Kobe University Repository : Kernel

PDF issue: 2026-11-24

Emotional Voice Conversion using Deep Learning Models

Luo, Zhaojie

(Degree)
博士（工学）

(Date of Degree)
2020-03-25

(Date of Publication)
2021-03-01

(Resource Type)
doctoral thesis

(Report Number)
甲第7783号

(URL)
https://hdl.handle.net/20.500.14094/D1007783

※ 当コンテンツは神戸大学の学術成果です。無断複製・不正使用等を禁じます。著作権法で認められている範囲内で、適切にご利用ください。

Doctoral Thesis

Emotional Voice Conversion using Deep Learning models

ディープラーニングモデルを用いた感情音声変換

Zhaojie Luo

Graduate School of System Informatics

Kobe University

A thesis submitted for Ph.D. degree

January 2020



Doctoral Thesis

Emotional Voice Conversion using Deep Learning Models

Zhaojie Luo

An artificial neural network is an important model for training features of voice conversion (VC) tasks. Typically, neural networks (NNs) are very effective in processing nonlinear features, such as Mel Cepstral Coefficients (MCC), which represent the spectrum features. However, a simple representation of fundamental frequency (F0) is not enough for NNs to deal with emotional voice VC, which is the most important feature in emotional voice representation, another problem is that there are insufficient emotional data for training. In this paper, I propose several emotional voice conversion (VC) models and the emotional voice synthesis model that can effectively learn emotional conversion functions or emotional synthesis functions with limited emotional voice data.

Chapter 1 introduces the study, novelty, and outline of this thesis to position our contributions. Chapter 2, briefly introduces the acoustic features and basics models that are frequently adopted in this study. Chapter 3 shows the conventional emotional VC methods such as using the Gaussian Mixture Model (GMM) to convert the prosody features for emotional VC, or applying the Non-Negative Matrix Factorization (NMF) in the emotional VC, or using the Deep Neural Networks (DNNs) method for emotional VC.

Chapter 4 proposes a novel and general emotional VC framework using neural networks based on the Continuous Wavelet Transform (CWT) method. In this study, I propose the arbitrary scales CWT (AS-CWT) method to systematically capture F0 features of different temporal scales, which can represent different prosodic levels ranging from micro-prosody to sentence levels. By utilizing these approaches, the proposed method can change the spectrum and the F0 for an emo-

tional voice simultaneously as well as outperform other state-of-the-art methods in terms of emotional VC.

Chapter 5 proposes a novel neutral-to-emotional voice synthesis model that can effectively learn a mapping from neutral to emotional speech with limited emotional voice data. The idea is 1) to decompose F0 into different temporal level representations using Continuous Wavelet Transform (CWT); 2) and use Cross Wavelet Transform (XWT) to combine different CWT-F0 features to synthesize interaction XWT-F0 features; 3) then, use both the CWT-F0 and corresponding XWT-F0 features for training emotional synthesis model.

Chapter 6 proposes the emotional VC model that using Dual Supervised Generative Adversarial Networks (Dual-SANs) With CWT-F0 Features. In an emotional VC task, each dataset is paired with the labeled emotional voice and neutral voice, which can be regarded as a dual task. Owing to 1) dual supervised learning's ability to improve the training performances by using the leveraging probabilistic connection between the dual tasks to enhance the learning from labeled data and 2) Generative Adversarial Networks' (GANs') ability to mitigate the over-smoothing problem caused in the low-level data space when converting the acoustic features, we further present a novel training framework for emotional VC using GANs combined with dual supervised learning, named as dual supervised adversarial networks (dual-SANs).

Finally, Chapter 7 draws the conclusions for this thesis.

Contents

List of Figures ix

List of Tables xiii

Publication List xvii

Glossary xxi

1 Introduction 1

1.1 Background 1

1.2 Approaches 2

1.3 Purpose of This Thesis 4

1.3.1 Adaptive Emotional Features Processing for Emotional VC 4

1.3.2 Increasing Emotional Features for Emotional VC 5

1.3.3 Robust Training Model for Emotional VC 6

1.4 Thesis Outline 6

2 Acoustic Features and Basics 9

2.1 Cepstrum 9

2.2 MFCC 11

2.3 F0 13

2.4 Continuous Wavelet Transform 13

2.5 STRAIGHT 14

3 Conventional Emotional VC methods 17

3.1 Overview 17

3.2 DBNs-based Emotional VC . . . 17
3.2.1 Deep Belief Networks (DBNs) . . . 18
3.2.2 Problems . . . 20
3.3 GMM-CWT-based Emotional VC . . . 20
3.3.1 Gaussian Mixture Model (GMM) . . . 20
3.3.2 Problems . . . 22
3.4 NMF-CWT-based Emotional VC . . . 22
3.4.1 Non-negative Matrix Factorization (NMF) . . . 22
3.4.2 Method . . . 25
3.4.3 Problems . . . 27

4 Emotional VC using Arbitrary-Scales Wavelet Transform 29
4.1 The Motivation and Related Work . . . 29
4.2 Proposed Method . . . 30
4.2.1 Arbitrary Scales Continuous Wavelet Transform . . . 30
4.2.2 Training Model . . . 35
4.2.2.1 Neural Networks . . . 35
4.2.2.2 Deep Belief Networks . . . 36
4.2.3 Framework of the emotional VC using deep learning models 38
4.3 Experiments . . . 40
4.3.1 Experimental Setup . . . 40
4.3.2 Objective Experiment . . . 42
4.3.3 Subjective Experiment . . . 46
4.4 Chapter Conclusions . . . 48

5 Emotional VC using Cross-Wavelet Transform 51
5.1 The Motivation and Related Work . . . 51
5.2 Proposed Method . . . 52
5.2.1 Cross Wavelet Transform . . . 52
5.2.2 Applying CWT and XWT in F0 features extraction . . . 54
5.3 Experiments . . . 56
5.3.1 Selecting Cross Wavelet Transform Features . . . 57
5.3.2 Objective experiment . . . 59
5.3.3 Subjective experiment . . . 61

5.4 Chapter Conclusions 62

6 Emotional VC using Dual Supervised GANs **63**
6.1 The Motivation and Related Work 63
6.2 Proposed Method 64
6.3 Experiments 69
6.3.1 Training Procedure 71
6.3.2 Objective Experiment 73
6.3.3 Subjective Experiment 78
6.3.3.1 Similarity 78
6.3.3.2 Naturalness 80
6.4 Chapter Conclusions 81

7 Conclusions **83**

References **85**

Acknowledgements **95**

BibTeX Citation for This Thesis **97**

List of Figures

1.1 Comparisons of the scaling method in our proposed AS-CWT and basic CWT methods. s, τ, and ω represent the scaling factor, translating factor and the central frequency of rescaled mother, respectively. (a) Original logarithmically-spaced scales CWT (from 1 octave (2^1) to 10 octaves (2^{10})) for most existing emotional VC approaches; (b) Duration based scales calculated with the optimized duration represent from phone to sentence levels; and (c) Proposed Arbitrary-scales CWT (AS-CWT) calculated by sub-decomposing optimized duration with the optimized numbers of scales for each temporal level. Here, λ is the numbers of the scale. 4

1.2 Flow of this thesis 5

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). 9

2.2 Formant extraction using cepstrum analysis. This example uses a speech signal where “a” is uttered. 11

2.3 Mel-scale filter banks 12

2.4 Modifying a speech signal using STRAIGHT. 15

LIST OF FIGURES

3.1 Our proposed voice conversion architecture, combined with two different DBNs and concatenating NNs. A source feature vector x is fed to DBN_x, NN_{xy}, and DBN_y in order, and then converted to a target vector y. This figure shows an example of architectures with two hidden layers in the DBNs and with one hidden layer in the NNs. σ indicates a standard sigmoid function, i.e. $\sigma(x) = 1./(1 + \exp(-x))$. 18

3.2 Example of a Gaussian mixture model ($M = 2$). 21

3.3 Non-negative matrix factorization. 23

3.4 The framework of the proposed exemplar-based spectrum and F0 conversion. The spectrum (SP) and the five-scale fundamental frequency representation (F0) are recombined following Eq. (3.34) to build source features. Dictionary is constructed according to Eq. (3.33), and the activation matrix is estimated using the update rule in Eq. (3.32) 25

4.1 Log-normalized F0 (A) and interpolated log-normalized F0 (B). The Red curve: target F0; The Blue curve: source F0. 31

4.2 Example of performing segmentation in the training data. Here, X_s, X_p and X_w represent the durations of sentence, phrase and word, respectively. 32

4.3 Gaussian distributions of duration in sentence, phrase, and word. 33

4.4 Example of sentence, phrase, word, syllable, and phone level scales when the number of each level (λ) is set to 3. The red, bule, and yellow curves represent the scales in each level when temporal duration (D_i) is calculated with i=1, i=2, and i=3, respectively. . 35

4.5 Training model of the AS-CWT method. 37

4.6 Framework of the proposed method. 39

4.7 RMSE as a function of number of scales 40

4.8 Examples of original F0 signal and F0 signal reconstructed with different numbers of scales. The top graph shows the original F0 signal (Os) and the F0 signal reconstructed with eight scales in each temporal level (R8 has 8×5 scales in total); The second graph shows Os and the reconstructed signal with five scales in each temporal level (R5 has 5×5 scales in total); The bottom pan shows Os and reconstructed signal with three scales in each temporal level (R3 has 3×5 scales in total). 41
4.9 Mel-cepstral distortion evaluation of spectral features conversion from emotional to neutral voices using the DBNs model 44
4.10 Root mean squared error evaluation of F0 features conversion from emotional to neutral voices using the AS-CWT method 45
4.11 Mel-cepstral distortion evaluation of spectral features conversion from neutral to emotional voices using the DBNs model 45
4.12 Root mean squared error evaluation of F0 features conversion from neutral to emotional voices using the AS-CWT method 47
4.13 Examples of source F0, target F0 and converted F0 47
4.14 MOS evaluation of emotional voice conversion using the AS-CWT method 49

5.1 (1) Linear interpolation and log-normalized processing for source and target emotional voice. (2) The continuous wavelet transforms for normalized LogF0 contours, the top pan and bottom pan show five examples of the different level's decomposed CWT-F0 features of the source and target F0, respectively. (3) First and second pans show the continuous wavelet spectrums of the source and target CWT-F0 features and the bottom pan represents the cross-wavelet spectrums of the two CWT-F0 features. The relative phase relationship is shown as arrows in the cross-wavelet spectrum (with in-phase pointing right ($\rightarrow$), anti-phase pointing left ($\leftarrow$), and source emotional features leading target emotional features by 90° pointing straight down. 55

5.2 The listening experiment setup for evaluating emotional voice generated by different XWT-F0 features. The small letters in the lower right of CWT-F0 and XWT-F0 represent the emotion. For instance, (n) and (a) represent the neutral and angry. (n,a) represents the XWT-F0 features generated by CWT-F0 features of angry and neutral. 58
5.3 Similarity to different emotional voices (neutral, angry, happy and sad) of synthesized emotional voices with different XWT-F0 features. 58
5.4 MOS evaluation of emotional voice conversion using the XWT method 61

6.1 Illustration of calculating the loss of dual-SANs, $\|G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y}\|_1$ and $\|G_B(\boldsymbol{y}, \hat{\boldsymbol{z}}) - \boldsymbol{x}\|_1$ represent the distance losses for conversion function from data $\boldsymbol{x}$ to $\boldsymbol{y}$ and the dual task (data $\boldsymbol{y}$ to $\boldsymbol{x}$), respectively. $|\boldsymbol{x} \times G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y} \times G_B(\boldsymbol{y}, \hat{\boldsymbol{z}})|$ represents the dual loss. . . 66
6.2 Spectrograms and F0 of the source, target and converted emotional voices by different methods. 77
6.3 Results of XAB test for similarity evaluation of the conversion from angry to neutral voice by dual-SANs with different features. . . . 78
6.4 MOS of the naturalness evaluation for the conversion of emotional voice to neutral voice using the dual-SANs model 80
6.5 MOS of the naturalness evaluation for the conversion of neutral voice to emotional voice using the dual-SANs model 81

List of Tables

4.1 MCD and F0-RMSE results for different emotions. A2N, S2N and H2N represent the datasets angry to neutral voice, sad to neutral voice and happy to neutral voice, respectively. 43

4.2 MCD and F0-RMSE results for different emotions. N2A, N2S and N2H represent the datasets neutral to angry voice, neutral to sad voice and neutral to happy voice, respectively. 46

4.3 MOS results for different emotions. A2N, S2N and H2N represent the datasets angry to neutral voice, sad to neutral voice and happy to neutral voice, respectively. 48

5.1 F0-RMSE results for different emotions. N2A, N2S and N2H represent the datasets from neutral to angry, sad and happy voice, respectively. (40) and (80) denote the number of training examples. "+" means using both the CWT-F0 features and selected XWT-F0 features for training model 60

5.2 Results of emotion classification for recorded (original) voices and converted voice by different methods, N, A, S and H represents Neutral, Angry, Sad and Happy, respectively [%]. 61

6.1 Details of network architectures of G_A, G_B, D_x, D_y. 72

6.2 MCD and F0-RMSE results for the conversion of emotional voice to neutral voice. 73

6.3 MCD and F0-RMSE results for the conversion of neutral voice to emotional voice. 74

6.4 MCD and F0-RMSE results for the conversion of angry voice to neutral voice with different data size. 74
6.5 Results of classification for recorded (original) voices and converted voice by different methods [%]. 79

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.* xvii-xx.

Publication List

Journal Papers

1. Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki: "Emotional Voice Conversion using Dual Supervised Adversarial Networks with Continuous Wavelet Transform F0 Features", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 27, No. 10, pp. 1535-1548, 2019.

2. Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki: "Neutral-to-emotional voice conversion with cross-wavelet transform F0 using generative adversarial networks ", *APSIPA Transactions on Signal and Information Processing*, doi.org/10.1017/ATSIP.2019.3, Vol. 8, 11 pages, March 2019.

3. Zhaojie Luo, Xiaojing Cai, Katsuyuki Tanaka, Tetsuya Takiguchi, Takuji Kinkyo and Shigeyuki Hamori: "Can We Forecast Daily Oil Futures Prices? Experimental Evidence from Convolutional Neural Networks", *Journal of Risk and Financial Management*, Vol. 12, Issue 1, 13 pages, DOI 10.3390, 2019.

4. Jinhui Chen, Zhaojie Luo (co-first author), Zhihong Zhang, Faliang Huang, Zhiling Ye, Tetsuya Takiguchi, Edwin R. Hancock: "Polar Transformation on Image Features for Orientation-Invariant Representations", *IEEE Transactions on Multimedia*, Vol. 21, Issue 2, pp. 300-313, DOI: 10.1109, 2018.

5. Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi and Yasuo Ariki: "Emotional Voice Conversion Using Neural Networks with Arbitrary-Scales F0 based on Wavelet Transform", *EURASIP Journal on Audio, Speech, and Music Processing*, 13 pages, doi.org/10.1186/s13636-017-0116-2, Aug. 2017

6. Jinhui Chen, Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "Multithreading Cascade of SURF for Facial Expression Recognition", *EURASIP Journal on Image and Video Processing*, Vol. 2016 (1), 2016:37, doi: 10.1186, 13 pages, Oct. 2016.

