mathbf{W}_{n+1} \quad = \quad & \mathbf{W}_n . * ((\mathbf{X}^t./(\mathbf{W}_n(\mathbf{A}^s\mathbf{H}^s)))(\mathbf{A}^s\mathbf{H}^s)^\mathsf{T}) \\ & ./(\mathbf{1}^{(D\times L)}(\mathbf{A}^s\mathbf{H}^s)^\mathsf{T}).\end{aligned} \tag{6.3}$$

The new parallel target dictionary is given by $\hat{\mathbf{A}}^t = \mathbf{W}\mathbf{A}^s$.

In the test stage, the noisy input source speaker's spectra matrix $\mathbf{X}$ is decomposed into the multiplication of dictionary $\mathbf{A} = [\mathbf{A}^s\mathbf{A}^n]$ by its activity $\mathbf{H} = [\mathbf{H}^{s\mathsf{T}}\mathbf{H}^{n\mathsf{T}}]^\mathsf{T}$ as follows:

$$\mathbf{X} = \mathbf{A}\mathbf{H}. \tag{6.4}$$

The converted spectra matrix $\hat{\mathbf{X}}^t$ is constructed from the estimated target dictionary $\hat{\mathbf{A}}^t$, and the clean activity $\mathbf{H}^s$ as follows:

$$\hat{\mathbf{X}}^t = \hat{\mathbf{A}}^t\mathbf{H}^s. \tag{6.5}$$

In this method, we do not have to consider the difference between parallel dictionaries because the parallel utterances are used as adaptation data, not as a dictionary. The activity matrix estimated from the source dictionary contains both the phoneme information and speaker information of the input utterance.In this method, the adaptation matrix is estimated from the fixed source dictionary and source activity matrix, and target speaker information is extracted using the adaptation matrix in this procedure. In other words, the adaptation matrix is independent of the phoneme, and it is the conversion matrix from the source to the target speaker.

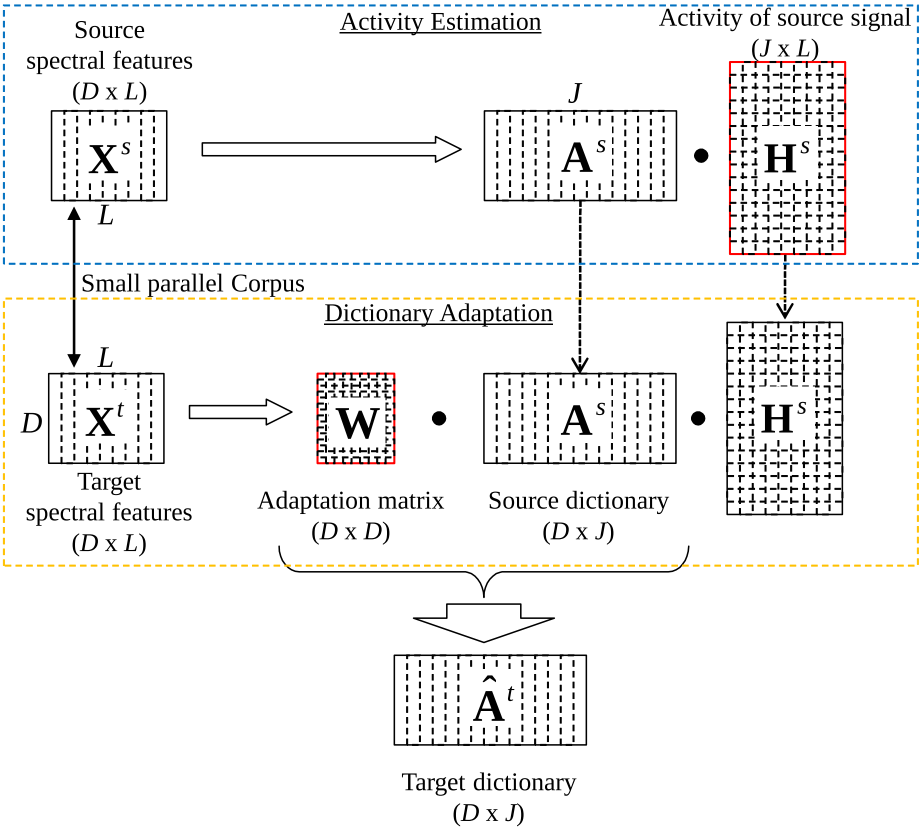


Figure 6.1: Estimation of parallel dictionary using a speaker transformation matrix

6.3 Experiments

6.3.1 Experimental Conditions

The new VC technique was evaluated by comparing it with conventional techniques based on GMM [3] and NMF [23] in a speaker conversion task using noisy speech data. Speaker MMY, MAU, MNM, FTK, FYN and FMS were selected from the ATR Japanese speech database [61], and we conducted male-to-female (MMY→FTK and MAU→FYN), male-to-male (MMY→MAU and MNM→MMY), and female-to-female (FTK→FYN and FMS→FTK) conversions. The sampling rate, frame shift, and window length are 8 kHz, 5 ms, and 25 ms, respectively.

We used 216 words of clean speech per each speaker to construct a source dictionary in NMF with a speaker adaptation and to train the GMM using the conventional method. Table 6.1 shows the average number of frames in each parallel dictionary. These training data were taken from the ATR Japanese speech database set A (all of task code B). The number of adaptation words was 10, 25, and 50 per each speaker. These adaptation words were randomly chosen from the ATR Japanese speech database set A (task code A). Fifty words, which are included in the ATR Japanese speech database set A (task code A) and are different from the training and adaptation data, were randomly chosen as test data.

The noisy speech signals were obtained by adding noise signals to clean speech data. We used three types of noise signal (restaurant, station, and exhibition) and these are randomly taken from the non-utterance section of CENSREC-1-C database [62]. The SNRs for each noise was set to 20 and 10 dB. They are added to a test word independently to each other. (A noisy speech signal includes one type of noise signal.) The average number of exemplars in the noise dictionary for each word was 104.

In objective evaluation, all 50 test words with three types of noisy signal at two different types of SNRs were converted. Therefore, total 1,800 words (6 pairs × 50 words × 3 noise types × 2 SNRs) were used for subjective evaluation. In subjective evaluation, the half of test data with restaurant-noise at 10 dB were used. Therefore, total amount of test words was 150 (6 pairs × 25 words) in subjective evaluation.

The spectrum, F0, and aperiodic components were extracted using STRAIGHT [92]. In the NMF-based method, a 513-dimensional spectrum extracted using STRAIGHT, was used as the feature vector in the input signal and source dictionary. The number of iterations used to estimate the activity was 300 [27]. The activity and the transformation matrix were initialized with non-negative random values. In the GMM-based method, 40 linear-cepstral coefficients obtained from the STRAIGHT [92] spectrum were used as the feature vectors. The number of Gaussian mixtures was 64 which was chosen to obtain minimum distortion on test data. In this study, F0 information was converted using conventional linear regression in all VC methods based on the mean and standard deviation [19] as follows:

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)}, \tag{6.6}$$

where $\mathbf{x}_t$ and $\hat{y}_t$ denote the log-scaled F0 of the source speaker and the converted word at frame t, respectively. $\mu^{(x)}$ and $\sigma^{(x)}$ denote the mean and standard deviation of the log-scaled F0, as calculated from source speaker's training data. $\mu^{(y)}$ and $\sigma^{(y)}$ are the mean and standard deviation of the target speaker data. We made no conversions to the aperiodic components. With STRAIGHT, we used converted spectra, F0, and source aperiodic components for synthesizing the target voice.

Table 6.1: Number of frames in each parallel dictionary

# training data	216	50	25	10
# frames	61,168	13,839	6,782	2,826

6.3.2 Experimental Results

Objective tests were carried out using the Normalized Spectrum Distortion (NSD) [93]

$$NSD = \sqrt{||\mathbf{X}^t - \hat{\mathbf{X}}^t||^2 / ||\mathbf{X}^t - \mathbf{X}^s||^2}, \tag{6.7}$$

where $\mathbf{X}^s$, $\mathbf{X}^t$, and $\hat{\mathbf{X}}^t$ denote the source, target, and converted spectrum, respectively.

Figs. 6.2 to 6.4 show the NSD for each speaker at 10 dB. "NMF" shows the result using conventional NMF without speaker adaptation and "A-NMF" shows the result using NMF with speaker adaptation. As shown in these figures, the performance of NMF without speaker adaptation decreases as the number of words used for the parallel dictionaries decreases. On the other hand, the performance of NMF with speaker adaptation does not decrease in comparison to the conventional NMF without speaker adaptation. Our A-NMF method obtained a better result than NMF when we used 216 parallel words for most speakers. We assume this to be due to the fact that the difference between parallel dictionaries degrades the performance of NMF.

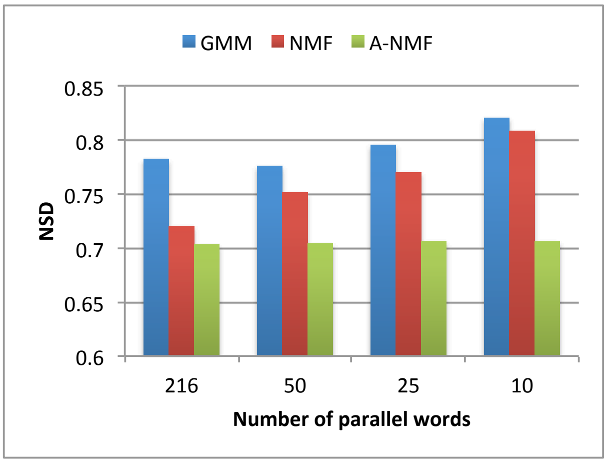


Figure 6.2: NSD for male-to-female conversion at 10 dB

Figs. 6.5 to 6.7 show the NSD for each speaker at 20 dB. Because of the low SNR conditions, the noise robustness of NMF-based VC is lower, compared to Figs. 6.2 to 6.4. However, the performance of NMF with speaker adaptation does not decrease as the number of words used for the parallel dictionaries decreases in comparison with the conventional NMF without speaker adaptation. Moreover, the performance of NMF with speaker adaptation is better than conventional GMM-based VC. These results show the effectiveness of our NMF-based speaker adaptation technique.

