Doctoral Thesis

Zero-Shot Recognition of Generic Objects

Tristan HASCOET

Graduate School of System Informatics

Kobe University

A thesis submitted for PhD. degree

June 2019

Doctoral Thesis

Zero-Shot Recognition of Generic Objects

Tristan HASCOET

In its essence, machine perception aims to extract interpretable structured infortmation from unstructured signals. Object recognition is a foundational task for computer vision and machine perception more broadly. Since the remarkable success of AlexNet in the ILSVRC2012 competition, Convolutional Neural Networks (CNN) have allowed for unprecedent progress in object recognition, which has opened the door for new applications that were previously though impossible. CNN-based classifiers have become the backbone of modern computer vision. Complex vision systems from object detection and image segmentation systems to higher level models such as image captioning and Visual Question Answering systems, have all been built on top of the backbone architecture of CNN classifiers.

Given this success and the central place of CNN-based object recognition components in vision systems, it is important to think about their limitations. On the conceptual side, object recognition is currently framed as a supervised classification problem. This classification setting induces a closed world assumption: The set of object categories a model can recognize is finite and fixed both by the architecture and the available training data. Outside Data annotation problem. Closed world

In comparison, humans flexibility. Open world. Because we combine perceptual abilities with higher abstraction formalisms and reasoning. For instance, children can recognize zebra.

The analogy to the human ability to recognize unknown obects has motivated ZSL. ZSL do XXX From a practical perspective, XXX. From a research point of view, XXX.

Despite its great potential impact and after a decade of active research, XXX. In this paper, we XXX

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.* ix-xv.

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”, *INTER-SPEECH 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, SP2014-53pp. 39-442014. (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

ZSL	Zero-Shot Learning
CNN	Convolutional Neural Network
NLP	Natural Language Processing
GCN	Graph Convolution Network

0. GLOSSARY

Chapter 1

Introduction

1.1 Background

1.2 Approaches

1.3 Purpose of This Thesis

1.3.1 Four Practical VC Tasks

1.3.1.1 Noise-robust VC

1.3.1.2 Assistive Technology for Articulation Disorders

1.3.1.3 VC Using Small-parallel Training Data

1.3.1.4 Many-to-many VC

1.3.2 Novelties of This Thesis

1.4 Outline

1. INTRODUCTION

Chapter 2

Visual Feature Extraction

The related publications for this chapter are [].

2.1 The Motivation and Related Work

2.1.1 Motivation

Chapter 3

Visual Feature Extraction

The related publications for this chapter are [].

3.1 The Motivation and Related Work

3.1.1 Motivation

Chapter 4

Visual Feature Extraction

The related publications for this chapter are [].

4.1 The Motivation and Related Work

4.1.1 Motivation

Chapter 5

Conclusions

In []

5. CONCLUSIONS

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.

[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.

[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.

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

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

[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.

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

[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.

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

[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.

[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.

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

[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.

[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.

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

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

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

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

[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.

[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.

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

[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.

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

[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.

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

[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.

[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.

[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.

[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.

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

[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.

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

[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.

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

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

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

[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.

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

[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.

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

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

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

[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.

[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.

[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.

[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.

[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.

Appendix

some section

5. APPENDIX

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.

5. ACKNOWLEDGEMENTS

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.