International Conference Papers

1. Zhaojie Luo, Jinhui Chen, Xiao Jing Cai, Katsuyuki Tanaka, Tetsuya Takiguchi, Takuji Kinkyo, Shigeyuki Hamori: "Oil Price Forecasting Using Supervised GANs with Continuous Wavelet Transform Features", *ICPR 2018*, pp.850-835, 2018.

2. Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, Yasuo Ariki: "Emotional Voice Conversion with Adaptive Scales F0 based on Wavelet Transform using Limited Amount of Emotional Data", *INTERSPEECH 2017* , pp. 3399-3403, Aug. 2017.

3. Jianfei Xue, Zhaojie Luo (co-first author), Koji Eguchi, Tetsuya Takiguchi, and Tsukasa Omoto: "A BAYESIAN NONPARAMETRIC MULTIMODAL DATA MODELING FRAMEWORK FOR VIDEO EMOTION RECOGNITION", *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 601-606, July 2017.

4. Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, Yasuo Ariki: "FACIAL EXPRESSION RECOGNITION WITH DEEP AGE", *The Second Workshop on Human Identification in Multimedia (HIM' 17), ICMEW 2017*, pp. 657-662, 2017.

5. Jinhui Chen, Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "Expression Recognition with Ri-HOG Cascade ", *Third Workshop on Computer Vision for Affective Computing*, 14 pages, Nov. 24, 2016.

6. Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki, Toru Nakashika: 'Emotional Voice Conversion Using Neural Networks with Different Temporal Scales of F0 based on Wavelet Transform", *9th ISCA Speech Synthesis Workshop*, pp. 153-158, Sept. 13-15th, 2016.

7. Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "Emotional Voice Conversion Using Deep Neural Networks with MCC and F0 Features", *IEEE/ACIS International Conference on Computer and Information Science (ICIS)*, pp. 977-981, June 26-29, 2016.

8. Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki: "Rotation-invariant Histograms of Oriented Gradients for Local Patch Robust Representation", *APSIPA Annual Summit and Conference*, pp. 196-199, Dec. 2015.

Domestic Conference Papers

1. Zhaojie Luo, Ryoichi Takashima, Tetsuya Takiguchi, Yasuo Ariki: "Speech-to-Speech Translation using Dual Learning and Prosody Conversion ", *Acoustical Society of Japan 2019 Autumn Meeting*, 1-P-30, pp. 1037-1040, 2019. (in Japanese)

2. Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "Speech Prosody Conversion using Sequence Generative Adversarial Nets with Continuous Wavelet Transform F0 features", *Acoustical Society of Japan 2019 Spring Meeting*, 1-P-35, pp. 1125-1128, 2019. (in Japanese)

3. Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "Neutral-to-Emotional Voice Conversion with Latent Representations of F0 using Generative Adversarial Networks", *Acoustical Society of Japan 2018 Autumn Meeting*, 1-R-43, pp. 1191-1194, 2018. (in Japanese)

0. PUBLICATION LIST

4. Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "EMOTIONAL VOICE CONVERSION WITH WAVELET TRANSFORM USING DUAL SUPERVISED ADVERSARIAL NETWORKS", *Acoustical Society of Japan 2018 Spring Meeting*, 2-Q-35, pp. 365-368, 2018. (in Japanese)

5. Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "Emotional Voice Conversion with Adaptive Scales F0 based on Wavelet Transform using Limited Amount of Emotional Data", *Acoustical Society of Japan 2017 Autumn Meeting*, 1-R-31, pp. 227-230, 2017. (in Japanese)

6. Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "Arbitrary-scales continuous wavelet transform for emotional voice conversion", *Acoustical Society of Japan 2017 Spring Meeting*, 2-P-41, pp. 377-380, 2017. (in Japanese)

7. Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki: "Emotional Speech Conversion Using Deep Neural Networks", *Acoustical Society of Japan 2016 Spring Meeting*, 1-R-46, pp. 351-354, 2016. (in Japanese)

Glossary

ANN	Artificial Neural Network
AS-CWT	Arbitrary-Scales Continuous Wavelet Transform
CWT	Continuous Wavelet Transform
DNN	Deep Neural Networks
Dual-SANs	Dual-Supervised Generative Adversarial Networks
DP	Dynamic Programing
DTW	Dynamic Time Warping
GANs	Generative Adversarial Networks
GMM	Gaussian Mixture Model
DBNs	Deep Belief Networks
LG	Logarithm Gaussian
MCD	Mel-Cepstral Distortion
MFCC	Mel-Frequency Cepstral Coefficients
MOS	Mean Opinion Score
NMF	Non-negative Matrix Factorization
RBM	Restricted Boltzmann Machine
RMSE	Root Mean Squared Error
VC	Voice Conversion
XWT	Cross Wavelet Transform

Chapter 1

Introduction

1.1 Background

Emotional voice conversion (VC) is a voice conversion technique for converting prosody in speech, which can represent different emotions while retaining the linguistic information. Recently, the study of VC has attracted wide attention in the field of speech processing. This technology can be applied in various domains, for example, emotion conversion [1], speech assistance [2], and other such applications [3] [4]. Therefore, it has continued to motivate related studies each year.

Normal VC tasks are designed to transform the speech of the source speaker to that of the target speaker, making the conversion speech sound like the voice of the target speaker. Many statistical approaches have been proposed in the past few decades [5] [6], and existing VC methods can be roughly divided into two categories: a) conventional shallow approaches and b) deep neural networks models. Specifically, among the shallow approaches, a Gaussian Mixture Model (GMM) has been commonly used and successfully built upon over many years [7] [8]. Other VC methods, such as those based on non-negative matrix factorization (NMF) [9], have also been proposed. However, these approaches have some limitations, specifically, they are time-consuming and require a degree of background knowledge to implement.

In recent years, deep learning has also markedly improved the performance of VC systems through learning hierarchies of features optimized for the task at

hand. One such example of a deeper VC method was proposed by Desai *et al.* [10] and based on Neural Networks (NNs). Nakashika *et al.* [11] also proposed a VC method using speaker-dependent restricted Boltzmann machines [12] or deep belief networks (DBNs) [13] to achieve non-linear deep transformation. However, deep learning models are restricted to problems with moderate dimensions and sufficient available data. Therefore, most deep learning-based VC networks do not focus on the emotional VC, which is mainly affected by low-dimensional fundamental frequency (F0) features. Thus, conventional VC models [11, 14] usually convert F0 using Logarithm Gaussian (LG) normalized transformation [15].

However, in VC tasks, the spectral and F0 features can affect the acoustic and prosodic features, respectively. Particularly in emotional VC tasks where the prosody plays an important role in conveying various types of non-linguistic information that represent the mood of the speaker, such as identity, intention, and attitude. Previous studies (*ex.* [16]) have shown that prosody conversion is affected by both short- and long-term dependencies in different temporal levels such as the phones, syllables, and words, within an utterance. The LG-based method is insufficient to convert prosody effectively because of the constraints of their linear models and low-dimensional F0 features.

It has been shown that continuous wavelet transform (CWT) can effectively model F0 in different temporal scales and significantly improve speech synthesis performance [17]. Therefore, Suni *et al.* [18] applied CWT for intonation modeling in Hidden Markov Model speech synthesis. Ming *et al.* used CWT in F0 modeling within the NMF model [19] or the DBLSTM model [20] for converting emotional VC. Thus, in this thesis, we also applied the CWT in the F0 features processing.

1.2 Approaches

The CWT was originally proposed by Goupillaud *et al.* [21], and for a 1-D input signal, the result is a 2-D description of the signal with respect to time-scale

parameter (s, τ) of the CWT function

$$CWT\{x(t); s, \tau\} = \int x(t)\psi^*_{s,\tau}(t)dt$$
$$\psi_{s,\tau}(t) = s^{-1/2}\psi\left(\frac{t-\tau}{s}\right) \tag{1.1}$$

where $*$ stands for complex conjugate, $\psi_{s,\tau}(t)$ is mother wavelet with the scaling factor s and translating factor τ. As shown in Figure 1.1, scaling factor s controls the width of the wavelet and translating factor τ decides the location of the wavelet. The scale s is inversely proportional to the central frequency (ω) of the rescaled mother wavelet. As shown in Figure 1.1 (a), in most of the existing emotional VC models [18, 19, 20], they decompose the 1-D F0 to 2-D CWT-F0 features using logarithmically-spaced scales, each separated by one octave (2^n).

In this thesis, we assume that only those features that can represent the utterance, phrase, word, syllable, and phone levels are useful for emotional VC training. Thus, we proposed the Arbitrary Scales Continuous Wavelet Transform (AS-CWT) method to replace the logarithmically-spaced scales, which are shown in Figure 1.1 (c). These arbitrary scales are better than the original logarithmically-spaced scales at decomposing F0 to CWT-F0 features, which represent different prosodic levels ranging from micro-prosody to the sentence levels. Then, to overcome the limitation of sufficient emotional voice data, in this thesis, we outline the practical elements of the cross-wavelet transform (XWT) method, which can combine different emotional voice features to process the new emotional features.

Moreover, to improve the emotional VC effect with a limited amount of training data, we propose dual supervised adversarial networks (dual-SANs), which combine generative adversarial networks (GANs) [22] with dual supervised learning [23] to train the Mel Cepstral Coefficients (MCC) and CWT-F0 features. The effectiveness of GANs is due to the fact that an adversarial loss forces the generated data to be indistinguishable from real data. This is particularly powerful for image generation tasks, and has begun to be applied in speech synthesis [24] and VC [25]. Dual supervised learning is a technique for supervised learning tasks in dual forms, which was proposed by Xia *et al.* [23] in 2017. Prior to dual supervised learning, He *et .al* proposed the dual-learning mechanism [26] in 2016. The dual learning method uses one agent to represent the model for the primal task

and another agent to represent the model for the dual task. They are then asked to teach each other through a reinforcement learning process.

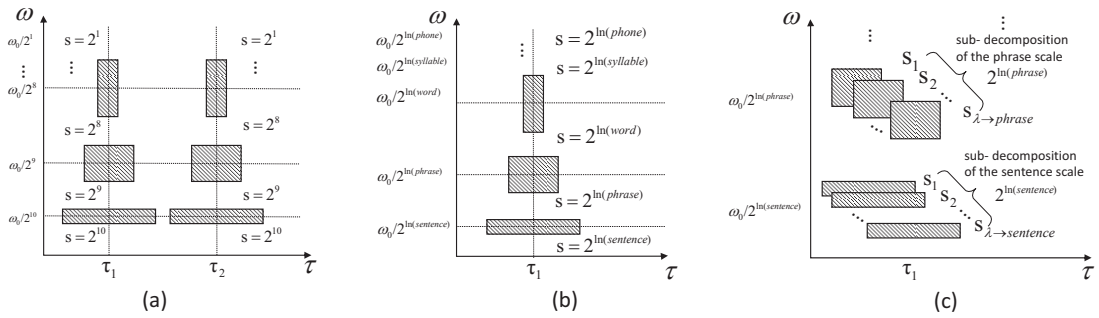

Figure 1.1: Comparisons of the scaling method in our proposed AS-CWT and basic CWT methods. s, τ, and ω represent the scaling factor, translating factor and the central frequency of rescaled mother, respectively. (a) Original logarithmically-spaced scales CWT (from 1 octave (2^1) to 10 octaves (2^{10})) for most existing emotional VC approaches; (b) Duration based scales calculated with the optimized duration represent from phone to sentence levels; and (c) Proposed Arbitrary-scales CWT (AS-CWT) calculated by sub-decomposing optimized duration with the optimized numbers of scales for each temporal level. Here, λ is the numbers of the scale.

1.3 Purpose of This Thesis

This paper proposes three emotional VC algorithms for processing the F0 features or improving the training models. All our methods are based on deep learning models using the CWT features, which provide a natural-sounding converted emotional voice. The three methods correspond to the following keywords; adaptive scaling features, cross wavelet transforming and dual supervised learning. Figure 1.2 shows the relationships of our proposed method.

1.3.1 Adaptive Emotional Features Processing for Emotional VC

In emotional voice conversion (VC) tasks, it is difficult to deal with a simple representation of fundamental frequency (F0), which is the most important

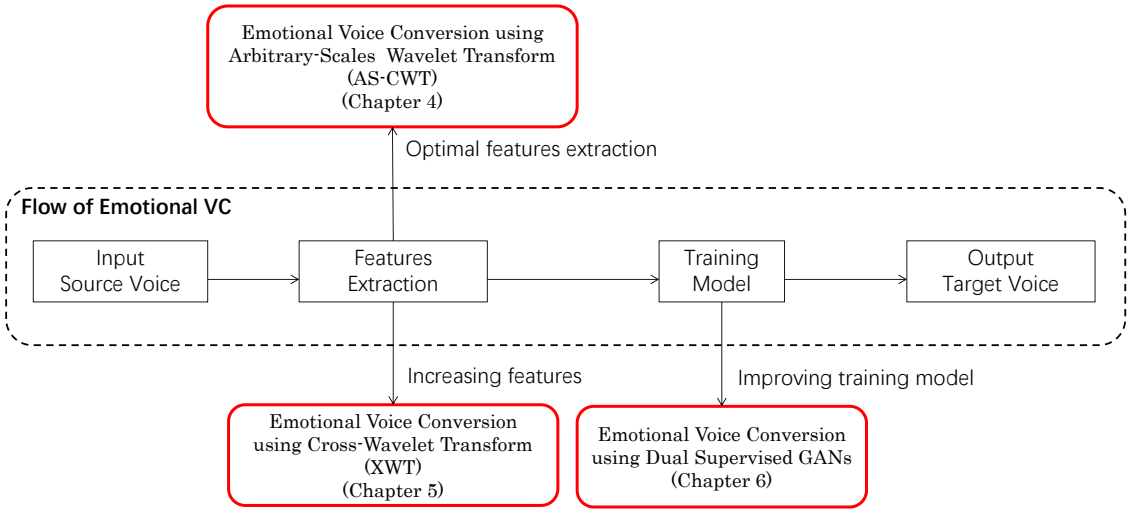

Figure 1.2: Flow of this thesis

feature in emotional voice representation. In order to address this issue, we propose the arbitrary scales CWT (AS-CWT) method to systematically capture F0 features of different temporal scales, which can represent different prosodic levels ranging from micro-prosody to sentence levels.

1.3.2 Increasing Emotional Features for Emotional VC

We propose a novel neutral-to-emotional voice conversion (VC) model that can effectively learn a mapping from neutral to emotional speech with limited emotional voice data. Although conventional VC techniques have achieved tremendous success in the spectral conversion, the lack of representations in fundamental frequency (F0), which explicitly represents prosody information, is still a major limiting factor for emotional VC. To overcome this limitation, in our proposed model, we outline the practical elements of the cross-wavelet transform (XWT) method, highlighting how such a method is applied in synthesizing diverse representations of F0 features in emotional VC. The idea is 1) to decompose F0 into different temporal level representations using Continuous Wavelet Transform (CWT); 2) to use Cross Wavelet Transform (XWT) to combine different CWT-F0 features to synthesize interaction XWT-F0 features; 3) and then use

both the CWT-F0 and corresponding XWT-F0 features to train the emotional VC model.