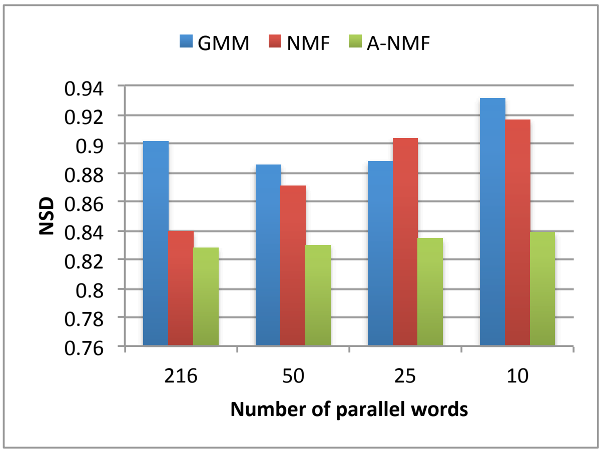


Figure 6.3: NSD for male-to-male conversion at 10 dB

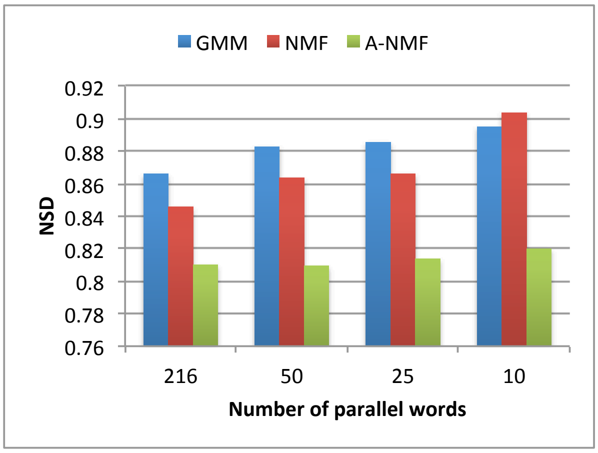


Figure 6.4: NSD for female-to-female conversion at 10 dB

For the speech quality evaluation, a mean opinion score (MOS) [63] test was performed. The opinion score was set to a five-point scale (5 excellent, 4 good, 3 fair, 2 poor, 1 bad). The number of participants was 8, and the SNR was 10 dB. Fig. 6.8 (left side) shows the MOS test on the speech quality. As shown in Fig. 6.8, NMF-based VC with speaker adaptation (25 adaptation words) obtained a better score than conventional NMF-based VC (25 words). The result was confirmed by a p-value test of 0.05.

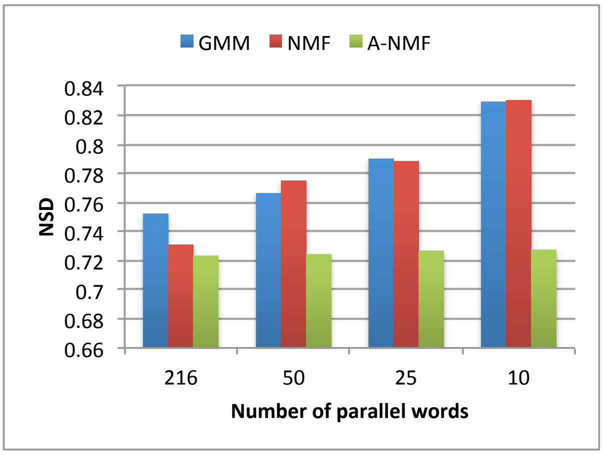


Figure 6.5: NSD for male-to-female conversion at 20 dB

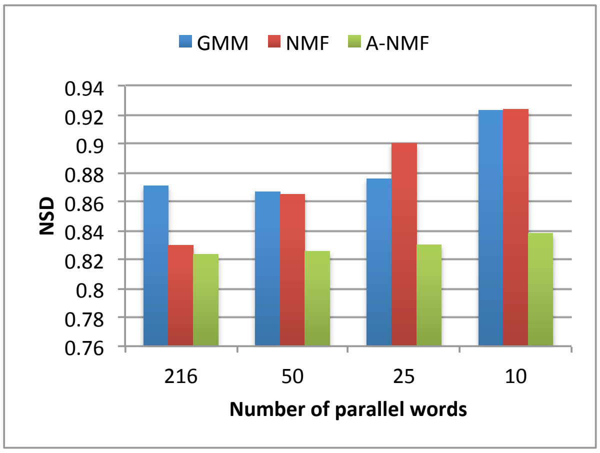


Figure 6.6: NSD for male-to-male conversion at 20 dB

For the evaluation of speaker individuality, an XAB test was carried out. In the XAB test, each participant listened to the target speech. The participant then listened to the speech converted by the two methods and selected the sample that sounded more similar to the target speech. Fig. 6.8 (right side) shows that the NMF-based VC with speaker adaptation obtained a higher score than the conventional NMF-based VC without speaker adaptation. We confirmed this result by a 0.05 p-value test.

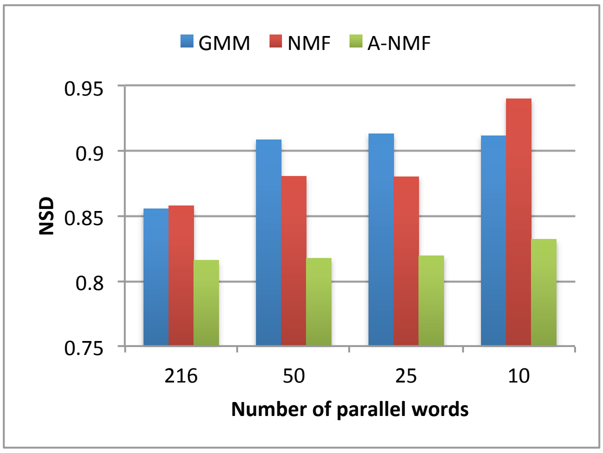


Figure 6.7: NSD for female-to-female conversion at 20 dB

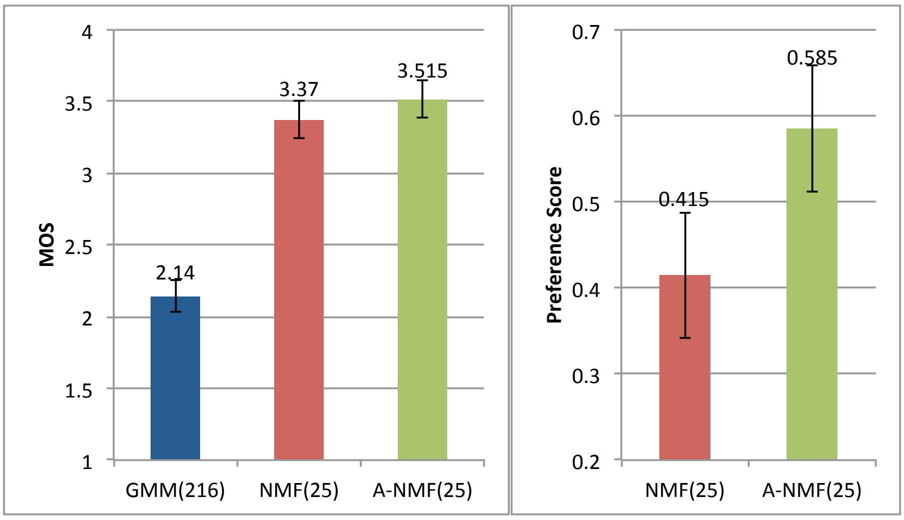


Figure 6.8: Results of MOS test and XAB test

6.4 Chapter Conclusions

In this chapter, an exemplar-based VC technique using speaker adaptation was presented. This method requires only a small amount of parallel data, where a linear regression transformation matrix is used to adapt a source dictionary to a target dictionary and it is estimated in an NMF framework. In comparison experiments between GMM-based VC, NMF without speaker adaptation, and NMF with speaker adaptation, the NMF-based VC with speaker adaptation showed better performance.

Some problems remain with this method. The proposed method requires higher computation times than the GMM-based method. In [59], we proposed a frame-work that reduces computational time for NMF-based VC. In future work, we will investigate the optimal number of bases and evaluate performance under other noise conditions. In addition, this method is limited to only one-to-one voice conversion because it requires a small amount of parallel data. Hence, we will research a method for many-to-many VC within this framework, and apply this method to other VC applications, such as assistive technology [67] or emotional VC [4].

Chapter 7

Many-to-many Voice Conversion Using Multi-NMF

The related publications for this chapter are [94, 95, 96].

7.1 The Motivation and Related Work

7.1.1 Motivation

Many-to-many VC is approaches that do not require the use of source or target speaker utterances as training data. In this approach, parallel training data from many speakers are used as training data and the source and the target speakers are represented as a linear combination of trained speakers.

Within the GMM-based framework, eigen-voice GMM (EV-GMM) [89] has been proposed for many-to-many VC. However, an exemplar-based approach without a parallel data set has not yet been demonstrated. Because exemplar-based approach enables to create natural-sounding converted voice [27], exemplar-based many-to-many VC method have been needed.

This chapter proposes multiple non-negative matrix factorization (Multi-NMF) which allows many-to-many VC to be applied. The parallel dictionaries that are needed in a conventional NMF-based VC are replaced by dictionaries that represent the voice of many speakers. In a conventional GMM-based many-to-many VC approach [97], the reference voice must be chosen from the training data, because the method combines many-to-one VC and one-to-many VC. Our NMF-based approach does not require this. Moreover, because our approach is exemplar-based, we assume our proposed VC method can create natural sounding voice compared to GMM-based VC method.