1.3.3 Robust Training Model for Emotional VC

In an emotional VC task, each dataset is paired with the labeled emotional voice and neutral voice, which can be regarded as a dual task. Owing to 1) dual supervised learning's ability to improve the training performances by using the leveraging probabilistic connection between the dual tasks to enhance the learning from labeled data and 2) generative adversarial networks' (GANs') ability to mitigate the over-smoothing problem caused in the low-level data space when converting the acoustic features, we further present a novel training framework for emotional VC using GANs combined with dual supervised learning, named as dual supervised adversarial networks (dual-SANs). In emotional VC experiments, we confirmed the high similarity performance of our method when using limited labeled data for emotional VC. Our method achieves good and consistent performance, in both objective and subjective evaluations.

1.4 Thesis Outline

The objective of this thesis is to outline the progress of emotional voice conversion, which is a great challenge in the area of voice conversion. For dealing with the insufficient emotional voice data, advantages of creating the AS-CWT F0 features and XWT F0 features, and the dual-SANs training models for improving the naturality and clarity of emotional voice conversion are constant motivation for exploring this field of research and will be mentioned in this thesis.

The content of this thesis is divided into 6 parts. We list some basic ideas and technologies related to emotional voice conversion in Chapter 2. Chapter 3 is about related works, state-of-the-art for emotional voice conversion researches will be introduced, the summary of these research results and their advantages and disadvantages will be listed. Later, in Chapter 4, AS-CWT F0 features extraction for emotional VC are described. Chapter 5 describes the cross-wavelet

transform method for emotional VC. In Chapter 6, a novel training model named dual-SANs is proposed. We draw the conclusions in Chapter 7.

Chapter 2

Acoustic Features and Basics

In this chapter, we review common acoustic features used in emotional voice conversion or other speech applications. First, cepstrum analysis, which is the basic idea to extract important information from a speech signal, is presented. Secondly, we will talk about MFCC, which is well-known features in the normal voice conversion processing. Thirdly, we will introduced the basic theory of continuous wavelet transform F0 features. Finally, analyzing method of speech using STRAIGHT is presented.

2.1 Cepstrum

Speech signal is generated from vibration of the vocal cord. The vibration passes a vocal tract of the speaker, and arrives at the listener's ears or micro-

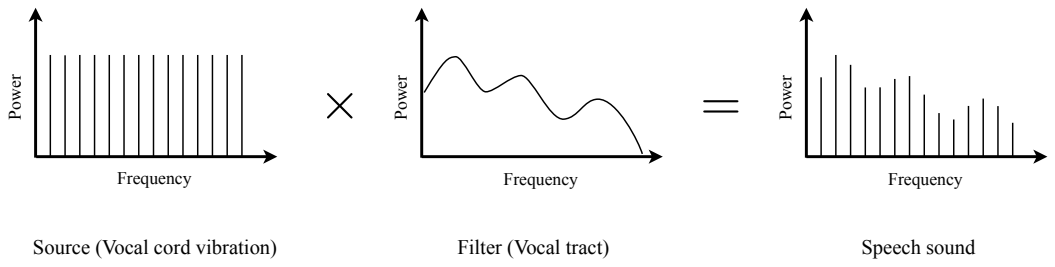

Figure 2.1: Source filter model: speech sound can be generated by multiplying spectra of the vocal cord vibration (source) and spectra of the vocal tract (filter).

phones. When we vibrate our vocal cords and change the shapes of our mouths 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 formant, and the fundamental frequency from the vibration is called pitch. In speech signal processing, the information related to the formant that determines phonemes is fairly important. Therefore, when the system recognizes a given speech, we can obtain better results with the formant information than with the original spectrum.

One well-known approach for extracting formants is cepstrum analysis [27]. 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 spectrum of vocal cord and spectrum of vocal 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 logarithm and inverse Fourier transform to Eq. (2.1), we obtain

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

and

$$S_{cep}(d) = DFT^{-1}\{\log|S(\omega)|\} \tag{2.3}$$

$$= DFT^{-1}\{\log|G(\omega)|\} + DFT^{-1}\{\log|H(\omega)|\}, \tag{2.4}$$

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

When we regard each spectra in Figure 2.1 as a signal, we notice that the vocal cord $G(\omega)$ is a signal that rapidly changes, and the vocal tract $H(\omega)$ is a signal that slowly changes, in contrast. By applying 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 a sum of the vocal cord information and the vocal tract information. Therefore, the vocal tract information $DFT^{-1}\{\log|H(\omega)|\}$ can be easily obtained by simple substruction. In other words, the formant information is extracted by liftering (low-pass filtering in quefrency) as shown in Figure 2.2.

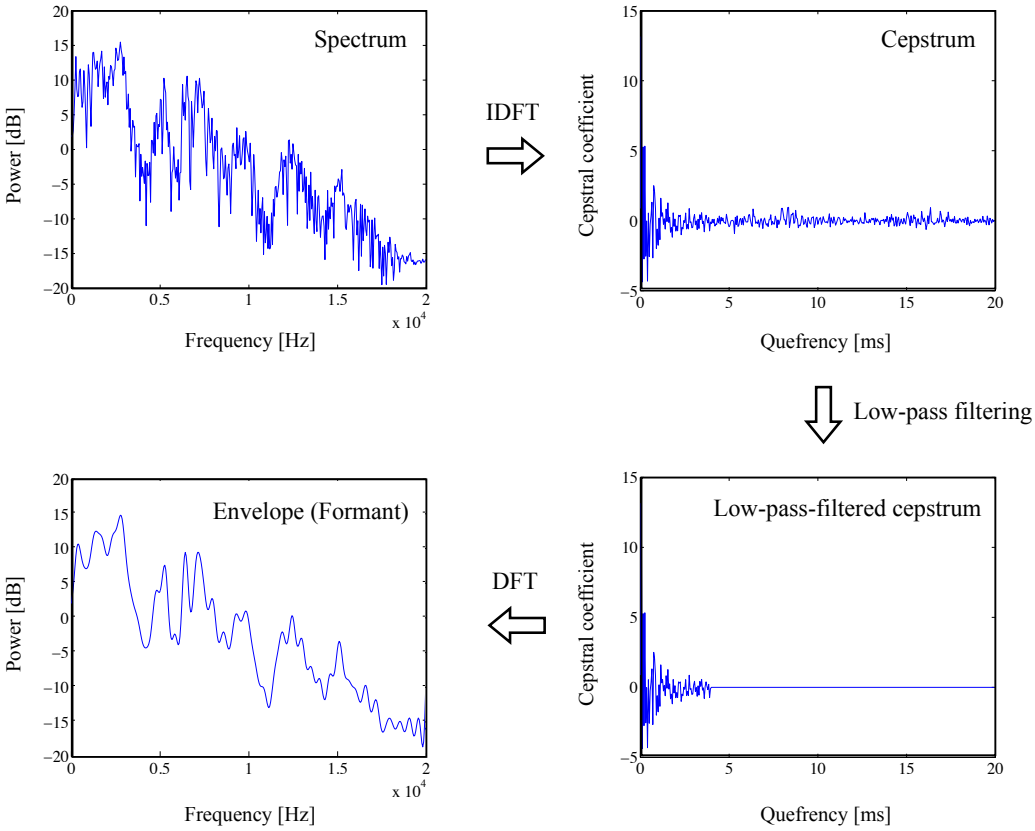

Figure 2.2: Formant extraction using cepstrum analysis. This example uses a speech signal where "a" is uttered.

2.2 MFCC

The cepstrum features introduced in the previous section was obtained as linear log spectrum (therefore, it is also called linear frequency cepstral coefficient; LFCC). On the other hand, there is another formant extraction method called mel-frequency cepstrum coefficients (MFCC) [28]. MFCC is extracted from the transformation on the mel-scale, which approximates the human-auditory-sensitive scale to pitches.

When we, human-beings, hear something, the auditory sensitivity gets poorer as the pitch is higher. That means our frequency resolution is very low in high frequency and high in low frequency. The relationship is not linear; but non-linear

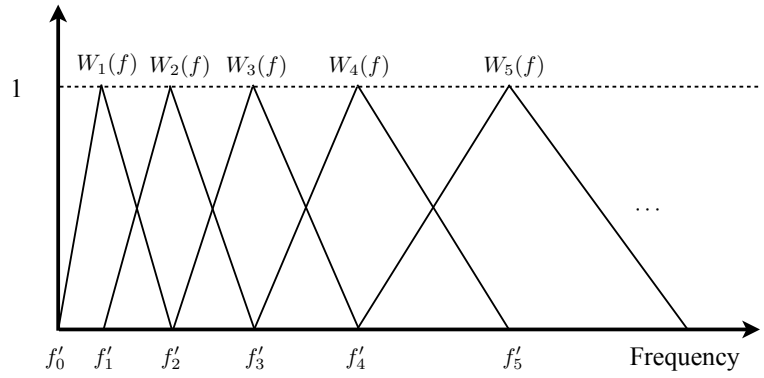

Figure 2.3: Mel-scale filter banks

as approximated by

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

where f' is called 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 has a triangle shape, and outputs the sum value of the multiplication. The power spectrum on 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 to represent the speech with small numbbers of dimensions; therefore, most works in speech signal processing, specifically in speech recognition, deal with the features. However, as the MFCC is based on filter-bank calculation, it is in general difficult to reconstruct 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 the speech reconstruction from MFCC using a source-filter model [29]. Other approaches such as an approach using z-transform [30, 31] and an approach using a mel log spectral approximation (MLSA) filter [32] have also been proposed.

2.3 F0

The fundamental frequency (also called the fundamental) of a periodic signal is the inverse (reciprocal) of the pitch period length. Fundamental frequency can be represented as F0.

F0 can represent the relationship between the fundamental frequency and the duration of the glottal pulse as follows:

$$F_0 = 1/\tau_s. \tag{2.8}$$

Here, τ_s represents the length in time of the glottal pulse or pitch period length.

The fundamental frequency is a measure of how high or low the frequency of a person's voice sounds. Its psychological correlate is pitch. It is the frequency of vocal fold vibration and correlates with changes in vocal fold tension and subglottal air pressure. A bass voice has a lower fundamental frequency than a soprano voice. A typical adult male will have a fundamental frequency of from 85 to 155Hz, and that of a typical adult female from 165 to 255Hz. Children and babies have even higher fundamental frequencies. Infants show a range of 250 to 650Hz, and in some cases go over 1000Hz. A 10-year-old boy or girl might have a fundamental frequency around 400Hz. When we speak, it is natural for our fundamental frequency to vary within a range of frequencies. This is heard as the intonation pattern or melody of natural speech. When we sing a song, we are controlling the fundamental frequency of our voice according to the melody of that song. Since a person's voice typically varies over a range of fundamental frequencies, it is more accurate to speak of a person having a range of fundamental frequencies, rather than one specific fundamental frequency, Nevertheless, a person's relaxed voice usually can be characterized by a "natural" fundamental frequency that is comfortable for that person.

2.4 Continuous Wavelet Transform

It is well established that prosody is influenced both at a supra-segmental level by long-term dependencies and at a segmental-level by short-term dependencies. It has been proven that CWT can decompose the F0 and energy contour into

several temporal scales that model prosody at different temporal levels. However, the wavelet transform is sensitive to the gaps in the F0 contour; therefore, a simple linear interpolation method is applied to obtain a continuous F0 trajectory [18]. Given a continuous F0 trajectory $f_0(t) \in L^2(\mathbb{R})$, its CWT representation is

$$W_{f_0;\psi}(s,\tau) = \langle f_0(t), \psi_{s,\tau}(t) \rangle = \int_{-\infty}^{\infty} f_0(t)\, \psi_{s,\tau}(t) dt \tag{2.9}$$

where $f_0(t)$ is the input signal and $\psi_{s,\tau}(t)$ is the mother wavelet function. Furthermore, we can reconstruct $f_0(t)$ using its CWT, $W_{f_0;\psi}(s,\tau)$, by its inverse function. (See Appendix A for a more detailed calculation of CWT-F0 and its reconstruction.)

In the computer processing of CWT, time remains continuous but time-scale parameters (s,τ) are sampled on dyadic scales [33]. Thus, the CWT used in voice conversion is computed at dyadic scales, 2^j, where $j = 1, 2, ..., 10$ [18]. The CWT can be represented as

$$\begin{aligned} W_{f_0;\psi}(j,k) &= \int_{-\infty}^{\infty} f_0(t)\, \psi_{j,k}(t) dt \\ \psi_{j,k}(t) &= 2^{-j/2} \psi(2^{-j}t - k) \\ j &= 1, 2, ..., 10, k \in \mathbb{Z} \end{aligned} \tag{2.10}$$

The original signal is approximately reconstructed by the inverse CWT transform

$$f_0(t) = \sum_{j=1}^{10} W_{f_0;\psi}(j,k) \psi_{j,k}(t) \tag{2.11}$$

2.5 STRAIGHT

STRAIGHT [34, 35] is a package tool provided by Kawahara that analyzes, modify, and synthesizes speech, written as MATLAB codes. The tool extracts three components from a speech signal: spectrum parameters, fundamental frequencies and aperiodic parameters. The analysis and synthesis quality is fairly high, so many researchers in speech signal processing loves to use it. Figure 2.4 depicts the common way to use the STRAIGHT for modifying a speech signal.

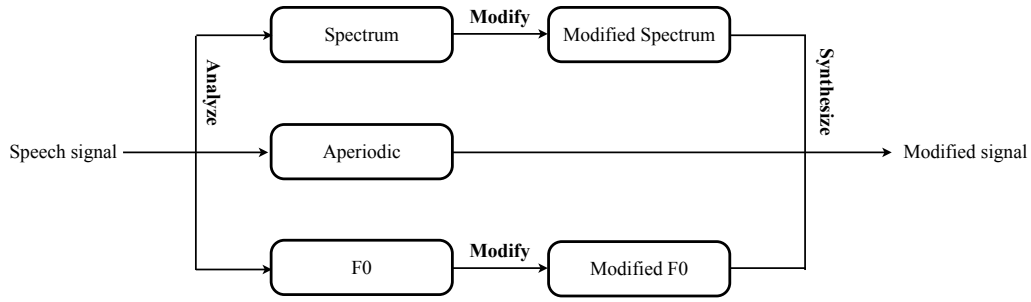

Figure 2.4: Modifying a speech signal using STRAIGHT.

The spectrum features extracted using STRAIGHT resembles formant information in cepstrum analysis. Therefore, in voice conversion, we usually modify the spectrum features and the F_0 to as the desired ones (we do not change the aperiodic features in the synthesis stage). The spectrum features are also reffered as STRAIGHT spectrum, STRAIGHT parameters, or just spectrum. In the remaining of this thesis, we refer to spectrum as the spectrum parameters extracted from a speech signal using STRAIGHT.