Experimental results show that the conversion quality of the proposed method is almost the same as that of conventional one-to-one VC, suggesting that it can be applied to voice quality control and noise-robust VC.

7.1.2 Related Work

Inspired by a speaker adaptation technique using hidden Markov models (HMM), Toda *et al.* [89] proposed eigen-voice GMM (EV-GMM). In this approach, the eigen-vector of an arbitrary speaker is estimated using parallel training data from many speakers. EV-GMM enables one-to-many VC and many-to-one VC. Saito *et al.* [90] used tensor representation to improve the performance of one-to-many GMM-based VC. Masuda *et al.* [98] proposed multistep VC, which was the first approach to many-to-many VC. Many-to-many VC using EV-GMM was proposed by Ohtani *et al.* [97] and EV-GMM has been expanded to voice quality control [99]. In these approach, many-to-one VC and one-to-many VC are combined by using a reference speaker's voice. They noted that the conversion quality was sensitive to the choice of reference speaker.

7.2 Proposed Method

Speaker conversion is a method which converts a source speaker's individual voice to that of a target speaker while preserving the linguistic information. Our proposed method is based on the following assumptions:

1. The spectra of an arbitrary speaker are represented by a linear combination of the bases of many speakers.

2. An activity matrix captures speaker-independent information.

Multi-NMF decomposes three factors (speaker weight vector, dictionary, and activity) from the observed spectra. Our approach is exemplar-based, and so the spectra of training data are used as the dictionary, and only the speaker weight vector and activity are estimated. Based on the above assumptions, the speaker weight vector captures the speaker individuality, while the activity captures the speaker-independent information. The detailed flow is described in the following subsections.

7.2.1 Dictionary Construction

A multi-speaker dictionary is constructed from all parallel training utterances of multiple speakers. Linear spectra are extracted from the training utterance as exemplars using STRAIGHT [36]. A pivot speaker is randomly chosen from multiple speakers and training exemplars of the other speakers are aligned with the exemplars of the pivot speaker by using DTW. Therefore, the dictionary is a 3-dimensional matrix that consists of (the number of dimensions of the spectrum) × (the number of frames) × (the number of speakers).

7.2.2 Flow of the Proposed Method

In our proposed method, the flow is different for inter-gender conversion and cross-gender conversion.

7.2.2.1 Inter-gender Conversion

Fig. 7.1 shows the flow of the proposed method. $\mathbf{V}^s$, $\mathbf{V}^t$, $\hat{\mathbf{V}}^t$, $\mathbf{a}^s$, $\mathbf{a}^t$, $\mathbf{H}^s$, and $\mathbf{H}^t$ denote the matrix of input source spectra, the matrix of the adaptation target speaker's spectra, the matrix of the converted spectra, the source speaker's weight vector, the target speaker's weight vector, the activity matrix of the source speaker, and the activity matrix of the target speaker, respectively. D, L, L', and J denote the number of dimensions of a spectrum, the frame of the source spectra, the frame of the adaptive spectra, and the frame of the dictionary, respectively. $\mathbf{W}^M \in \mathbb{R}^{(D \times J \times K)}$ denotes the dictionary matrix, which consists of the parallel exemplars of many speakers, where K is the number of speakers in the corpus. The superscript on $\mathbf{W}^M$ means that it combines the dictionaries of many speakers.

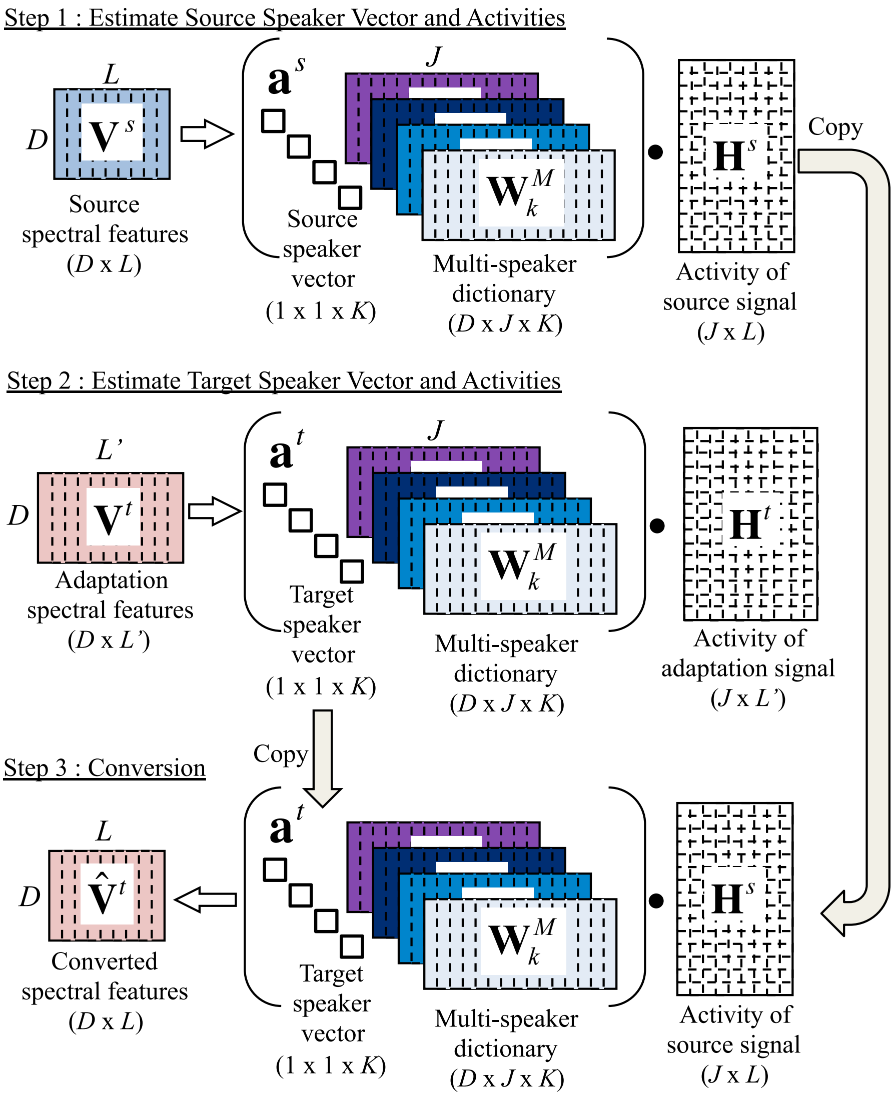


Figure 7.1: Many-to-many VC using Multi-NMF

In inter-gender conversion, the dictionary consists of either male only or female only spectra. The k-th speaker's dictionary is denoted by $\mathbf{W}_k^M \in \mathbb{R}^{(D \times J)}$.

First, the matrix of input source spectra $\mathbf{V}^s$ is represented as follows, based on Assumption 1,

$$\mathbf{V}^s \approx \left(\sum_{k=1}^{K} a_k^s \mathbf{W}_k^M \right) \mathbf{H}^s \tag{7.1}$$

where a_k^s denotes the k-th element of $\mathbf{a}^s$. We stress that each speaker's dictionary is multiplied by the same activity matrix element of $\mathbf{H}^s$ in (7.1).

Next, sample frames of the target speaker spectra $\mathbf{V}^t$ are used as adaptive spectra, and the target speaker's weight vector and the activity matrix for the adaptive data are estimated as $\mathbf{a}^t$ and $\mathbf{H}^t$, respectively.

$$\mathbf{V}^t \approx \left(\sum_{k=1}^{K} a_k^t \mathbf{W}_k^M \right) \mathbf{H}^t \tag{7.2}$$

Finally, the converted spectra $\hat{\mathbf{V}}^t$ are constructed from the estimated target speaker's weight vector $\mathbf{a}^t$ and the source speaker's activity matrix $\mathbf{H}^s$, based on Assumption 2.

$$\hat{\mathbf{V}}^t = \left(\sum_{k=1}^{K} a_k^t \mathbf{W}_k^M \right) \mathbf{H}^s \tag{7.3}$$

7.2.2.2 Cross-gender Conversion

In order to reduce computational time and memory usage, two dictionaries are used in the case of cross-gender conversion. The dictionary of source gender is represented as $\mathbf{W}^{M_s} \in \mathbb{R}^{(D \times J \times K_s)}$ and the dictionary of target gender is represented as $\mathbf{W}^{M_t} \in \mathbb{R}^{(D \times J \times K_t)}$. K_s and K_t denote the number of speakers whose samples are included in the source and target dictionaries.

First, the activity matrix is estimated from the source spectra and source dictionary as follows, based on Assumption 1,

$$\mathbf{V}^s \approx \left(\sum_{k=1}^{K_s} a_k^s \mathbf{W}_k^{M_s} \right) \mathbf{H}^s \tag{7.4}$$

Next, the target speaker's weight vector and the activity matrix for adaptive data are estimated from adaptive data and the target dictionary.

$$\mathbf{V}^t \approx \left(\sum_{k=1}^{K_t} a_k^t \mathbf{W}_k^{M_t} \right) \mathbf{H}^t \tag{7.5}$$

Finally, the converted spectra $\hat{\mathbf{V}}^t$ are constructed from the estimated target speaker's weight vector, the source speaker's activity matrix, and the target dictionary based on Assumption 2.

$$\hat{\mathbf{V}}^t = \left(\sum_{k=1}^{K_t} a_k^t \mathbf{W}_k^{M_t} \right) \mathbf{H}^s \tag{7.6}$$

7.2.3 Multi-NMF

We are proposing Multi-NMF, which estimates a speaker vector $\mathbf{a} \in \mathbb{R}^{(1\times 1\times K)}$ and an activity matrix $\mathbf{H} \in \mathbb{R}^{(J\times L)}$ from input spectra $\mathbf{V} \in \mathbb{R}^{(D\times L)}$, given a dictionary $\mathbf{W}^M \in \mathbb{R}^{(D\times J\times K)}$. The cost function of Multi-NMF is defined as follows:

$$\begin{aligned} & d(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}) + \lambda ||\mathbf{H}||_1 \\ & \qquad\qquad s.t. \quad \mathbf{H} \geq 0, \mathbf{a} \geq 0, \sum_{k=1}^{K} a_k = 1. \end{aligned} \tag{7.7}$$

where the first term is the KL divergence between $\mathbf{V}$ and $\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}$, and the second term is the L1-norm regularization term that causes the activity matrix to be sparse. λ represents the weight of the sparse constraint.