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 interest features, which leads to high performance in estimation of a model, and obtain high-quality speech sound.

Chapter 3

Conventional Emotional VC methods

3.1 Overview

This chapter presents conventional voice conversion (VC) methods—Nakashika *et al.* converts spectral features by DBNs and converts the F0 features through the LG method [11], Suni *et al.* [18] applied CWT for intonation modeling in hidden Markov model (HMM) speech synthesis, and Ming *et al.* [19] used CWT in F0 modeling within the NMF model for emotional VC. These approaches will be compared with the proposed methods in the remaining chapters. The problems of each method are also declared in this chapter.

3.2 DBNs-based Emotional VC

This system proposed by Nakashika *et al.* converts spectral features by Deep belief networks (DBNs) and converts the F0 features through the Logarithm Gaussian method [11], which can be expressed with the equation

$$\log\left(f0_{\text{conv}}\right) = \mu_{\text{tgt}} + \frac{\sigma_{\text{tgt}}}{\sigma_{\text{src}}} \times \left(\log\left(f0_{\text{src}}\right) - \mu_{\text{src}}\right), \tag{3.1}$$

where μ_{src} and σ_{src} are the mean and variance of the F0 in logarithm for the source speaker, respectively; μ_{tgt} and σ_{tgt} are the mean and variance of the F0

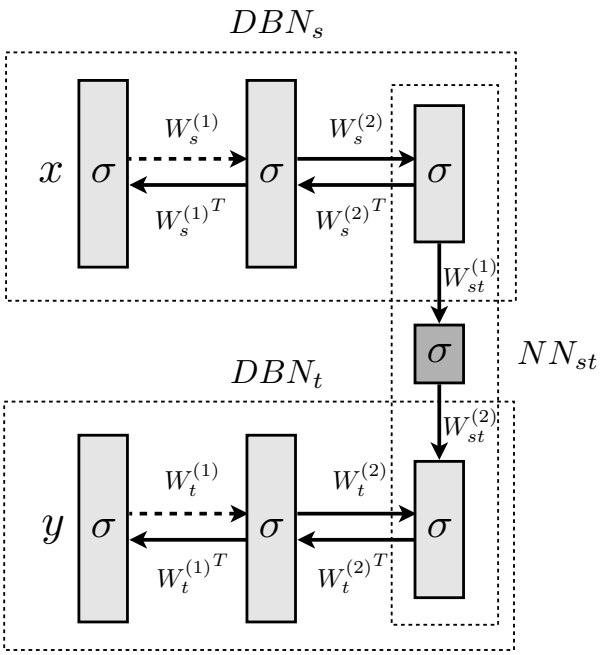

Figure 3.1: Our proposed voice conversion architecture, combined with two different DBNs and concatenating NNs. A source feature vector x is fed to DBN_x, NN_{xy}, and DBN_y in order, and then converted to a target vector y. This figure shows an example of architectures with two hidden layers in the DBNs and with one hidden layer in the NNs. σ indicates a standard sigmoid function, i.e. $\sigma(x) = 1./(1 + \exp(-x))$.

for target speaker, respectively; ($f0_{\text{src}}$) is the source speaker pitch; and ($f0_{\text{conv}}$) is the converted fundamental frequency for the target speaker.

3.2.1 Deep Belief Networks (DBNs)

Figure 3.1 shows a flow of our proposed method, where an input vector (cepstrum) of the source speaker is converted to that of a target speaker in the high-order space by using DBNs and NNs. We prepare different DBNs for source speech and target speech (DBN_s and DBN_t, respectively) so as to capture the speaker-individuality information. All the DBNs are trained using the corresponding speaker's training data. As shown in Figure 3.1, DBNs stack multiple

layers (L layers) and share the weights bottom-up and top-down[1]. Given weight parameter matrices $\boldsymbol{W}_s^{(l)}$ and $\boldsymbol{W}_t^{(l)}(l = 1, 2, \ldots, L)$ for DBN_s and DBN_t, respectively, bottom-up conversion functions $\zeta_s(\boldsymbol{z})$ and $\zeta_t(\boldsymbol{z})$ can be represented by:

$$\zeta_i(\boldsymbol{z}) = (\zeta_i^{(1)} \circ \zeta_i^{(2)} \circ \cdots \circ \zeta_i^{(L)})(\boldsymbol{z}) \tag{3.2}$$

$$= \bigodot_{l=1}^{L} \zeta_i^{(l)}(\boldsymbol{z}) \tag{3.3}$$

$$\zeta_i^{(l)}(\boldsymbol{z}) = \sigma(\boldsymbol{W}_i^{(l)}\boldsymbol{z}), \quad i \in \{s, t\}, \tag{3.4}$$

where $\bigodot_{l=1}^{L}$ denotes composition of L functions. For instance, $\bigodot_{l=1}^{2} \zeta_s^{(l)}(\boldsymbol{z}) = \sigma(\boldsymbol{W}_s^{(2)}\sigma(\boldsymbol{W}_s^{(1)}\boldsymbol{z}))$.

Similarly, given a high-order feature vector at the highest layer, a top-down conversion function $\zeta_i^{-1}(\boldsymbol{z})$ that brings it back to the original (cepstrum) space is given by:

$$\zeta_i^{-1}(\boldsymbol{z}) = \bigodot_{l=1}^{L} \sigma(\boldsymbol{W}_i^{(L-l+1)^T}\boldsymbol{z}). \tag{3.5}$$

In our approach, the compact-represented input vector calculated by Eq. (3.2) is converted into the high-order target space using $(I+1)$ layers perceptron NN_{st} (in Figure 3.1). Once the weight parameters $\boldsymbol{W}_{st}^{(l)}(l = 1, 2, \ldots, I)$ of NN_{st} are estimated beforehand, an input vector can be converted to:

$$\eta_{st}(\boldsymbol{z}) = \bigodot_{l=1}^{I} \sigma(\boldsymbol{W}_{st}^{(i)}\boldsymbol{z}) \tag{3.6}$$

Summarizing the above, a conversion function of our method from a source speech $\boldsymbol{x}$ to a target speech $\boldsymbol{y}$ is given by:

$$\boldsymbol{y} = \zeta_t^{-1}(\eta_{st}(\zeta_s(\boldsymbol{x}))) \tag{3.7}$$

$$= \bigodot_{l=1}^{2L+I} \sigma(\boldsymbol{W}^{(l)}\boldsymbol{x}), \tag{3.8}$$

[1]Technically, each stack is not a bidirectional model except for the highest layer; however, the architecture is approximately regarded as being a bidirectional model in this paper.

where $\boldsymbol{W}^{(l)}$ denotes an element of a set of weight parameters $\boldsymbol{W}$, where

$$\boldsymbol{W} = \{\boldsymbol{W}^{(l)}\}_{l=1}^{2L+1} \tag{3.9}$$

$$= \{\boldsymbol{W}_s^{(1)}, \cdots, \boldsymbol{W}_s^{(L)}, \boldsymbol{W}_{st}^{(1)}, \cdots, \tag{3.10}$$

$$\boldsymbol{W}_{st}^{(I)}, \boldsymbol{W}_t^{(L)^T}, \cdots, \boldsymbol{W}_t^{(1)^T}\}. \tag{3.11}$$

3.2.2 Problems

The DBNs based method is effective with spectral conversion but not the F0 features. As we know, the spectral can affect the voice's acoustic while the F0 features affect emotional features mostly. F0 features comprise one of the most important parameters for representing emotional speech, because they can clearly describe the variation of voice prosody from one pitch period to another. However, F0 features extracted from STRAIGHT are low-dimensional features that cannot be processed well by deep models, such as NMF models or DBN models. Therefore, in [11], F0 features are converted by the LG-based method. The LG-based method is insufficient to convert the prosody effectively owing to constraints of their linear models and low-dimensional F0 features.

3.3 GMM-CWT-based Emotional VC

Suni *et al.* present a system that uses wavelets to decompose the pitch contour into ve temporal scales ranging from microprosody to the utterance level. Each component is then individually trained within GMM-HMM framework and used in a superpositional manner at the synthesis stage.

3.3.1 Gaussian Mixture Model (GMM)

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

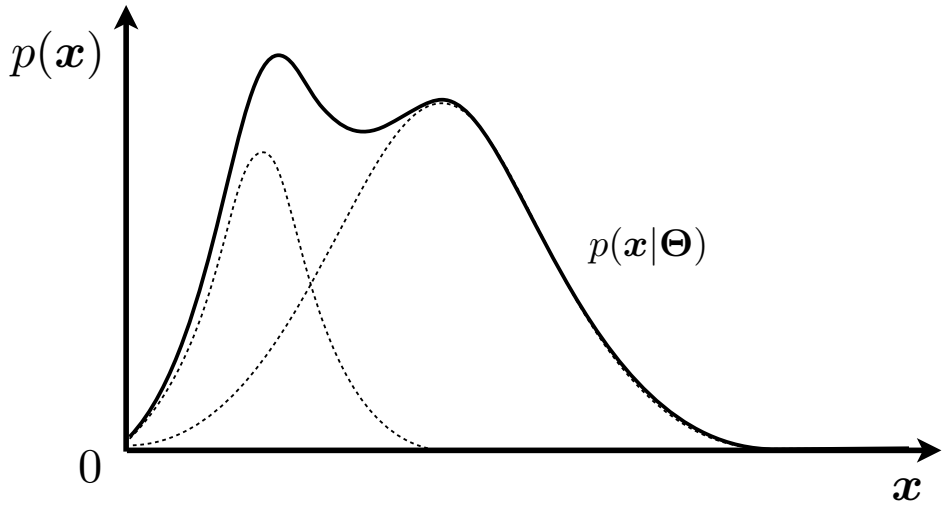

Figure 3.2: Example of a Gaussian mixture model ($M = 2$).

which is defined as

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{3.12}$$

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

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_D \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1D}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{D1}^2 & \cdots & \cdots & \sigma_{DD}^2 \end{bmatrix}, \tag{3.13}$$

and $(\cdot)^T$ and $|\cdot|$ indicate transpose and determinant of a matrix, respectively.

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

$$p(\boldsymbol{x}|\boldsymbol{\Theta}) = \sum_{m=1}^{M} \alpha_m \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{3.14}$$

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 3.2 shows an example of a one-dimensional GMM that has 2 components with a solid line. The GMM was obtained by weighted sum of two Gaussian distributions depicted with the dotted lines.

The GMM parameters can be estimated using the expectation-maximization (EM) algorithm [36].

3.3.2 Problems

It is often reported that the GMM-based voice conversion includes over-smoothing problems and over-fitting problems. The over-smoothing arises because the parameters of multiple Gaussian components are estimated by averaging the observations with similar context descriptions. As a result, the outputs distribute near the modes of each component. The over-fitting problems come from its complexity.

3.4 NMF-CWT-based Emotional VC

3.4.1 Non-negative Matrix Factorization (NMF)

Non-negative matrix factorization (NMF) [37] is a matrix decomposition method under non-negative constraints. In audio signal processing, NMF is widely used for source separation [38, 39], speech enhancement [40, 41], speech recognition [42], automatic music transcription [43, 44, 45, 46], and voice conversion [2, 9, 47, 48, 49, 50]. Common use of NMF in these approaches is that the observed signal is represented as a linear combination of atoms (also called exemplars or dictionaries), which are trained beforehand. Thanks to the non-negative constraints, a small number of atoms are selected and combined. Therefore, NMF can also be seen as a tool of sparse coding [51, 52], though there exist NMF algorithms that contains explicit sparse constraints [53, 54, 55].

The basic idea to decompose a matrix $\boldsymbol{X} \in \mathbb{R}^{D \times T}$ is to find two matrices $\boldsymbol{W} \in \mathbb{R}^{D \times R}$ and $\boldsymbol{H} \in \mathbb{R}^{R \times T}$ that minimize the distance between $\boldsymbol{X}$ and $\boldsymbol{WH}$ under non-negative constraints. Figure 3.3 depicts the decomposition process of NMF. In NMF, $\boldsymbol{W}$ is called a basis matrix, and contains R number of atoms (dictionaries, or bases) in columns. $\boldsymbol{H}$ is called an activity matrix, and indicates

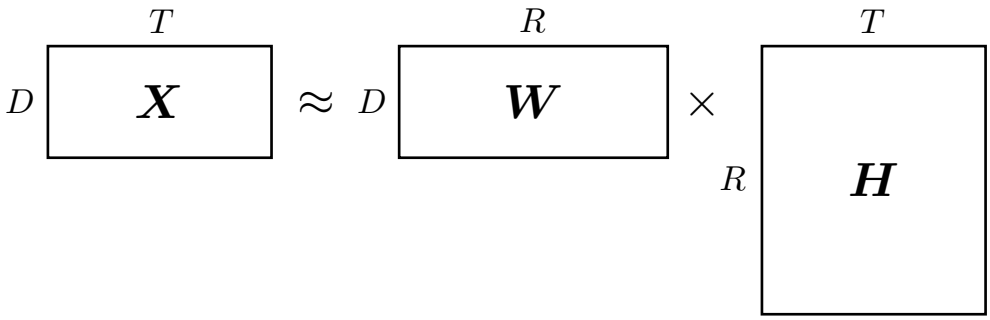

Figure 3.3: Non-negative matrix factorization.

the activity of each atom in time. The decomposition process is formulated as

$$\boldsymbol{W}, \boldsymbol{H} = \underset{\boldsymbol{W},\boldsymbol{H}}{\operatorname{argmin}} D(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H}) \tag{3.15}$$

$$\text{s.t. W, H} \geq 0, \tag{3.16}$$

where $\cdot \geq 0$ indicates that all elements in the matrix are greater than or equal to zero. $D(\boldsymbol{A}|\boldsymbol{B})$ denotes the distance between the two matrices; typically, the Euclidian distance $D_{EUC}(\boldsymbol{A}|\boldsymbol{B})$ or the Kullback-Leibler (KL) divergence $D_{KL}(\boldsymbol{A}|\boldsymbol{B})$ is used. They are, respectively, defined as

$$D_{EUC}(\boldsymbol{A}|\boldsymbol{B}) = ||\boldsymbol{A} - \boldsymbol{B}||^2 = \sum_{ij}(A_{ij} - B_{ij})^2 \tag{3.17}$$

$$D_{KL}(\boldsymbol{A}|\boldsymbol{B}) = \sum_{ij} A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}, \tag{3.18}$$

where A_{ij} denotes the ith-row jth-column element in the matrix $\boldsymbol{A}$.

Now let us consider the case where the Euclidian distance is used. Using Eq. (3.17), the objective function of NMF becomes

$$\begin{aligned} D_{EUC}(\boldsymbol{X}|\boldsymbol{W}\boldsymbol{H}) &= \sum_{dt}(X_{dt} - (\boldsymbol{W}\boldsymbol{H})_{dt})^2 \\ &= \sum_{dt}(X_{dt}^2 - 2X_{dt}(\boldsymbol{W}\boldsymbol{H})_{dt} + (\boldsymbol{W}\boldsymbol{H})_{dt}^2) \\ &= \sum_{dt}(X_{dt}^2 - 2X_{dt}\sum_{r} W_{dr}H_{rt} + (\sum_{r} W_{dr}H_{rt})^2) \end{aligned} \tag{3.19}$$

The third term in Eq. (3.19) $(\sum_r W_{dr}H_{rt})^2$ is difficult to derive partial differential. Therefore, an auxiliary function G_{EUC} that satisfies $G_{EUC}(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\lambda}) \geq$

$D_{EUC}(\boldsymbol{X}|\boldsymbol{WH})$ is used, which is defined as follows

$$G_{EUC}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\lambda}) \triangleq \sum_{dt}(X_{dt}^2 - 2X_{dt}\sum_r W_{dr}H_{rt} + \sum_r \frac{W_{dr}^2 H_{rt}^2}{\lambda_{drt}}) \tag{3.20}$$

$$s.t.\ \forall\lambda_{drt} \geq 0,\ \sum_{d,r,t}\lambda_{drt} = 1. \tag{3.21}$$

Applying Lagrange multipliers, we obtain the following update rules:

$$\lambda_{drt} \leftarrow \underset{\lambda_{drt}}{\operatorname{argmin}}\, G_{EUC} = \frac{W_{dr}H_{rt}}{\sum_{r'} W_{dr'}H_{r't}} \tag{3.22}$$

$$W_{dr} \leftarrow \underset{W_{dr}}{\operatorname{argmin}}\, G_{EUC} = \frac{\sum_t X_{dt}H_{rt}}{\sum_t \frac{H_{rt}^2}{\lambda_{drt}}} \tag{3.23}$$

$$H_{rt} \leftarrow \underset{H_{rt}}{\operatorname{argmin}}\, G_{EUC} = \frac{\sum_d X_{dt}W_{dr}}{\sum_d \frac{W_{dr}^2}{\lambda_{drt}}}. \tag{3.24}$$

Replacing λ_{drt} in Eqs. (3.23)(3.24) with Eq. (3.22), we finally obtain the common update rules for Euclidian-distance-based NMF as follows:

$$W_{dr} \leftarrow W_{dr}\frac{\sum_t X_{dt}H_{rt}}{\sum_t H_{rt}\sum_{r'} W_{dr'}H_{r't}} \tag{3.25}$$

$$H_{rt} \leftarrow H_{rt}\frac{\sum_d X_{dt}W_{dr}}{\sum_d W_{dr}\sum_{r'} W_{dr'}H_{r't}}. \tag{3.26}$$

Similarly, we can derive update rules for KL-based NMF using an auxiliary function G_{KL}, that is

$$G_{KL}(\boldsymbol{W},\boldsymbol{H},\boldsymbol{\lambda})$$

$$\triangleq \sum_{dt}(X_{dt}\log X_{dt} - X_{dt}\sum_r \lambda_{drt}\log\frac{W_{dr}H_{rt}}{\lambda_{drt}} - X_{dt} + \sum_r W_{dr}H_{rt}) \tag{3.27}$$

$$s.t.\ \forall\lambda_{drt} \geq 0,\ \sum_{d,r,t}\lambda_{drt} = 1. \tag{3.28}$$

The derived update rules for the KL-based NMF are:

$$W_{dr} \leftarrow W_{dr}\frac{\sum_t \frac{X_{dt}H_{rt}}{\sum_{r'} W_{dr'}H_{r't}}}{\sum_t H_{rt}} \tag{3.29}$$

$$H_{rt} \leftarrow H_{rt}\frac{\sum_d \frac{X_{dt}W_{dr}}{\sum_{r'} W_{dr'}H_{r't}}}{\sum_d W_{dr}}. \tag{3.30}$$

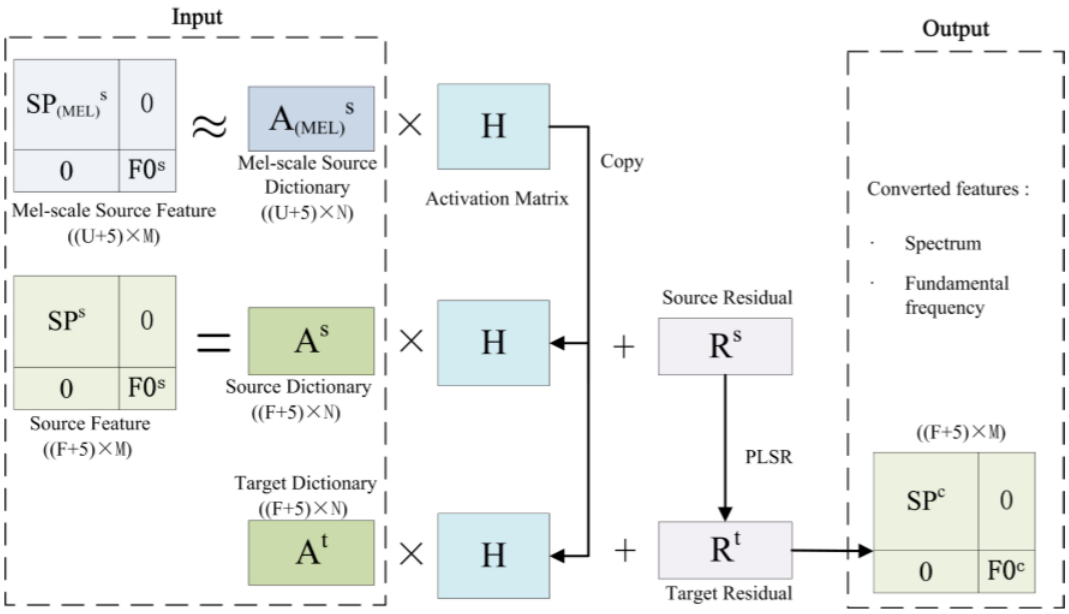

Figure 3.4: The framework of the proposed exemplar-based spectrum and F0 conversion. The spectrum (SP) and the five-scale fundamental frequency representation (F0) are recombined following Eq. (3.34) to build source features. Dictionary is constructed according to Eq. (3.33), and the activation matrix is estimated using the update rule in Eq. (3.32)

Sometimes the update rules in matrix-wise representation is useful and easy to see. They are:

$$\boldsymbol{W} \leftarrow \boldsymbol{W} \odot \frac{\frac{\boldsymbol{X}}{\boldsymbol{W}\boldsymbol{H}}\boldsymbol{H}^{\mathrm{T}}}{\mathbf{1}^{D\times T}\boldsymbol{H}^{\mathrm{T}}} \tag{3.31}$$

$$\boldsymbol{H} \leftarrow \boldsymbol{H} \odot \frac{\boldsymbol{W}^{\mathrm{T}}\frac{\boldsymbol{X}}{\boldsymbol{W}\boldsymbol{H}}}{\boldsymbol{W}^{\mathrm{T}}\mathbf{1}^{D\times T}}, \tag{3.32}$$

where $\odot$ and $\frac{\cdot}{\cdot}$ denote element-wise multiplication and division, respectively, and $\mathbf{1}^{N\times M}$ indicates a matrix with the size of $N \times M$ whose elements are all one.

3.4.2 Method

The proposed exemplar-based spectrum and F0 conversion framework is illustrated in Figure 3.4 . There are three important parts in this framework: 1) paired joint dictionary construction; 2) activation matrix estimation and feature

conversion; 3) residual compensation. We will explain the details of each part in this section.

1) Dictionary Construction

Given the parallel source and target training data, we first extract spectrum SP and fundamental frequency f0 from speech signals using STRAIGHT analysis method. Secondly, the CWT representation of F0 is obtained according to CWT method described above in Section 2.4. Then, we can get the acoustical aligned source and target spectral and prosody features SP,F0. Denote the frame aligned source and target spectrum and the five-scale representation of fundamental frequency as SP^s, $F0^s$, SP^t, $F0^t$ respectively, we can build the paired joint exemplar as:

$$a_n^s = \begin{bmatrix} sp_n^s & 0 \\ 0 & f0_n^s \end{bmatrix}, a_n^t = \begin{bmatrix} sp_n^t & 0 \\ 0 & f0_n^t, \end{bmatrix} \tag{3.33}$$

where $sp_n \in \mathbb{R}^{F\times 1}$ is the n-th column vector from matrix SP, and $f0_n \in \mathbb{R}^{5\times 1}$ is the 5 dimensional representation of fundamental frequency of the n-th speech frame, and 0 is an all zero column vector.

2) Spectrum and F0 Conversion

Both spectrum and fundamental frequency are non-negative, and they can be represented as a linear combination of basis exemplars. Moreover, the assumption that acoustically aligned source and target dictionary can share the same activation matrix is still valid here. So we can convert spectrum and fundamental frequency simultaneously in the voice conversion using NMF.

Corresponding to the new exemplar, we need to define each column of matrix X and Y as a joint super-vector:

$$x = \begin{bmatrix} sp^s & 0 \\ 0 & f0^s \end{bmatrix}, y = \begin{bmatrix} sp^t & 0 \\ 0 & f0^t, \end{bmatrix} \tag{3.34}$$

where **sp** and **f0** are the spectral and prosody features extracted from one frame of speech signal.

3) Residual Compensation

There is inevitably some modeling error between the source matrix X and the reconstructed AH. During offline training, we use a development set to calculate source and target reconstruct residual in logarithmic scale:

$$R^{(X)} = \log(X) - \log(A^s H), R^{(Y)} = \log(Y) - \log(A^t H). \tag{3.35}$$

The source-target residual pairs can be obtained using the alignment information. Then, a mapping implemented by partial least square regression (PLSR) [56] can be established as

$$R^{(Y)} \approx \mathbb{P}(R^{(X)}). \tag{3.36}$$

Take residual model in to consideration, the process of generate target feature can be formulated as

$$Y = \exp(\log(A^t H) + \mathbb{P}(R^{(X)})). \tag{3.37}$$

3.4.3 Problems

The NMF-based voice conversion in the exemplar-based approach does not use any statistical training. Therefore, an over-smoothing problem does not occur unlike GMM-based or NN-based approaches. However, this approach tends to be time-consuming and difficult to execute in real-time, because thousands of dictionaries are used in the conversion process. On the other hand, a spectral-mapping-based approach uses much fewer dictionaries in conversion. In this approach, however, over-smoothing arises due to the stastical training. The other serious problem is that NMF-based approaches cannot treat with negative-valued data. For instance, MFCCs, which allows negative values, cannot be fed as input features in these approaches. In [2, 9, 47, 48, 49, 50], they used the magnitude (power) spectrum or STRAIGHT spectrum for the acoustic features.

Chapter 4

Emotional VC using Arbitrary-Scales Wavelet Transform

The related publications for this chapter are [57, 58, 59].

In this chapter, emotional voice conversion using neural networks with arbitrary-scales (AS-CWT) F0 based on the Wavelet Transform method is described. The method is viewed as combining deep neural networks with AS-CWT F0 features, which is related to the DBNs-based and CWT-based emotional VC discussed in Chapter 3.

4.1 The Motivation and Related Work

Most of the VC-related works focus on the conversion of spectral features, rather than on the conversion of the fundamental frequency (F0). The spectral and F0 features obtained from STRAIGHT [60] can affect the voice ' s acoustic and emotional features, respectively. F0 features comprise one of the most important parameters for representing emotional speech, because they can clearly describe the variation of voice prosody from one pitch period to another. However, F0 features extracted from STRAIGHT are low-dimensional features that cannot be processed well by deep models, such as NMF models or DBN models.

Therefore, F0 features are usually converted by logarithm Gaussian (LG) normalized transformation [15] in these models. However, previous studies have shown that prosody conversion is affected by both short- and long-term dependencies, such as the sequence of segments, syllables, words within an utterance, as well as lexical and syntactic systems of a language [16]. The LG-based method is insufficient to convert the prosody effectively owing to constraints of their linear models and low-dimensional F0 features. Recently, it was proved that continuous wavelet transform (CWT) can effectively model F0 in different temporal scales and significantly improve the speech synthesis performance [17]. For this reason, Suni *et al.* [18] applied CWT for intonation modeling in a hidden Markov model (HMM) speech synthesis. Ming *et al.* [19] used CWT in F0 modeling within the NMF model for emotional VC and obtained a better result than the LG method in terms of F0 conversion. Thus, in this thesis, inspired by the ability of deep learning models to perform well in complex nonlinear feature conversion [12] and the ability of CWT to improve F0 feature conversion [19], we proposed a novel method that used NNs to train the CWT- F0 for converting the prosody of the emotional voice.

4.2 Proposed Method

4.2.1 Arbitrary Scales Continuous Wavelet Transform

The STRAIGHT is frequently used to extract features from a speech signal. Generally, the smoothing spectrum and instantaneous-frequency-based F0 are derived as excitation features for every 5 ms from the STRAIGHT [60]. To obtain the same number of frames, a dynamic time warping method is used to align the extracted features (spectrum and F0) of the source voice and target voice. Then, the aligned spectral features are translated into MCC. The F0 features produced by STRAIGHT are one dimensional and discrete. Modeling the variations of F0 in all temporal scales using linear models is difficult. Inspired by the work in [18], before training the F0 features by NNs, we adopted CWT to decompose the F0 contour into several temporal scales, which can then be used to model different prosodic levels ranging from micro-prosody to the sentence levels. In an earlier

work [57], we adopted CWT to decompose the F0 contour into 30 temporal scales before training the F0 features by NNs. The decomposed 30 dimensional features are linearly spaced scales, each separated by one third of an octave. However, only the features that can represent the utterance, phrase, word, syllable, and phone levels are useful for training. Thus, in the current paper, we apply the AS-CWT method to decompose F0 features before training them. The steps for processing details are described below.

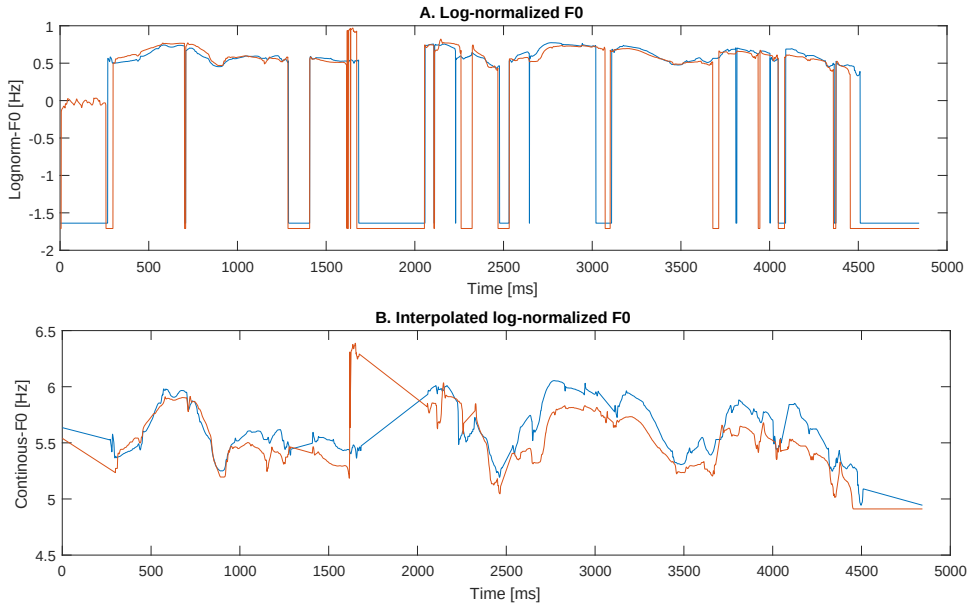

Figure 4.1: Log-normalized F0 (A) and interpolated log-normalized F0 (B). The Red curve: target F0; The Blue curve: source F0.