$\mathbf{W}_k^M$ is fixed, and $\mathbf{H}$ and $\mathbf{a}$ are estimated by minimizing (7.7). The updating rule is determined by adapting Jensen's inequality. The derivation of (7.8) and (7.9) is given in Appendix 8.

$$a_k \leftarrow \frac{a_k}{\sum_{d,l} (\mathbf{W}_k^M \mathbf{H})_{dl}} \sum_{d,l} \left(\frac{v_{dl} (\mathbf{W}_k^M \mathbf{H})_{dl}}{\sum_k a_k (\mathbf{W}_k^M \mathbf{H})_{dl}} \right) \tag{7.8}$$

$$\begin{aligned} \mathbf{H} \leftarrow \; & \mathbf{H}. * ((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^\mathsf{T} (\mathbf{V}./(\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}))) \\ & ./((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^\mathsf{T} \mathbf{1}^{D\times L} + \lambda \mathbf{1}^{J\times L}) \end{aligned} \tag{7.9}$$

where v_{dl} denotes the element of $\mathbf{V}$ and $.*$, $./$, and $\mathbf{1}$ denote element-wise multiplication, element-wise division, and an all-ones matrix, respectively. A speaker vector is normalized in each iteration so that the sum of the vector equals unity.

7.3 Experiments

7.3.1 Experimental Conditions

We used the ATR Japanese speech database set C [61], a corpus of the speech of 20 males and 20 females. In the case of inter-gender conversion, samples of 10 males or 10 females were used as training data and those from the other 10 males or 10 females were used as test data. In the case of cross-gender conversion, 10 males and 10 females were used as training data and the rest were used as test data. The sampling rate was 12 kHz. We compared our method with conventional one-to-one NMF-based VC and one-to-one GMM-based VC [19], which use parallel data from the source and the target speakers as training data. For each method, 50 parallel sentences spoken by each speaker were used for dictionary construction or training of the GMM. In the proposed method, two sentences uttered by the target speaker, which were not included in the test or training data, were used as adaptation data.

In the proposed and conventional NMF-based methods, the dimension number of the spectral features was 2,565. It consisted of a 513-dimensional STRAIGHT spectrum and its consecutive frames (the 2 frames coming before and the 2 frames coming after). These features are used instead of Δ features because we cannot use a negative value in a NMF-based method. The number of iterations for NMF and Multi-NMF was 300, and λ in (7.7) was set to 0.1. The activity matrix in the proposed and conventional NMF-based methods was initialized with positive random values. The speaker weight vector in the proposed method was initialized with the same positive value, $1/K$, where K is the number of speakers in the dictionary.

In the conventional GMM-based method, mel-cepstrum+Δ+$\Delta\Delta$ was used as a spectral feature. Its number of dimensions was 60. In order to reduce the number of parameters, diagonal covariance matrices were used in GMM [19] . In this paper, GV is not used. The number of Gaussian mixtures was set to 64, which was chosen experimentally.

In the conventional GMM and NMF-based methods, F0 information was converted using conventional linear regression based on the mean and standard deviation [19] as follows:

$$\hat{y}_t = \frac{\sigma^{(y)}}{\sigma^{(x)}}(x_t - \mu^{(x)}) + \mu^{(y)}, \tag{7.10}$$

where x_t and $\hat{y}_t$ denote the log-scaled F0 of the source speaker and the converted word at frame t, respectively. $\mu^{(x)}$ and $\sigma^{(x)}$ denote the mean and standard deviation of the log-scaled F0, as calculated from the source speaker's training data. $\mu^{(y)}$ and $\sigma^{(y)}$ are the mean and standard deviation of the target speaker data, respectively. In our proposed method, in order to focus on the spectra conversion, F0 information was converted using (7.10) and the mean and standard deviation of the log-scaled F0 were calculated form the parallel training data of the source and target speakers. We made no conversions to the aperiodic components.

In order to evaluate our proposed method, we conducted both objective and subjective evaluations. In section 7.3.3, 50 sentences that were not in the training data were used for the objective evaluation and the half of them are used in section 7.3.2. We used mel-cepstrum distortion (Mel-CD) [dB] [100] as a measure of the objective evaluations, defined as follows:

$$\mathit{Mel\text{-}CD} = (10/\ln 10)\sqrt{2\sum_{d=1}^{24}(mc_d^{conv} - mc_d^{tar})^2} \tag{7.11}$$

where mc_d^{conv} and mc_d^{tar} denote the d-th dimension of the converted and target mel-cepstral coefficients, respectively.

The subjective evaluation was based on "speech quality" and "similarity to the target speaker". For the subjective evaluation, 25 sentences were evaluated by 10 Japanese native speakers. For the evaluation of speech quality, we performed a mean opinion score (MOS) test [63]. The opinion score was calibrated by a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). For the similarity evaluation, an XAB test was carried out, in which each participant listened to the voice of the target speaker then to the voice converted by the two methods.

The participant was asked to judge which sample sounded most similar to the target speaker's voice.

7.3.2 Parameters

In this section, we evaluated the performance of our proposed method using different parameters. The default setting is explained in section 7.3.1.

First, we evaluated the performance of our proposed method using different numbers of bases in the multiple-dictionary. The results are shown in Table 7.1. M→M, F→F, M→F, and F→M denote male-to-male conversion, female-to-female conversion, male-to-female conversion, and female-to-male conversion, respectively. In "All bases", the default multiple-dictionary was used (its mean number of bases is 35,439). In "20,000 bases" and "10,000 bases", the indicated numbers of bases, which are randomly chosen from the default dictionary, were used. There is no significant difference between the number of bases in the dictionary and the performance; however we did confirm that there is some performance variation between speakers when we use a small dictionary. Therefore, we used all the bases in the following experiments.

Next, we evaluated the performance of our proposed method using different amounts of adaptation data. The results are shown in Table 7.2. There are no significant differences between the performance and the amount of adaptation data. Therefore, we assume the default setting (2 sentences) is good enough to estimate the target speaker weight vector.

Finally, we evaluated the performance of our proposed method using different numbers of speakers in the multiple-dictionary. In "10 speakers", all 10 speakers were stored in the multiple-dictionary. In "5 speakers", 5 speakers were randomly chosen from the full multiple-dictionary. Table 7.3 indicates that the reduced number of training speakers tends to degrade the conversion performance.

7.3.3 Results

Fig. 7.2 shows a mean vector calculated from the speaker weight vectors of a male, and Fig. 7.3 shows that of a female speaker. The error bar shows its variance. These figures show the robustness of the estimation of speaker weight vectors.

Table 7.1: Mel-CD [dB] using a different number of bases

	M→M	F→F	M→F	F→M
Source	4.79	4.79	5.42	5.42
All bases	4.42	4.38	4.66	4.63
20,000 bases	4.42	4.37	4.66	4.63
10,000 bases	4.43	4.38	4.65	4.63

Table 7.2: Mel-CD [dB] using the differing amounts of adaptation data

	M→M	F→F	M→F	F→M
Source	4.79	4.79	5.42	5.42
2 sentences	4.42	4.38	4.66	4.63
1 sentence	4.41	4.39	4.66	4.63

Table 7.3: Mel-CD [dB] using different number of speakers

	M→M	F→F	M→F	F→M
Source	4.79	4.79	5.42	5.42
10 speakers	4.42	4.38	4.66	4.63
5 speakers	4.45	4.48	4.70	4.60

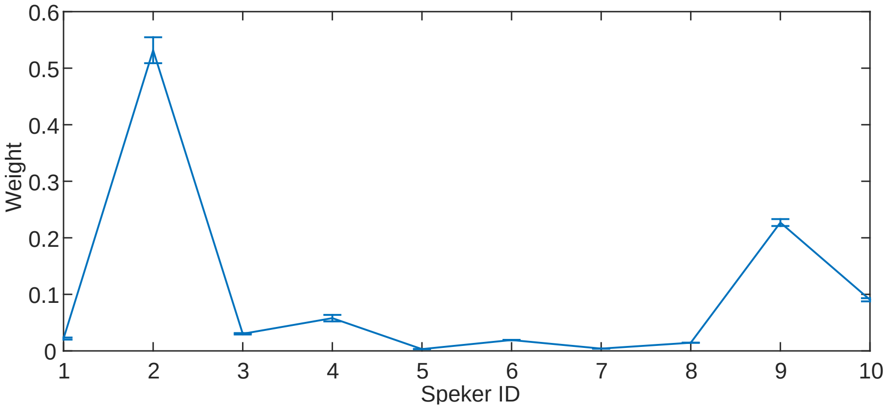


Figure 7.2: Mean and variance of the speaker weight vector (M105)