1) To explore the perceptually relevant information, F0 contour is transformed from the linear to logarithmic semitone scale, which is referred to as logF0. As shown in Figure 4.1 (A), the logF0 is discrete. As the wavelet method is sensitive to the gaps in the F0 contours, we must fill in the unvoiced parts in the logF0 via linear interpolation to reduce discontinuities in voice boundaries. Finally, we normalize the interpolated logF0 contour to zero mean and unit variance. An example of an interpolated pitch contour is depicted in Figure 4.1 (B).

2) Next, we calculate the scales of different prosodic levels ranging from sentence level to micro-prosody using AS-CWT method. In order to find the scales

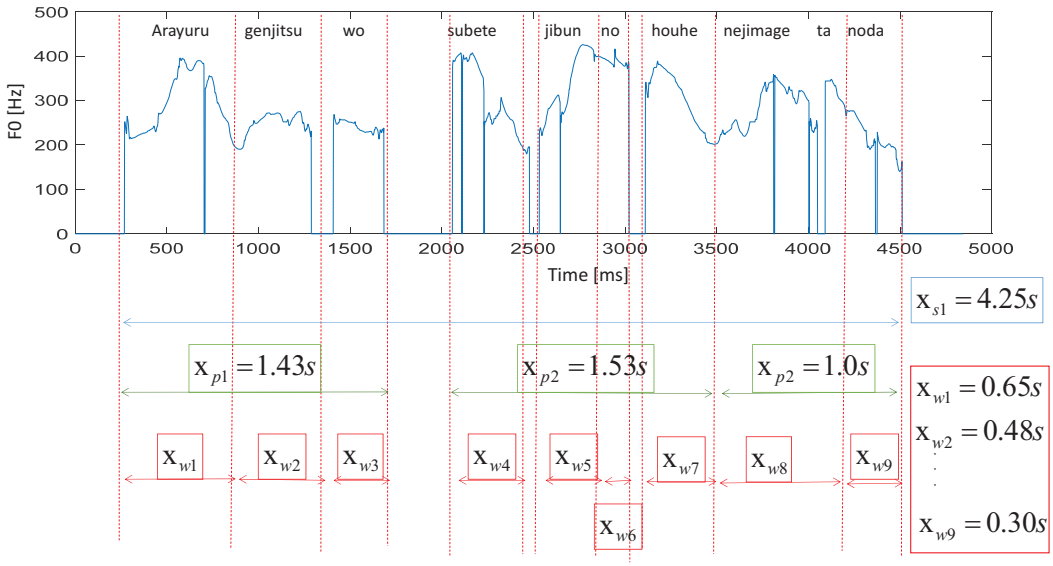

Figure 4.2: Example of performing segmentation in the training data. Here, X_s, X_p and X_w represent the durations of sentence, phrase and word, respectively.

of sentence, phrase, and word levels, we first perform segmentation in some sentences of the training data. As shown in Figure 4.2, the ranges of duration vary in sentence, phrase, and word. We use the Gaussian function to separately calculate the probability densities of the duration in the sentence, phrase, and word using

$$\begin{aligned} S(x) &= N(x, \mu_s, \sigma_s^2) \\ P(x) &= N(x, \mu_p, \sigma_p^2) \\ W(x) &= N(x, \mu_w, \sigma_w^2), \end{aligned} \tag{4.1}$$

where $S(x)$, $P(x)$ and $W(x)$ represent the probability density of duration in the sentence, phrase and word, separately. The means and standard deviations of the durations in the sentence, phrase, and word are calculated from the pre-segmented training data, as shown in Figure 4.2. Figure 4.3 represents the probability density curves of the sentence, phrase, and word. We choose the parts over 60% to process the scales of the sentence, phrase, and word levels. Each temporal duration is defined by

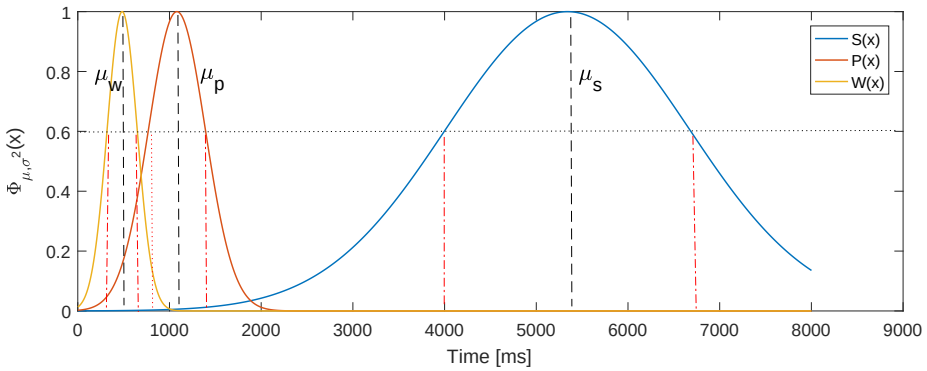

Figure 4.3: Gaussian distributions of duration in sentence, phrase, and word.

$$
\begin{aligned}
s_i &= \min(x_s) + \frac{\max(x_s) - \min(x_s)}{\lambda} \times i \\
p_i &= \min(x_p) + \frac{\max(x_p) - \min(x_p)}{\lambda} \times i \\
w_i &= \min(x_w) + \frac{\max(x_w) - \min(x_w)}{\lambda} \times i,
\end{aligned}
\tag{4.2}
$$

where s_i, p_i and w_i represent the durations of sentence, phrase, and word, respectively; x_s, x_p and x_w are the values when probability densities $S(x)$, $P(x)$, and $W(x)$, respectively, are over 60%; $i = 0, ..., \lambda$; and λ is the number of scales in sentence, phrase, and word. The average duration of non-emphasized syllables was found between 50ms and 180ms [61], and that of phone level is 20ms to 40ms. Thus, the durations of syllable and phone can be represented as $syl_i = 50 + ((180 - 50)/\lambda) \times i$ and $pho_i = 20 + ((40 - 20)/\lambda) \times i$, respectively.The scales can then be represented by

$$
\begin{aligned}
\theta_i &= \log_2(D_i/\tau_0) \\
D_i &\in \{s_i, p_i, w_i, syl_i, pho_i\},
\end{aligned}
\tag{4.3}
$$

where, $\tau_0 = 5$ms, and $\{D_i\}_{i=0,...,\lambda}$ represents all the durations of the sentence, phrase, word, syllable, and phone levels. θ_i represents each scale calculated by Eq. 4.3.

3) After calculating the scales that can model prosody at different temporal levels, we adopt CWT to decompose the F0 contour with several temporal scales. The continuous wavelet transform of F0 is defined by

$$W\left(\tau,t\right)=\tau^{-1/2}\int_{-\infty}^{\infty}f_{0}\left(x\right)\psi\left(\frac{x-t}{\tau}\right)dx \tag{4.4}$$

$$\psi\left(t\right)=\frac{2}{\sqrt{3}}\pi^{-1/4}\left(1-t^{2}\right)e^{-t^{2}/2}, \tag{4.5}$$

where $f_0\left(x\right)$ is the input signal and ψ is the Mexican hat wavelet. The original signal f0 can be recovered from the wavelet representation $W(f0)$ by inverse transform [62]:

$$f_{0}(t)=\int_{-\infty}^{\infty}\int_{0}^{\infty}W(f_{0})(\tau,x)\tau^{-5/2}\psi\left(\frac{x-t}{\tau}\right)dxd\tau \tag{4.6}$$

As described in [18], the reconstruction is incomplete if all information on $W(f0)$ is not available. In that study, the authors performed the decomposition and reconstruction by choosing ten scales, all of which are one octave apart. In our recent work in [57], we decomposed the continuous logF0 with 30 discrete scales, each separated by one third of an octave. Increasing the number of scales can result in a better reconstruction after the decomposition. However, we want to select the features to better represent the utterance, phrase, word, syllable, and phone levels, so we apply the non-linear scales θ_i calculated in Eq. 4.3, which can better represent the duration (D_i) of all levels of linguistic structure. Therefore, our F0 is represented by separate components given by

$$W_{\theta_{i}}(f_{0})(t)=W_{\theta_{i}}(f_{0})(2^{\theta_{i}+1}\tau_{0},t)\left(\theta+2.5\right)^{-5/2}, \tag{4.7}$$

The original signal is approximately recovered by

$$f_{0}=\sum_{i=0}^{\lambda}W_{\theta_{i}}f_{0}(t)(\theta_{i}+2.5)^{-5/2}+\epsilon(t), \tag{4.8}$$

where $\epsilon(t)$ is the reconstruction error, and λ represents the number of scales in each temporal level. We evaluated the accuracy of the reconstruction by decomposing and reconstructing several training sentences with different values of λ. The correlation between the original and the reconsturcted F0 signal was calculated with root mean square reconstruction error. In Section 4.3, we will describe the experiments to obtain the optimum value of λ. Figure 4.4 shows the example of $\lambda = 3$. As shown in the figure, the sentence, phrase, word, syllable and phone levels each have three scales. In this example, one-dimensional F0 feature is decomposed to 15 streams in distinct scales.

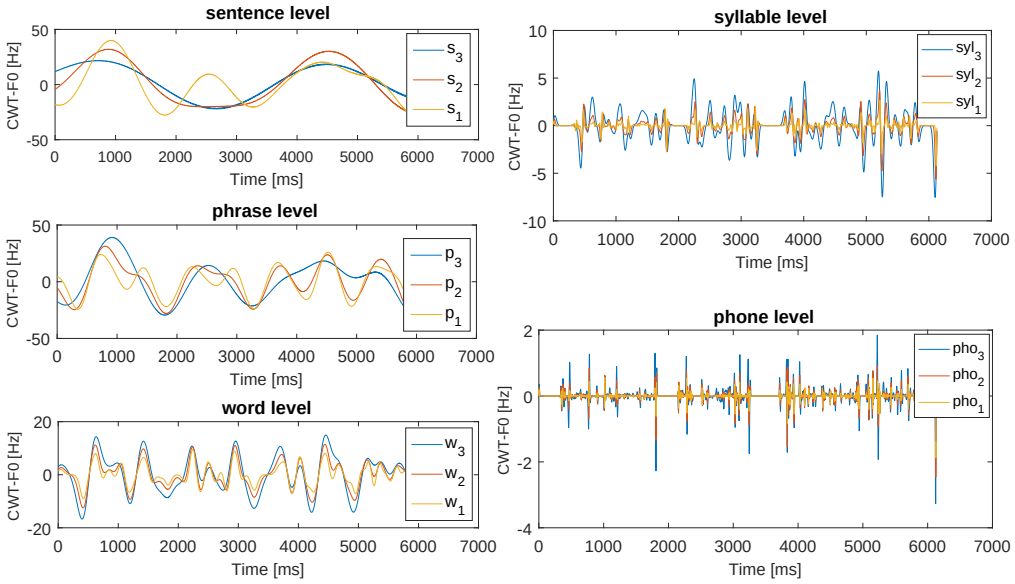

Figure 4.4: Example of sentence, phrase, word, syllable, and phone level scales when the number of each level (λ) is set to 3. The red, bule, and yellow curves represent the scales in each level when temporal duration (D_i) is calculated with i=1, i=2, and i=3, respectively.

4.2.2 Training Model

4.2.2.1 Neural Networks

Neural networks (NNs) are trained on a frame error (FE) minimization criterion and the corresponding weights are adjusted to minimize the error squares

over the whole source-target, stereo training data set. As shown in Eq. 4.9, the error of mapping is given by

$$\epsilon = \sum_{t} ||y_t - G(x_t)||^2, \tag{4.9}$$

where $G(x_t)$ denotes the NNs mapping of x_t and is defined as shown below:

$$G(x_t) = (G^1 \circ G^2 \circ \cdots \circ G^L) = \bigodot_{l=1}^{L} G^{(l)}(x_t) \tag{4.10}$$

$$G^l(x_t) = \sigma(W^l x_t) \tag{4.11}$$

In the equations above, $\bigodot_{l=1}^{L}$ denotes the composition of L functions. For instance, $\bigodot_{l=1}^{2} W^{(l)}(z) = \sigma(W^{(2)}\sigma(W^{(1)}(x_t))$. $W^{(l)}$ represents the weight matrices of layer l in NNs. σ denotes a standard tanh function which is defined as:

$$\sigma(x) = \tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}, \tag{4.12}$$

As shown in the training model of Figure 4.5, we use a 4-layer NN model for prosody training. w1, w2, and w3 represent the weight matrices of first, second and third layers of NN, respectively.

4.2.2.2 Deep Belief Networks

Deep belief networks (DBNs) have an architecture that stacks multiple Restricted Boltzmann Machines (RBMs), which are composed of a visible layer and a hidden layer with full, two-way inter-layer connections but no intra-layer connections. As an energy-based model, the energy of a configuration (v, h) is defined as :

$$E(v, h) = -a^T v - b^T h - v^T W h, \tag{4.13}$$

where $W \in R_{I\times J}$, $a \in R_{I\times 1}$, and $b \in R_{J\times 1}$ denote the weight parameter matrices between visible units and hidden units, a bias vector of visible units, and a bias

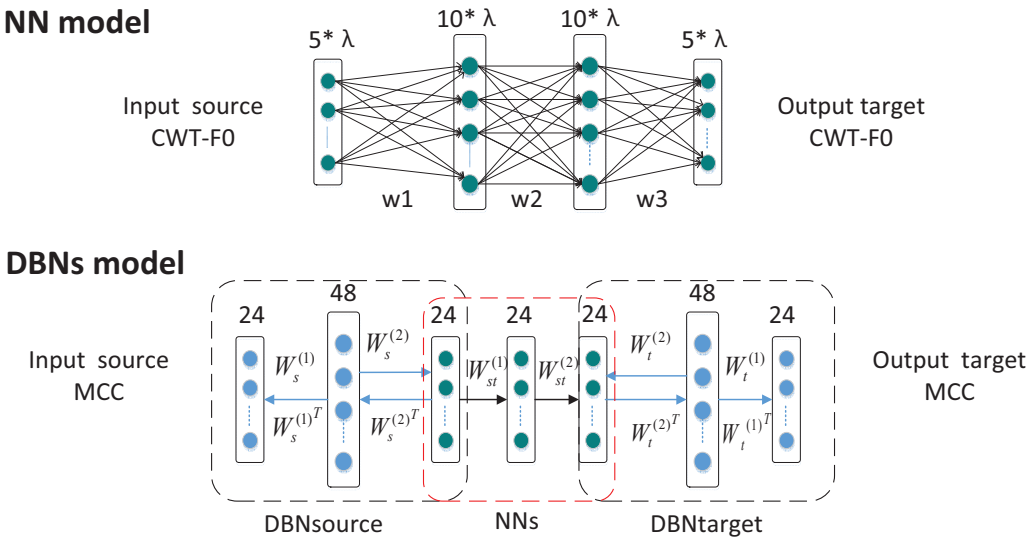

Figure 4.5: Training model of the AS-CWT method.

vector of hidden units, respectively. The joint distribution over v and h is defined as

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)}. \tag{4.14}$$

The RBM has the shape of a bipartite graph and has no intra-layer connections. Consequently, the individual activation probabilities are obtained via

$$P(h_j = 1|v) = \sigma\left(b_j + \sum_{i=1}^{m} w_{i,j} v_i\right); \tag{4.15}$$

$$P(v_i = 1|h) = \sigma\left(a_i + \sum_{j=1}^{n} w_{i,j} h_j\right). \tag{4.16}$$

In DBNs, σ denotes a standard sigmoid function given by $(\sigma(x) = 1/(1+e^{-x}))$. For parameter estimation, RBMs are trained to maximize the product of probabilities assigned to some training set data. To calculate the weight parameter matrix, we use the RBM log-likelihood gradient method, which is defined as

$$L(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log P_\theta\left(v^{(n)}\right) - \frac{\lambda}{N} \|W\|. \tag{4.17}$$

In the equation, $P_\theta\left(v^{(n)}\right)$ is the probability of visible vectors in the inner model with the model parameters $\theta = (W, a, b)$. To differentiate the $L(\theta)$ via Eq. 4.18, we can obtain W when making the $L(\theta)$ be the largest.

$$\frac{\partial L(\theta)}{\partial W_{ij}} = E_{P_{data}}[v_i h_j] - E_{P_\theta}[v_i h_j] - \frac{2\lambda}{N} W_{ij}, \tag{4.18}$$

where $E_{P_{data}}\langle\cdots\rangle$ and $E_{P_\theta}\langle\cdots\rangle$ represent the average values of the input data and the inner model, respectively. As shown in the training model of Figures 4.5, our proposed method has two different DBNs for source speech and target speech (DBNsource and DBNtarget). This is intended to capture the speaker-individuality information and connect them by the NNs. The numbers of each node from input x to output y are [24 48 24] for DBNsource and DBNtarget, respectively. The connected NN is a 3-layers model. The whole training process of the DBNs is conducted as described in the steps below.

1) We train two DBNs for the source and target speakers. In the training of DBNs, the hidden units that are computed as a conditional probability ($P(h|v)$) in Eq. 4.15 are fed to the following RBMs. These are then trained layer-by-layer until the highest layer is reached.

2) After pre-training the two DBNs separately, we connect them by the NNs. The weight parameters of NNs are estimated in order to minimize the error between the output and the target vectors.

3) Finally, the entire network (DBNsource, DBNtarget, and NNs) is fine-tuned by back-propagation using the MCC features.

4.2.3 Framework of the emotional VC using deep learning models

Our proposed framework, as shown in Figure 4.6, transforms both the excitation and the filter features from the source voice to the target voice. As described in Section 4.2.1, we extracted the spectral features and F0 features from both the source voice and target voice by the STRAIGHT, and then used DTW to align them. Next, we processed the aligned F0 features into CWT-F0 features by AS-CWT method for NNs and transform the aligned spectral features into the MCC features, respectively. The conversion function training of our proposed method

has two stages. The first stage is the conversion of CWT-F0 using the NNs, the other is the MCC conversion using the DBNs. In the first stage, we used the high dimension CWT-F0 features for prosody features training. To achieve this, we transfered the parallel data consisting of the aligned F0 features of the source and target voices to CWT-F0 features by using the AS-CWT method. Then we used the 4-layers NN models to train the CWT-F0 features. The numbers of nodes from the input layer to output layer are [5λ 10λ 10λ 5λ]. In the second stage, we first transformed aligned spectral features of source and target voices to 24-dimensional MCC features. Then, we used these MCC features of the source and target voice as the input-layer data and output-layer data for DBNs. Finally, we connected them by the NNs for deep training. The conversion phrase of Figure 4.6 shows how our trained conversion function can be applied. The source voice is processed into spectral features and F0 features by the STRAIGHT, which are then transformed to MCC and CWT-F0 features, respectively. These features can then be fed into the conversion function to convert the features. Finally, we converted them back to spectrum and F0, and used these features to reconstruct the waveform with the STRAIGHT.

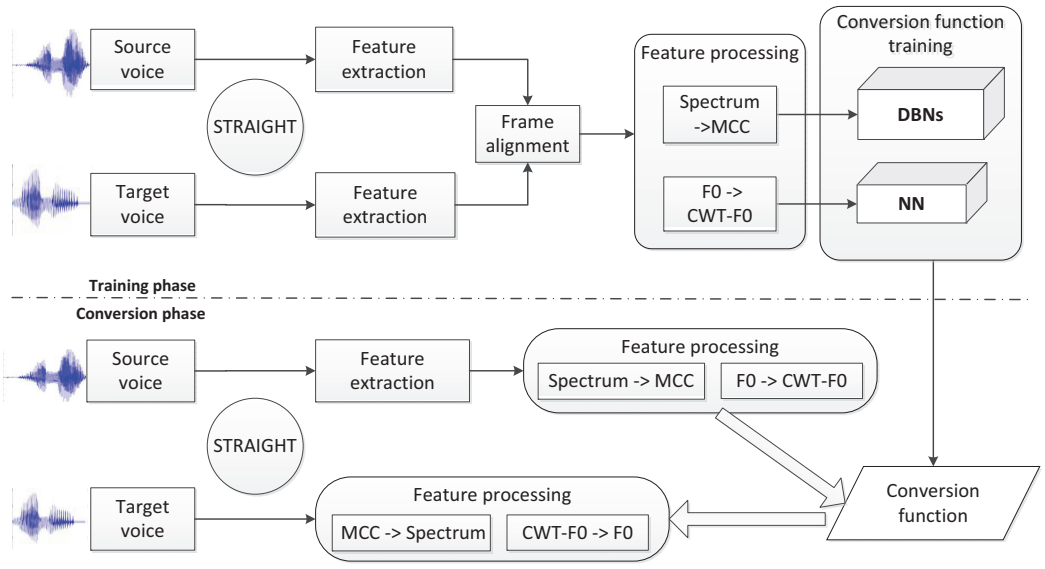

Figure 4.6: Framework of the proposed method.

4.3 Experiments

4.3.1 Experimental Setup

We used a database of emotional Japanese speech constructed in a previous study [63]. The waveforms used were sampled at 16kHz. Input and output data had the same speaker but expressing different emotions. We set the six datasets into the following: happy to neutral voices, angry to neutral voices, and sad to neutral voices, as well as their inverse conversion from neutral voices to each emotion voices. For each dataset, 50 sentences were chosen as training data and 10 sentences were choosen for the VC evaluation.

As described in Section 4.2.1, to obtain the optimum numbers of scales (λ) in each temporal level, we used the AS-CWT method to decompose and reconstruct several training sentences with the numbers of scales ranging from three to ten. We then calculated the RMSE between the original contours and the reconstructed F0 values. As shown in Figure 4.7, the value of RMSE decreases as the numbers of scales increases; however, when increasing to eight, the value of RMSE decreases slightly. Hence, we select eight scales in each temporal levels in our proposed model. As shown in Figure 4.8 shown, the reconstructed F0 with eight scales in each temporal level is more similar to original contours then three scales and five sales. To evaluate the proposed method, we compared the results with

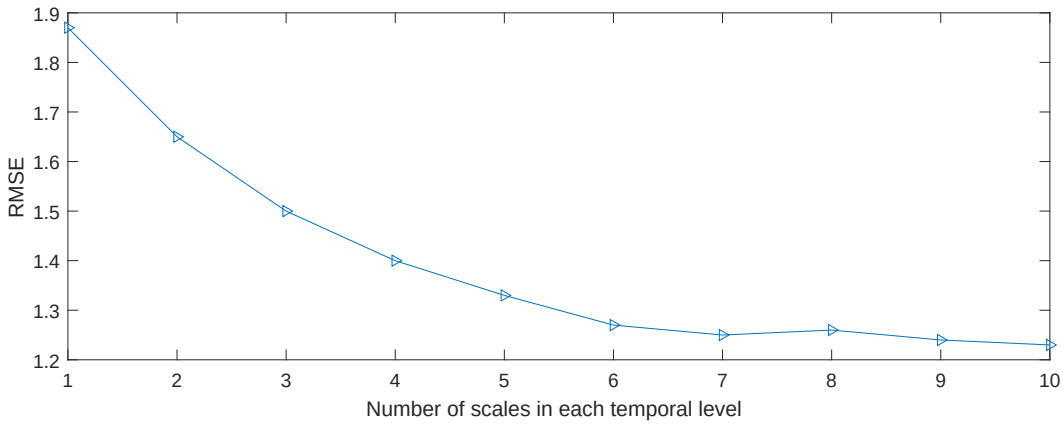

Figure 4.7: RMSE as a function of number of scales

several state-of-the-art methods listed below.

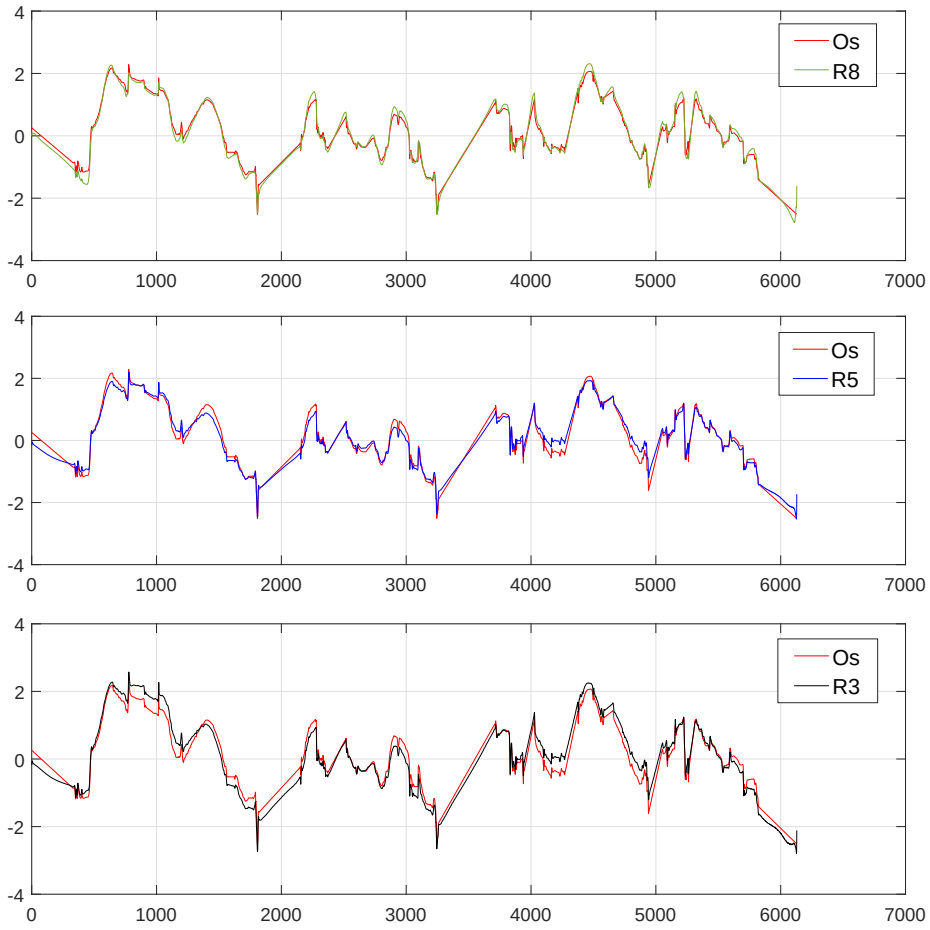

Figure 4.8: Examples of original F0 signal and F0 signal reconstructed with different numbers of scales. The top graph shows the original F0 signal (Os) and the F0 signal reconstructed with eight scales in each temporal level (R8 has 8×5 scales in total); The second graph shows Os and the reconstructed signal with five scales in each temporal level (R5 has 5×5 scales in total); The bottom pan shows Os and reconstructed signal with three scales in each temporal level (R3 has 3×5 scales in total).

- **DBNs+LG:** This system proposed by Nakashika *et al.* converts spectral features by DBNs and converts the F0 features through the LG method [11], which can be expressed with the equation

$$\log\left(f0_{\text{conv}}\right) = \mu_{\text{tgt}} + \frac{\sigma_{\text{tgt}}}{\sigma_{\text{src}}} \times \left(\log\left(f0_{\text{src}}\right) - \mu_{\text{src}}\right), \tag{4.19}$$

  where μ_{src} and σ_{src} are the mean and variance of the F0 in logarithm for the source speaker, respectively; μ_{tgt} and σ_{tgt} are the mean and variance of the F0 for target speaker, respectively; $(f0_{\text{src}})$ is the source speaker pitch; and $(f0_{\text{conv}})$ is the converted fundamental frequency for the target speaker.

- **DBNs+NMF:** Using the DBNs to convert spectral features while using the non-negative matrix factorization (NMF) to convert five-scale CWT-F0 features.

- **DBNs+CWT(30):** This is the proposed method in our previous work [57]that uses DBNs to convert spectral features while using the NNs to convert the 30-scales CWT-F0 features; each scale is separated by one third of an octave.

- **DBNs+CWT(40):** This method uses the same model as DBNs+CWT(30) with a different number of scales (40) in CWT-F0 features.

- **DBNs+AS-CWT (proposed method):** This is the proposed system that uses the DBNs to convert spectral features while using the NNs to convert the CWT-F0 features decomposed by AS-CWT method, each temporal level has eight scales (8×5 scales in total).

4.3.2 Objective Experiment

Mel Cepstral Distortion (MCD) was used for the objective evaluation of spectral conversion, MCD is defined below.

$$MCD = (10/\ln 10)\sqrt{2\sum_{i=1}^{24}\left(mc_i^t - mc_i^c\right)^2} \tag{4.20}$$

In Eq. 4.20, mc_i^t and mc_i^c represent the target and the converted mel-cepstral, respectively.
To evaluate the F0 conversion, we used the RMSE

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((F0_i^t) - (F0_i^c))^2}, \tag{4.21}$$

where $F0_i^t$ and $F0_i^c$ denote the target and the converted F0 features, respectively. A lower MCD and F0-RMSE value indicate smaller distortion or predicting error. Unlike the RMSE evaluation function used in [19], which evaluated the F0 conversion by calculating logarithmic scaled F0, we used original target F0 and converted F0 to calculate the RMSE values. Given that our RMSE function evaluates complete sentences that contain both voiced and unvoiced F0 features instead of the voiced logarithmic scaled F0, the RMSE values are expected to be high. For emotional voices, the unvoiced features also include some emotional information. Therefore, we choose the F0 of complete sentences for evaluation instead of the voiced logarithmic scaled F0.