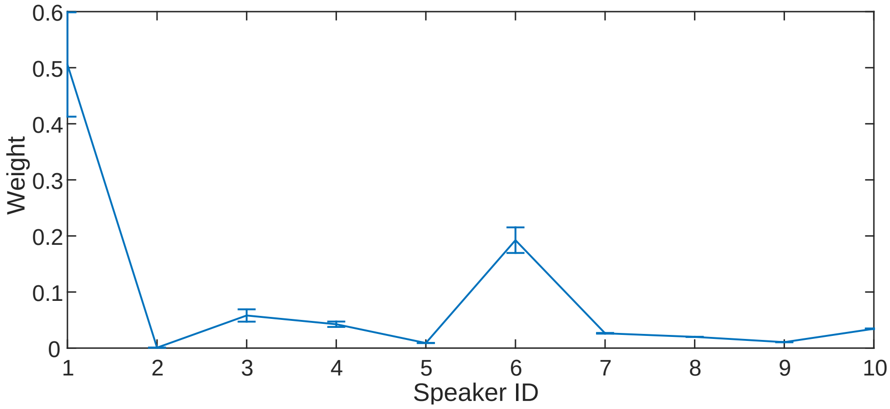


Figure 7.3: Mean and variance of the speaker weight vector (F105)

Tables 7.4 to 7.7 show the Mel-CD for male-to-male conversion, female-to-female conversion, male-to-female conversion, and female-to-male conversion, respectively. In these tables, a lower value indicates a better result. Source, Multi, NMF, and GMM denote Mel-CD between the target and the source speech, when converted by the proposed method, converted by one-to-one NMF, and by one-to-one GMM, respectively. Although the dictionaries in our proposed method included neither the source nor target speaker spectra, the difference in distortion between one-to-one VC methods and our proposed many-to-many VC method were small. For some pairs of speakers, the distortion of the proposed method was almost identical to that of the one-to-one conversions (for example, F5→F10 and F2→M2).

Table 7.4: Mel-CD [dB] of male-to-male conversion

	Source	Multi	NMF	GMM
M1→M6	4.76	4.16	4.06	3.93
M2→M7	5.29	4.92	4.71	4.74
M3→M8	4.68	4.47	4.15	4.23
M4→M9	4.59	4.18	3.92	3.92
M5→M10	4.29	4.02	3.69	3.62
Mean	4.72	4.35	4.11	4.09

Table 7.5: Mel-CD [dB] of female-to-female conversion

	Source	Multi	NMF	GMM
F1→F6	4.74	4.38	4.19	4.20
F2→F7	4.88	4.52	4.51	4.51
F3→F8	4.77	4.25	4.07	3.99
F4→F9	4.78	4.40	4.18	4.10
F5→F10	4.50	4.07	4.06	4.01
Mean	4.73	4.32	4.20	4.16

Table 7.6: Mel-CD [dB] of male-to-female conversion

	Source	Multi	NMF	GMM
M1→F1	5.46	4.59	4.32	4.59
M2→F2	5.05	4.59	4.32	4.37
M3→F3	5.22	4.44	4.24	4.27
M4→F4	5.89	4.95	4.83	4.73
M5→F5	5.05	4.39	4.04	4.06
Mean	5.34	4.57	4.35	4.41

Table 7.7: Mel-CD [dB] of female-to-male conversion

	Source	Multi	NMF	GMM
F1→M1	5.46	4.69	4.48	4.67
F2→M2	5.05	4.42	4.24	4.42
F3→M3	5.22	4.37	4.11	4.24
F4→M4	5.89	4.99	4.75	4.75
F5→M5	5.05	4.34	4.07	4.10
Mean	5.34	4.56	4.33	4.43

Fig. 7.4 shows the results of the MOS test on speech quality. The error bars show a 95% confidence score. Here, a higher value indicates a better result. M-to-M, F-to-F, M-to-F, and F-to-M denote male-to-male conversion, female-to-female conversion, male-to-female conversion, and female-to-male conversion, respectively. In inter-gender conversion, our proposed method achieved a significantly better score than both conventional one-to-one NMF and GMM-based VC, using a p-value test of 0.05. In cross-gender conversion, the difference between our proposed method and one-to-one NMF-based VC was non-significant. However, the proposed method achieved a significantly the better score than to one-to-one GMM-based VC.

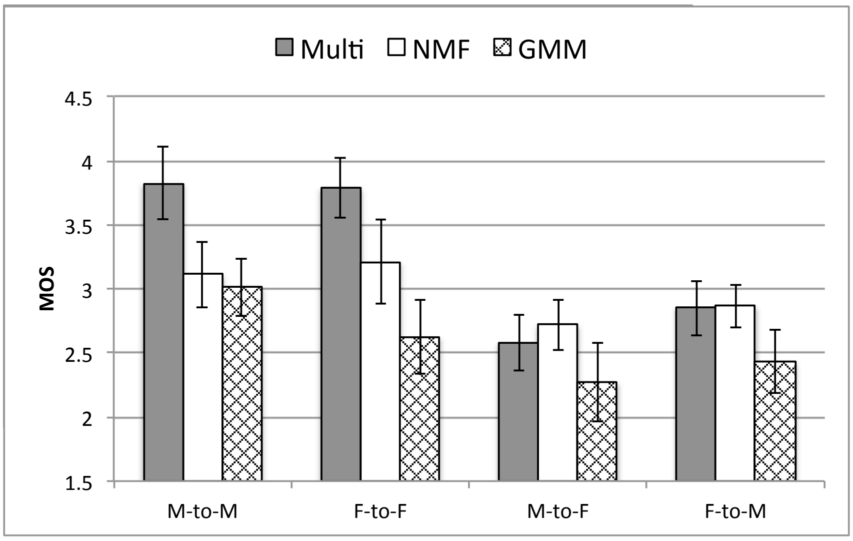


Figure 7.4: MOS of speech quality

Fig. 7.5 and Fig. 7.6 show the results of the XAB test. The error bars show a 95% confidence score. With this test, a higher value indicates a better result. Fig. 7.5 compares the results of the proposed method and one-to-one NMF-based VC. The difference in female-to-female conversion in Fig. 7.5 is not statistically significant because $p > 0.1$ in the p-value test. The difference in the other conversion in Fig. 7.5 is significant in $p < 0.05$. Fig. 7.6 compares the results of the proposed method and one-to-one GMM-based VC. The difference in Fig. 7.6 is not significant in $p > 0.1$. The score achieved by our proposed method was lower than that of the one-to-one NMF-based method except for female-to-female conversion. However, the difference between our proposed method and the one-to-one GMM-based method was not statistically significant. These speaker similarity tests show that our proposed many-to-many VC approach effectively converts the individuality of the source speaker's voice to the target speaker's voice.

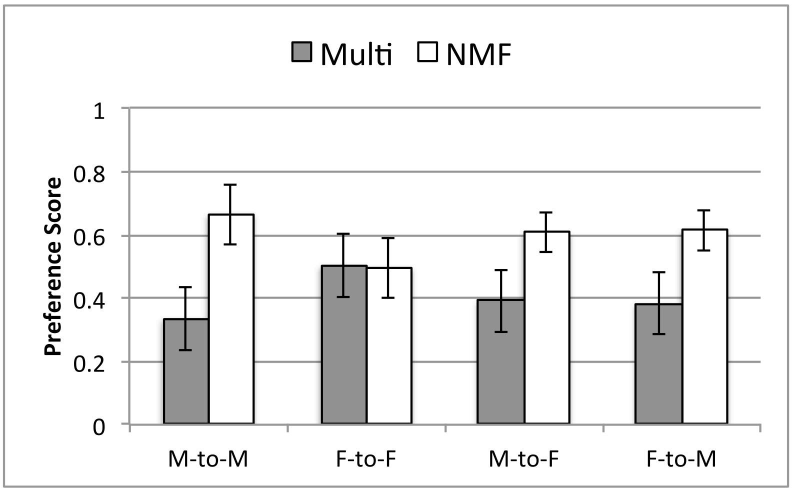


Figure 7.5: XAB test between proposed method and NMF

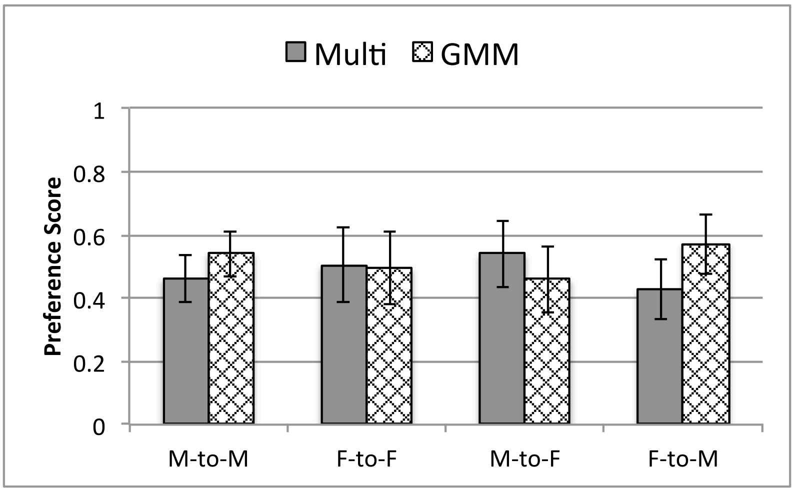


Figure 7.6: XAB test between proposed method and GMM

7.3.4 Discussion

In the objective evaluations, the difference in Mel-CD between our proposed method and the conventional one-to-one VC method was about 0.3 dB. We consider this difference quite small, considering that the Multi-NMF method stores neither the source nor target speakers' spectra in the dictionaries. We also assume that Mel-CD is superior to the GMM-based method because it uses mel-cepstrum as an input feature, whereas NMF-based methods use a STRAIGHT spectrum.

In the subjective evaluation of speech quality, the results were different for inter-gender conversion and cross-gender conversion. In inter-gender conversion, our Multi-NMF based method achieved a better score than conventional one-to-one VC methods, whereas in cross-gender conversion, the scores were almost the same. We assume that the performance difference across the two conditions reflects the difference between the dictionaries. In inter-gender conversion, the activity of the input spectra is estimated from a given dictionary, and the target spectra are constructed using the same dictionary and activity. In the case of cross-gender conversion, the activity of the input spectra is estimated from the source dictionary, while the target spectra are constructed from a different target dictionary. The same problem was also observed in one-to-one NMF-based VC. Our proposed method achieved a better score in inter-gender conversion because it avoids this problem by using Multi-NMF. In [101], we discussed this problem in one-to-one NMF-based VC and resolved it using an activity-mapping method.

In subjective evaluations of speaker similarity, our proposed method achieved almost the same score as the conventional GMM-based one-to-one VC method. The scores were slightly lower than those of the one-to-one NMF-based method, except for female-to-female conversion. We assume that this reflects the fact that the difference between Multi-NMF and NMF in female-to-female conversion is smaller than that for other conversions (Tables 7.4 to 7.7).

7.4 Chapter Conclusions

This study proposed a novel sparse representation of NMF-based framework, which allows exemplar-based VC for arbitrary speakers. Conventional one-to-one exemplar-based VC methods require source and target dictionaries, comprising parallel utterances by the source and target speakers. The input speaker's spectra are represented by linear combinations of spectra (bases) from a source dictionary and their weights. Converted spectra are constructed from the parallel bases from a target dictionary, and weights estimated by input spectra. Our proposed VC method does not require the utterances of the source and target speakers to be parallel. The proposed Multi-NMF estimates the source speaker weight vector and its activities from input spectra, using a dictionary of the spectra of many speakers. A target speaker weight vector is estimated from adaptation data, and the target speech is synthesized from the target speaker weight vector and the activities of the input speech. We assume that Multi-NMF makes it possible to decompose input speech into two parts: phonetic information, which is estimated as activities, and speaker information, which is estimated as the speaker weight vector. Moreover, our many-to-many VC method does not require the reference speaker's voice used in many-to-many EV-GMM-based VC. The experimental results demonstrate that the conversion quality of the proposed method is almost the same as that of conventional one-to-one VC, although our proposed method includes neither the source speakers' spectra, nor the target speakers' spectra.

Some problems remain. Our proposed method requires longer computation times than those of the GMM-based and one-to-one NMF-based methods. In NMF-based methods, the computation times reflect the size of the dictionary used. In [59], we proposed an NMF-based dictionary learning approach which reduces the computation time of NMF-based VC methods. In future research, we will therefore seek more compact representations of dictionaries for many-to-many VC. A more compact dictionary should also improve cross-gender conversion.

In future work, we will apply our method to noisy environments and to assistive technology for speakers with articulation disorders. We will also extend the comparisons between our method and other many-to-many VC methods. The proposed method should be easily applied to voice quality control, using regression of speaker weight vectors and voice expression words. We also plan to conduct research on speaker identification using the speaker weight vector.

Chapter 8

Conclusions

This thesis forces on VC methods based on NMF. We systematically propose methods for four practical VC tasks using NMF-based VC. We organize these novel and effective proposed methods in the thesis and cover major research themes this field of research.

In Chapter 4, we propose a framework to train the basis matrices of source and target exemplars to give then a common weight matrix. We discuss a noise-robust VC technique based on sparse representation. We propose a framework to train the basis matrices of source and target exemplars to give them a common activity matrix. The basis matrix of the source exemplars was trained using NMF. Then, the basis matrix of the target exemplars is trained using NMF, where the weight matrix was fixed to that obtained from the source exemplars. By using the basis matrices instead of the exemplars, the VC performed with less computation times than the exemplar-based method. In comparison experiments, exemplar-based and proposed methods showed better performance than the GMM-based method when evaluating noisy speech.

In Chapter 5, we present a VC method for a person with an articulation disorder resulting from athetoid cerebral palsy. We propose a spectral conversion method based on NMF for a voice with this articulation disorder. Our proposed method introduces a dictionary selection method to conventional NMF-based VC. The results demonstrated that our VC method improves the listening intelligibility of words uttered by a person with an articulation disorder. Moreover, compared to conventional GMM-based and conventional NMF-based VCs, our proposed VC preserves the individuality and the naturalness of the source speaker's voice. The idea of dictionary selection for articulation disorders are also adopted for TTS [102].

In Chapter 6, an adaptation matrix in an NMF framework is introduced to adapt the source dictionary to the target dictionary. This method requires only a small amount of parallel data, A linear regression transformation matrix is used to adapt a source dictionary to a target dictionary and is estimated in an NMF framework. In comparison experiments between GMM-based VC, NMF without speaker adaptation, and NMF with speaker adaptation, the NMF-based VC with speaker adaptation gives the best performance.

In Chapter 7, we propose Multi-NMF to allow the implementation of many-to-many, exemplar-based VC. Our proposed VC method does not require the utterances of the source and target speakers to be parallel. The proposed Multi-NMF estimates the source speaker weight vector and its activity from the input spectra, using a dictionary of the spectra of many speakers. A target speaker weight vector is estimated from adaptation data, and the target speech is synthesized from the target speaker's weight vector and activity of the input speech. We assume that Multi-NMF makes it possible to decompose input speech into two parts: phonetic information, estimated as activity, and speaker information, estimated as the speaker's weight vector.

As explained above, this paper proposes new algorithms for four important VC tasks. and expands the use of VC systems. These tasks are not only related to VC but also to speech recognition and other problems in the field of signal processing.

Some problems remain. Our proposed method requires longer computation times than those of the GMM-based and one-to-one NMF-based methods. In NMF-based methods, the computation times reflect the size of the dictionary used. In [103], we proposed NMF-based dictionary learning based on an Alternating Direction Method of Multipliers (ADMM) [50]. By combining with ADMM-based dictionary learning, our proposed algorithm can be adopted to real time VC. NMF-based dictionary learning using graph-embedded have also been proposed in [96].

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara. Esophageal speech enhancement based on statistical voice conversion with gaussian mixture models. *in Proc. ICASSP*, pages 655–658, 1988. 1

[2] H. Valbret, E. Moulines, and J. P. Tubach. Voice transformation using PSOLA technique. *Speech Communication, vol. 11, no. 2-3, pp. 175-187*, 1992. 1

[3] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, 6(2):131–142, 1998. 1, 4, 14, 34, 45, 61

[4] C. Veaux and X. Robet. Intonation conversion from neutral to expressive speech. *in Proc. Interspeech*, pages 2765–2768, 2011. 1, 67

[5] H. Kawanami, Y. Iwami, T. Toda, and K. Shikano. GMM-based voice conversion applied to emotional speech synthesis. *in Proc. EUROSPEECH*, 2003. 1

[6] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki. GMM-based emotional voice conversion using spectrum and prosody features. *American Journal of Signal Processing*, 2(5), 2012. 1

[7] R. Aihara, R. Ueda, T. Takiguchi, and Y. Ariki. Exemplar-based emotional voice conversion using non-negative matrix factorization. *in Proc. APSIPA*, 2014. 1

[8] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. *Speech Communication*, 54(1):134–146, 2012. 1, 41

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

[10] 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*, pages 2494–2498, 2014. 1

[11] Y. Ohtani, M. Tamura, M. Morita, and M. Akamine. GMM-based bandwidth extension using sub-band basis spectrum model. *in Proc. Interspeech*, pages 2489–2493, 2014. 1

[12] S. Möller. *Assessment and prediction of speech quality in telecommunications*. Springer, 2000. 2

[13] Y. Lavner, J. Rosenhouse, and I. Gath. The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, 4(1):63–74, 2001. 2

[14] E. Helander and J. Nurminen. On the importance of pure prosody in the perception of speaker identity. *in Proc. Interspeech*, pages 2665–2668, 2007. 2

[15] D. T. Chappell and J. Hansen. Speaker-specific pitch contour modeling and modification. *in Proc. ICASSP*, 2:885–888, 1998. 2

[16] B. Gillett and S. King. Transforming F0 contours. *in Proc. EUROSPEECH*, 2003. 2

[17] E. Helander and J. Nurminen. A novel method for prosody prediction in voice conversion. *in Proc. ICASSP*, 4:509–512, 2007. 2

[18] M. Müller. *Information retrieval for music and motion*, volume 6. Springer, 2007. 4, 14

[19] T. Toda, A. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(8):2222–2235, 2007. 4, 18, 19, 23, 47, 62, 75, 76

[20] 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*, pages 3052—3056, 2013. 4

[21] T. Nakashika. Voice conversion based on deep learning. *Doctral Thesis*, 2014. 4, 9

[22] 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*, 23(3):580–587, 2015. 4

[23] R. Takashima, T. Takiguchi, and Y. Ariki. Exemplar-based voice conversion in noisy environment. *in Proc. SLT*, pages 313–317, 2012. 4, 28, 34, 61

[24] 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.*, 22(10):1506–1521, 2014. 4

[25] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Neural Information Processing System*, pages 556–562, 2001. 4, 21, 22, 23, 27

[26] R. Takashima, T. Takiguchi, and Y. Ariki. Exemplar-based voice conversion using sparse representation in noisy environments. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E96-A(10):1946–1953, 2013. 4, 55

[27] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki. Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary. *in Proc. ICASSP*, pages 7944–7948, 2014. 5, 39, 62, 69

[28] 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. 5, 19

[29] B. P. Bogert, M. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *in Proc. the symposium on time series analysis*, 15:209–243, 1963. 10

[30] P. Mermelstein. Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388, 1976. 11

[31] B. Milner and X. Shao. Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model. *in Proc. Interspeech*, pages 2421–2424, 2002. 12

[32] Z. Tychtl and J. Psutka. Speech production based on the mel-frequency cepstral coefficients. *in Proc. EUROSPEECH*, 99:2335–2338, 1999. 12

[33] T. Ramabadran, J. Meunier, M. Jasiuk, and B. Kushner. Enhancing distributed speech recognition with back-end speech reconstruction. *in Proc. Interspeech*, pages 1859–1862, 2001. 12

[34] S. Imai. Cepstral analysis synthesis on the mel frequency scale. *in Proc. ICASSP*, 8:93–96, 1983. 12

[35] H. Kawahara, I. Masuda-Katsuse, and A.de 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*, 27(3-4):187–207, 1999. 12, 47

[36] H. Kawahara. STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, 27(6):349–353, 2006. 12, 71

[37] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. *in Proc. ICASSP*, pages 3933–3936, 2008. 13

[38] M. Morise, F. Yokomori, and K. Ozawa. World: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE transactions on information and systems*, E99-D(7):1877–1884, 2016. 13

[39] D. Erro, I. Sainz, E. Navas, and I. Hernaez. Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE J. Sel. Topics in Signal Process.*, 8(2):184–194, 2014. 13

[40] A. R. Toth and A. W. Black. Using articulatory position data in voice transformation. *in Proc. ISCA SSW6*, pages 182–187, 2007. 14

[41] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977. 15

[42] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. *in Proc. ICASSP*, pages 1315–1318, 2000. 18

[43] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modulation spectrum-constrained trajectory training for gmm-based voice conversion. *in Proc. ICASSP*, pages 4859–4863, 2015. 19

[44] S. de Jong. SIMPLS: An alternative approach to partial least squares regression. *Chemometrics Intell. Lab. Syst*, 18(3):251–263, 1993. 19

[45] 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*, 20(3):806–817, 2012. 19

[46] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, 2011. 22

[47] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(7):2067–2080, 2011. 22, 25, 28, 30

[48] P. O. Hoyer. Non-negative matrix factorization with sparseness constraint. *Journal of Machine Learning Research*, (5):1457–1469, 2004. 22

[49] J. Kim, R. D. C. Monteiro, and H. Park. Group sparsity in nonnegative matrix factorization. *in Proc. the SIAM International Conference on Data Mining*, pages 851–862, 2012. 22

[50] D. L. Sun and C. Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. *in Proc. ICASSP*, pages 6242–6246, 2014. 22, 86

[51] T. Virtanen, B. Raj, J. F. Gemmeke, and H. Van hamme. Active-set newton algorithm for non-negative sparse coding of audio. *in Proc. ICASSP*, (3116–3120), 2014. 22

[52] J. F. Gemmeke and T. Virtanen. Noise robust exemplar-based connected digit recognition. *in Proc. ICASSP*, pages 4546–4549, 2010. 22

[53] M. N. Schmidt and R. K. Olsson. Single-channel speech separation using sparse non-negative matrix factorization. *in Proc. Interspeech*, pages 2614–2617, 2006. 22, 27

[54] T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(3):1066–1074, 2007. 22, 27

[55] H. Sawada, H. Kameoka, S. Araki, and N. Ueda. Efficient algorithms for multichannel extensions of itakura-saito nonnegative matrix factorization. *in Proc. ICASSP*, pages 261–264, 2012. 22

[56] C. Févotte, N. Bertin, and J. L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence with application to music analysis. *Neural Computation*, 21(3):793–830, 2009. 22

[57] A. Cichocki, R. Zdnek, A. H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. WILKEY, 2009. 22

[58] T. Barker and T. Virtanen. Non-negative tensor factorisation of modulation spectrograms for monaural sound source separation. *in Proc. Interspeech*, pages 827–831, 2013. 22

[59] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki. Noise-robust voice conversion based on sparse spectral mapping using non-negative matrix factorization. *IEICE Transactions on Information and Systems*, E97-D(6):1411–1418, 2014. 27, 67, 84

[60] Bjorn Schuller, Felix Weninger, Martin Wollmer, Yang Sun, and Gerhard Rigoll. Non-negative matrix factorization as noise-robust feature extractor for speech recognition. *in Proc. ICASSP*, 2010. 28

[61] 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*, 9:357–363, 1990. 34, 45, 61, 75

[62] N. Kitaoka, T. Yamada, S. Tsuge, C. Miyajima, K. Yamamoto, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Matsuda,

T. Ogawa, S. Kuroiwa, K. Takeda, and S. Nakamura. CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments. *Acoustical Science and Technology*, 30 (2009)(5):363–371, 2009. 34, 61

[63] INTERNATIONAL TELECOMMUNICATION UNION. Methods for objective and subjective assessment of quality. *ITU-T Recommendation P.800*, 2003. 35, 50, 64, 76

[64] K. Masaka, R. Aihara, T. Takiguchi, and Y. Ariki. Multimodal exemplar-based voice conversion using lip features in noisy environments. *in Proc. Interspeech*, 1159-1163, 2014. 38

[65] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li. Exemplar-based voice conversion using non-negative spectrogram deconvolution. *in Proc. SSW8*, 2013. 38

[66] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki. Individuality-preserving voice conversion for articulation disorders based on Non-negative Matrix Factorization. *in Proc. ICASSP*, pages 8037–8040, 2013. 39, 40, 45

[67] 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*, 2014:5, doi:10.1186/1687-4722-2014-5, 2014. 39, 67

[68] R. Aihara, T. Takiguchi, and Y. Ariki. Individuality-preserving voice conversion for articulation disorders using phoneme-categorized exemplars. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):13:1–13:17, 2015. 39

[69] M. V. Hollegaard, K. Skogstrand, P. Thorsen, B. Norgaard-Pedersen, D. M. Hougaard, and J. Grove. Joint analysis of SNPs and proteins identifies regulatory IL18 gene variations decreasing the chance of spastic cerebral palsy. *Human Mutation, Vol. 34*, pages 143–148, 2013. 39, 40

[70] S. T. Canale and W. C. Campbell. Campbell's operative orthopaedics. Technical report, Mosby-Year Book, 2002. 40

[71] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayachi. Integration of metamodel and acoustic model for dysarthric speech recognition. *Journal of Multimedia, Volume 4, Issue 4, pp. 254-261*, 2009. 40

[72] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li. Multimodal speech recognition of a person with articulation disorders using AAM and MAF. *in Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP'10)*, pages 517–520, 2010. 40

[73] J. Lin, W. Ying, and T. S. Huang. Capturing human hand motion in image sequences. *in Proc. IEEE Motion and Video Computing Workshop*, pages 99–104, 2002. 41

[74] T. Starner, J. Weaver, and A. Pentland. Real-time American sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12), pp. 1371-1375*, 1998. 41

[75] G. Fang, W. Gao, and D. Zhao. Large vocabulary sign language recognition based on hierarchical decision trees. *in Proc. 5th International Conference on Multimodal Interfaces*, pages 125–131, 2003. 41

[76] N. Ezaki, M. Bulacu, and L. Schomaker. Text detection from natural scene images: Towards a system for visually impaired persons. *in Proc. ICPR*, pages 683–686, 2004. 41

[77] M. K. Bashar, T. Matsumoto, Y. Takeuchi, H. Kudo, and N. Ohnishi. Unsupervised texture segmentation via wavelet-based locally orderless images (wlois) and SOM. *in Proc. 6th IASTED International Conference COMPUTER GRAPHICS AND IMAGING*, 2003. 41

[78] V. Wu, R. Manmatha, and E. M. Riseman. Textfinder: an automatic system to detect and recognize text in images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(11), pp. 1224-1229*, 1999. 41

[79] K. Yabu, T. Ifukube, and S. Aomura. A basic design of wearable speech synthesizer for voice disorders [japanese]. *EIC Technical Report (Institute of Electronics, Information and Communication Engineers)*, 105(686):59–64, 2006. 41

[80] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. *in Proc. Interspeech*, pages 148–151, 2006. 41

[81] C. Veaux, J. Yamagishi, and S. King. Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders. *in Proc. Interspeech*, 2012. 42

[82] J. Yamagishi, Christophe Veaux, Simon King, and Steve Renals. Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology, Vol. 33 (2012) No. 1*, pages 1–5, 2013. 42

[83] A. Maier, T. Haderlein, F. Stelzle, E. Noth, E. Nkenke, F. Rosanowski, A. Schutzenberger, and M. Schuster. Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010. 42

[84] R. Aihara, Y. Takashima, T. Takiguchi, and Y. Ariki. Home appliance control using speech recognition for a person with an articulation disorder. *in Proc. The 17th International Symposium on Applied Electromagnetics and Mechanics (ISEM2015)*, 2015. 42

[85] R. Aihara, T. Fujii, T. Nakashika, T. Takiguchi, and Y. Ariki. Noise-robust voice conversion using a small parallel data based on non-negative matrix factorization. *in Proc. The 23rd European Signal Processing Conference (EUSIPCO)*, pages 315–319, 2015. 57

[86] R. Aihara, T. Fujii, T. Nakashika, T. Takiguchi, and Y. Ariki. Small-parallel exemplar-based voice conversion in noisy environments using affine non-negative matrix factorization. *EURASIP Journal on Audio, Speech, and Music Processing, doi:10.1186/s13636-015-0075-4*, 2015. 57

[87] C. H. Lee and C. H. Wu. MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training. *in Proc. Interspeech*, pages 2254–2257, 2006. 57, 58

[88] A. Mouchtaris, J. Van der Spiegel, and P. Mueller. Nonparallel training for voice conversion based on a parameter adaptation approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3):952–963, 2006. 57, 58

[89] T. Toda, Y. Ohtani, and K. Shikano. Eigenvoice conversion based on Gaussian mixture model. *in Proc. Interspeech*, pages 2446–2449, 2006. 57, 69, 70

[90] D. Saito, K. Yamamoto, N. Minematsu, and K. Hirose. One-to-many voice conversion based on tensor representation of speaker space. *in Proc. Interspeech*, pages 653–656, 2011. 57, 70

[91] E. M. Grais and H. Erdogan. Adaptation of speaker-specic bases in non-negative matrix factorization for single channel speech-music separation. *in Proc. Interspeech*, pages 569–572, 2011. 57

[92] H. Kawahara and H. Matsui. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. *in Proc. ICASSP*, I:256–259, 2003. 62

[93] T. En-Najjary, O. Roec, and T. Chonavel. A voice conversion method based on joint pitch and spectral envelope transformation. *in Proc. ICSLP*, pages 199–203, 2004. 62

[94] R. Aihara, T. Takiguchi, and Y. Ariki. Many-to-many voice conversion based on multiple non-negative matrix factorization. *in Proc. Interspeech*, pages 2749–2753, 2015. 69

[95] R. Aihara, T. Takiguchi, and Y. Ariki. Many-to-one voice conversion using exemplar-based sparse representation. *in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2015. 69

[96] R. Aihara, T. Takiguchi, and Y. Ariki. Multiple non-negative matrix factorization for many-to-many voice conversion. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 24(7):1175–1184, 2016. 69, 86

[97] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Many-to-many eigenvoice conversion with reference voice. *in Proc. Interspeech*, pages 1623–1626, 2009. 69, 70

[98] T. Masuda and M. Shozakai. Cost reduction of training mapping function based on multistep voice conversion. *in Proc. ICASSP*, 4:693–696, 2007. 70

[99] K. Ohta, T. Toda, Y. Ohtani, H. Saruwatari, and K. Shikano. Adaptive voice-quality control based on one-to-many eigenvoice conversion. *in Proc. Interspeech*, pages 2158–2161, 2010. 70

[100] J. Kominek, T. Schultz, and A. W Black. Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion. *in Proc. the International Workshop on Spoken Language Technology for Under-Resourced Languages (SLTU)*, 2008. 76

[101] R. Aihara, T. Takiguchi, and Y. Ariki. Activity-mapping non-negative matrix factorization for exemplar-based voice conversion. *in Proc. ICASSP*, pages 4899–4903, 2015. 83

[102] R. Ueda, R. Aihara, T. Takiguchi, and Y. Ariki. Individuality-preserving spectrum modification for articulation disorders using phone selective synthesis. *in Proc. SLPAT*, 2015. 86

[103] R. Aihara, T. Takiguchi, and Y. Ariki. Semi-non-negative matrix factorization using alternating direction method of multipliers for voice conversion. *in Proc. ICASSP*, pages 5170–5174, 2016. 86

Appendix

Update rule of NMF

The cost function of NMF (3.4) can be rewritten as

$$\begin{aligned} & d(\mathbf{V}, \mathbf{WH}) + \lambda||\mathbf{H}||_1 && (8.1) \\ =& \sum_{d,l} \left\{ v_{dl} \log v_{dl} - v_{dl} \log(\sum_j w_{dj} h_{jl}) - v_{dl} + \sum_j w_{dj} h_{jl} \right\} + \lambda \sum_{j,l} h_{jl} \\ \leq& \sum_{d,l} \left\{ v_{dl} \log v_{dl} - v_{dl} \sum_j \psi_j \log(\frac{w_{dj} h_{jl}}{\psi_j}) - v_{dl} + \sum_j w_{dj} h_{jl} \right\} + \lambda \sum_{j,l} h_{jl} \\ =& P(\mathbf{V}, \mathbf{W}, \mathbf{H}, \psi_j) && (8.2) \end{aligned}$$

where v_{dl}, w_{dl}, and h_{jl} represent the element of $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{H}$, respectively, using Jensen's inequality, ψ_j defined as follows:

$$\psi_j = \frac{w_{dj} h_{jl}}{\sum_n w_{dn} h_{nl}} \qquad (8.3)$$

The partial derivative of P in h_{jl} is written as

$$\frac{\partial P}{\partial h_{jl}} = \sum_d \left\{ -v_{dl} \psi_j \frac{1}{h_{jl}} + w_{dj} \right\} + \lambda \qquad (8.4)$$

Setting $\frac{\partial P}{\partial h_{jl}} = 0$, the updating rule is derived as

$$h_{jl} \leftarrow \sum_d v_{dl} \frac{w_{dj} h_{jl} / \sum_n w_{dn} h_{nl}}{\sum_f w_{fj} + \lambda} \qquad (8.5)$$

(8.5) can be rewritten to matrix form as

$$\mathbf{H} \leftarrow \mathbf{H}. * (\mathbf{W}^\mathsf{T}(\mathbf{V}./(\mathbf{WH}))) ./ (\mathbf{W}^\mathsf{T} \mathbf{1}^{D \times L} + \lambda \mathbf{1}^{J \times L}) \qquad (8.6)$$

Update rule of Multi-NMF

Activity matrix

The update rule of $\mathbf{H}$ in the proposed Multi-NMF is derived by replacing $\mathbf{W}$ in (8.6) with $\sum_{k=1}^{K} a_k \mathbf{W}_k^M$, as follows:

$$\begin{aligned}\mathbf{H} \;\leftarrow\; & \mathbf{H}.*((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^\mathsf{T}(\mathbf{V}./(\sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}))) \\ & ./((\sum_{k=1}^{K} a_k \mathbf{W}_k^M)^\mathsf{T} \mathbf{1}^{D\times L} + \lambda \mathbf{1}^{J\times L}) \qquad (8.7)\end{aligned}$$

Speaker weight vector

Here, the sparse function can be omitted. By using Jensen's inequality, the upper bound of the cost function (7.7) is derived as follows:

$$\begin{aligned}& d(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}) \qquad (8.8)\\ =& \sum_{d,l} \left\{ v_{dl} \log v_{dl} - v_{dl} \log(\sum_{k}^{K} a_k \mathbf{W}_k^M \mathbf{H})_{dl} - v_{dl} + \sum_{k}^{K} (a_k \mathbf{W}_k^M \mathbf{H})_{dl} \right\} \\ \leq& \sum_{d,l} \left\{ v_{dl} \log v_{dl} - v_{dl} \sum_{k}^{K} \alpha_k \log(\frac{a_k \mathbf{W}_k^M \mathbf{H}}{\alpha_k})_{dl} - v_{dl} + \sum_{k}^{K} (a_k \mathbf{W}_k^M \mathbf{H})_{dl} \right\} \\ =& Q(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}, \alpha_k) \qquad (8.9)\\ \leq& \sum_{d,l} \left\{ \log v_{dl} - v_{dl} \sum_{k} \alpha_k \sum_{j} \beta_j \log(\frac{a_k w_{dj}^k h_{jl}}{\beta_j}) + v_{dl} \alpha_k \log \alpha_k - v_{dl} \right. \\ & \left. + \sum_{k} (a_k \sum_{j} w_{dj}^k h_{jl}) \right\} \\ =& R(\mathbf{V}, \sum_{k=1}^{K} a_k \mathbf{W}_k^M \mathbf{H}, \alpha_k, \beta_j) \qquad (8.10)\end{aligned}$$

where v_{dl}, w^k_{dl}, h_{jl}, and a_k represent the element of $\mathbf{V}$, $\mathbf{W}^M_k$, $\mathbf{H}$, and $\mathbf{a}$, respectively. α_k and β_j are defined as follows:

$$\begin{aligned} \alpha_k &= \frac{a_k \sum_j (w^k_{dj} h_{jl})}{\sum_m a_m \sum_j (w^k_{dj} h_{jl})} && (8.11) \\ \beta_j &= \frac{w^k_{dj} h_{jl}}{\sum_n w^k_{dn} h_{nl}} && (8.12) \end{aligned}$$

The partial derivative of R in a_k is written as

$$\frac{\partial R}{\partial a_k} = \sum_{d,l} \left\{ -v_{dl} \sum_k \alpha_k \sum_j \beta_j \frac{1}{a_k} + \sum_k \sum_j w^k_{dj} h_{jl} \right\} \qquad (8.13)$$

Setting $\frac{\partial R}{\partial a_k} = 0$, the updating rule is derived as follows:

$$a_k \leftarrow \frac{a_k}{\sum_{d,l} (\mathbf{W}^M_k \mathbf{H})_{dl}} \sum_{d,l} \left(\frac{v_{dl} (\mathbf{W}^M_k \mathbf{H})_{dl}}{\sum_k a_k (\mathbf{W}^M_k \mathbf{H})_{dl}} \right) \qquad (8.14)$$

Acknowledgements

First, I would like to thank my supervisors, Emeritus Professor Yasuo Ariki and Associate Professor Tetsuya Takiguchi at Kobe University, who have given me helpful advice and continued support during my research and writing up. Their broad knowledge in the field and his down-to-earth attitude has been of great help to my study. I also thank Professor Ohkawa, Professor Tamaki, and Professor Matoba for their constructive comments and valuable suggestions, which helped to improve this thesis.

Meanwhile, I would like to thank the past and the present members in CS 17 Media Lab., where we have done efforts together and shared joys and sorrows of research life.

Finally, to my family and my friends, it cannot be described in words, yet I would like to show my full gratitude for their love, encouragements, and unconditional support throughout my study.

BibTeX Citation for This Thesis

```
@article  {R. AiharaKBUPhDThesis2017,
            title={{Voice Conversion Based on Non-negative Matrix Factorization
                    and Its Application to Practical Tasks}},
            journal={Doctoral Thesis},
            author={Ryo Aihara},
            institution={Kobe University},
            month={Mar.},
            year={2017}
          }
```

Doctor Thesis, Kobe University
“Voice Conversion Based on Non-negative Matrix Factorization and Its Application to Practical Tasks”, 130 pages
Submitted on January, 19, 2017.
The date of publication is printed in the cover of repository version published in Kobe University Repository Kernel.