Table 4.1: MCD and F0-RMSE results for different emotions. A2N, S2N and H2N represent the datasets angry to neutral voice, sad to neutral voice and happy to neutral voice, respectively.

	MCD [dB]			F0-RMSE [Hz]		
	A2N	S2N	H2N	A2N	S2N	H2N
Source	6.03	5.18	6.30	76.8	73.7	100.4
DBNs+LG	5.47	4.77	5.92	76.1	73.5	85.2
DBNs+NMF	5.46	4.78	5.93	69.4	66.9	74.3
DBNs+CWT(30)	5.47	4.77	5.93	61.6	64.2	75.9
DBNs+CWT(40)	5.47	4.77	5.93	62.3	67.2	76.1
DBNs+AS-CWT	5.47	4.77	5.93	51.1	52.1	64.4

The average MCD and F0-RMSE results from emotional to neutral pairs are reported in Table 4.1. The MCD results are presented in the left part of Table 4.1. Comparing DBNs with source, DBNs decrease the the value of MCD. As shown in Figure 4.9, among DBNs+LG, DBNs+NMF, DBNs+CWT(30),

DBNs+CWT(40), and DBNs+AS-CWT, MCD decreases or increases slightly, proving that the conversion of F0 does not have a significant impact on the spectral feature conversion. The F0-RMSE results are presented in the right part of Table 4.1. As shown in Table 4.1 and Figure 4.10, the conventional linear conversion logarithm Gaussian can affect the conversion of happy to neutral, but only slightly affect the conversion of angry voices and sad voices to neutral voices. The NMF method, previous CWT method, and the proposed AS-CWT method can affect the conversion of all emotional voice datasets. Comparing the results of DBN+CWT(30) and DBN+CWT(40), we see that 30 scales is the optimal number of scales for this model. Simply increasing the number of scales can not enhance the effect for DBNs+CWT. In addition, the proposed method can obtain significant improvement in F0 conversion as a whole.

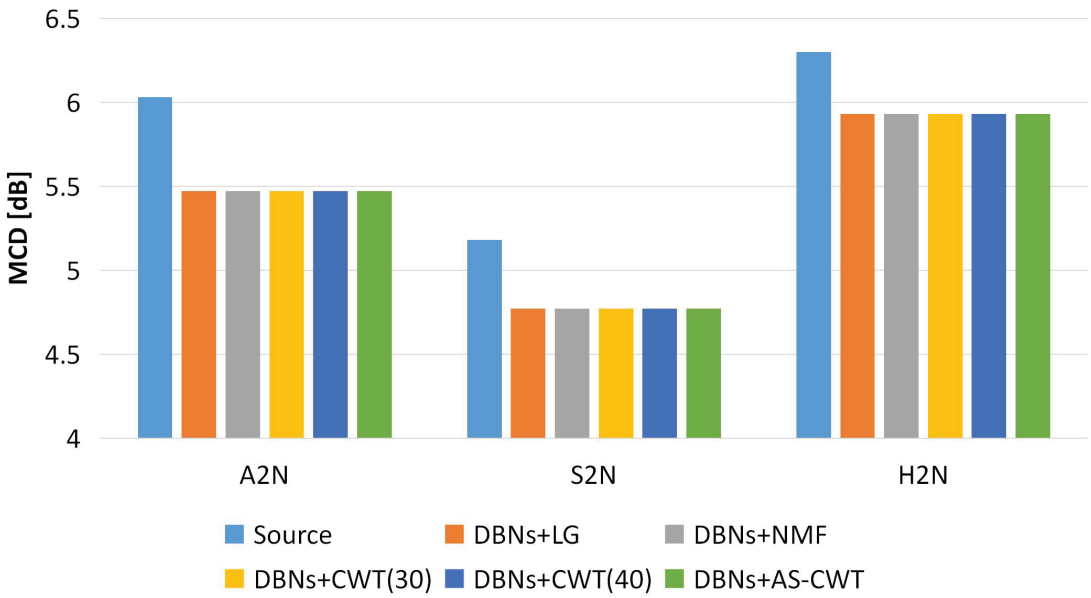

Figure 4.9: Mel-cepstral distortion evaluation of spectral features conversion from emotional to neutral voices using the DBNs model

Table 4.2 shows the MCD and F0-RMSE results from the neutral to emotional pairs. For spectral conversion, MCD values are deceased by the DBNs training model. However, when comparing Figure 4.11 with Figure 4.9, we can see that the effects of the conversion from neutral to angry and sad voices are not as significant

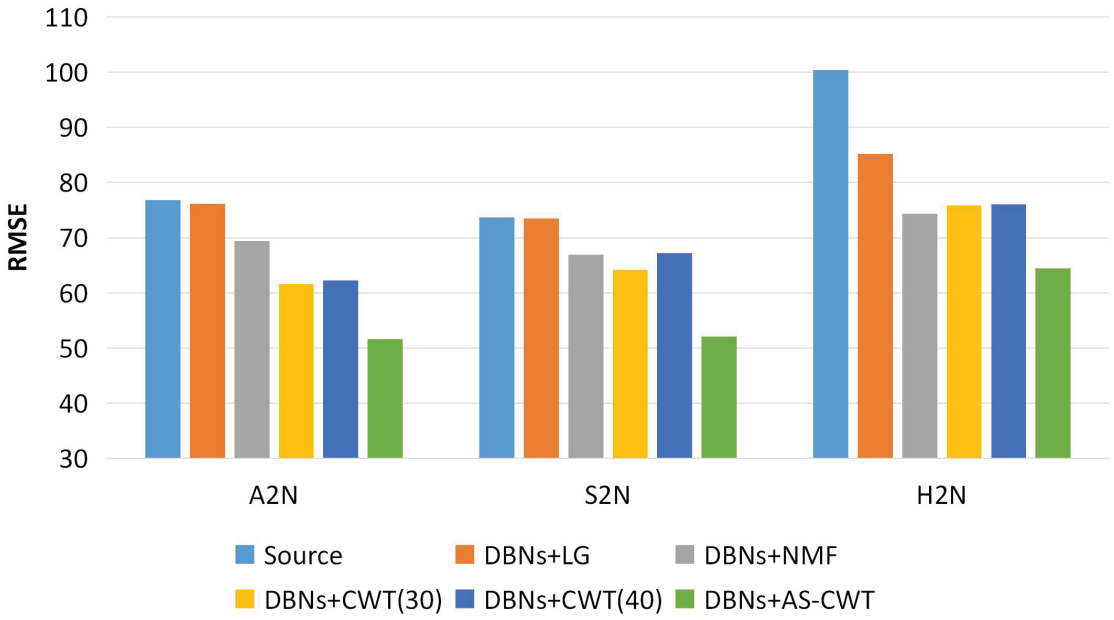

Figure 4.10: Root mean squared error evaluation of F0 features conversion from emotional to neutral voices using the AS-CWT method

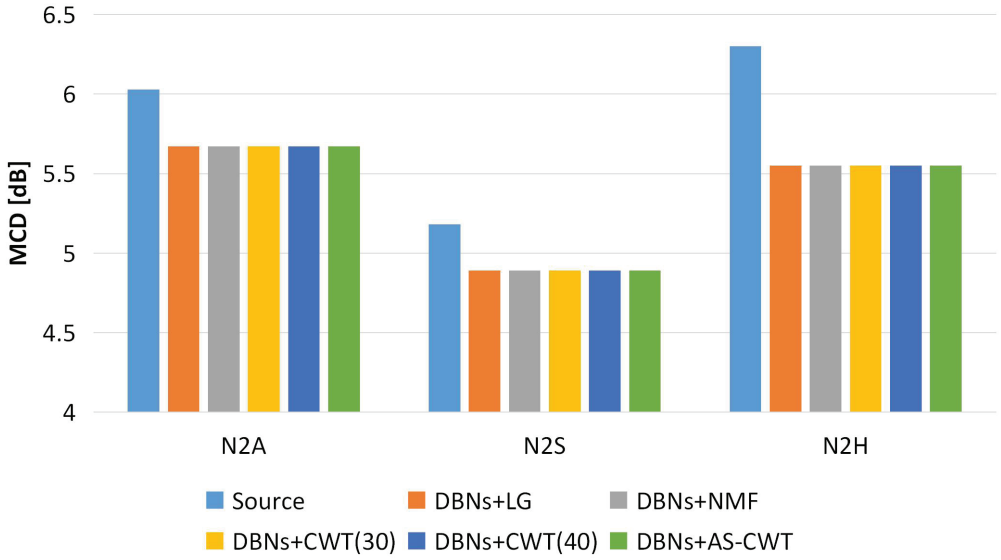

Figure 4.11: Mel-cepstral distortion evaluation of spectral features conversion from neutral to emotional voices using the DBNs model

Table 4.2: MCD and F0-RMSE results for different emotions. N2A, N2S and N2H represent the datasets neutral to angry voice, neutral to sad voice and neutral to happy voice, respectively.

	MCD [dB]			F0-RMSE [Hz]		
	N2A	N2S	N2H	N2A	N2S	N2H
Source	6.03	5.18	6.30	76.8	73.7	100.4
DBNs+LG	5.67	4.88	5.55	76.3	72.0	99.3
DBNs+NMF	5.67	4.88	5.54	70.4	62.3	75.2
DBNs+CWT(30)	5.68	4.88	5.55	39.5	40.1	64.5
DBNs+CWT(40)	5.68	4.88	5.55	41.6	40.5	67.5
DBNs+AS-CWT	5.68	4.88	5.55	41.5	39.4	63.2

as their inverse conversion, but the conversion from neutral to happy voice can generate better results than its inverse conversion. As shown in Figure 4.12, RMSE values decrease significantly when our proposed method is used compared with other methods. When comparing Figure 4.12 with Figure 4.10, we can see that the CWT method and AS-CWT method can effectively convert F0 features from neutral to emotional voices and obtain better results than the conversion from emotional to neutral voices. This is because the CWT and AS-CWT models decompose the one-dimensional F0 features into several scales containing further details of different temporal scales, which modeled and captured F0 features more appropriately than the other methods as well as alleviated estimation errors. As Figure 4.13 shows, the blue curve, red curve, and yellow curve represent the source, target, and converted F0, respectively. Here, we can see that after conversion using the proposed method, F0 becomes much similar to the target neutral voice.

4.3.3 Subjective Experiment

We conducted a subjective emotion evaluation via a mean opinion score test. The opinion score was set to a five-point scale (the more similar to the emotion of the sample voice the target speech sounded and the more different it is from the source speech, the larger the point given). Here, we test the emotional

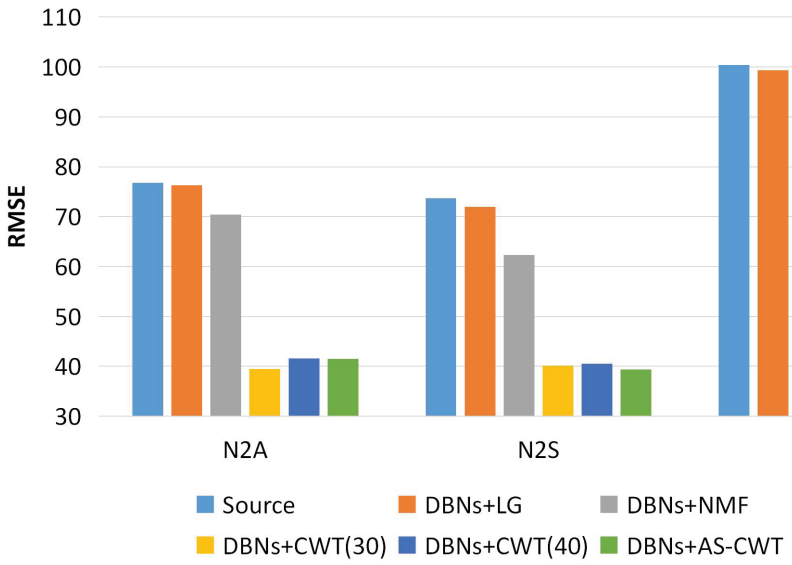

Figure 4.12: Root mean squared error evaluation of F0 features conversion from neutral to emotional voices using the AS-CWT method

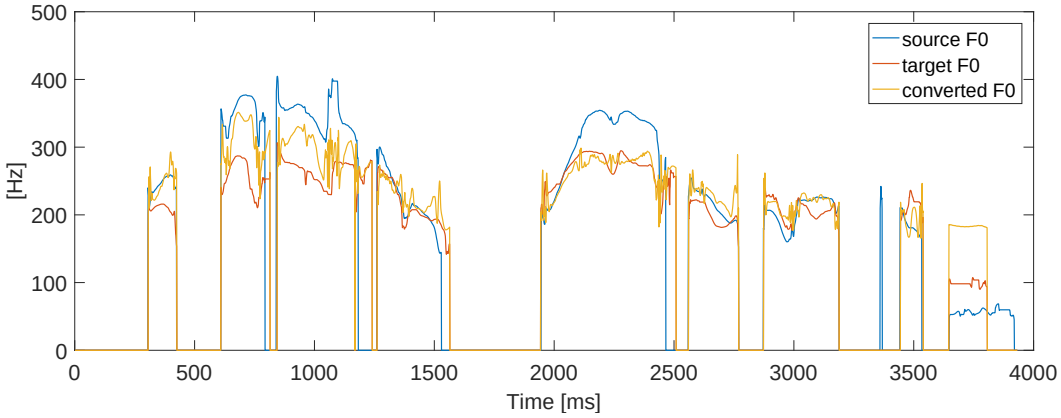

Figure 4.13: Examples of source F0, target F0 and converted F0

Table 4.3: MOS results for different emotions. A2N, S2N and H2N represent the datasets angry to neutral voice, sad to neutral voice and happy to neutral voice, respectively.

	A2N	S2N	H2N
DBNs+LG	2.03	2.63	2.76
DBNs+NMF	3.37	3.02	2.94
DBNs+AS-CWT	3.57	3.59	3.40

to neutral pairs (H2N, S2N, A2N). In each test, 50 utterances (10 for source speech, 10 for target speech, and 30 for converted speeches by the three methods) were selected, and 10 listeners were involved. Each subject listened to the source and target speeches. Then, the subject listened to the speech converted using the three methods and asked to give each conversion a point. As shown in Table 4.3 and Figure 4.14, the angry voice to neutral voice and the sad voice to neutral voice conversions can obtain better results than the happy voice to neutral voice conversion by the method DBNs+NMF and DBNs+AS-CWT, respectively. However, the conventional Gaussian method shows poor performance in the conversion of angery voice to neutral voice. The DBNs+AS-CWT (proposed method) also obtained a better score than the other two methods in each emotional VC.

4.4 Chapter Conclusions

In this paper, we proposed a method using DBNs to train the MCC features to construct the mapping relationship of the spectral envelopes, while using NNs to train the CWT-F0 features. Such features are conducted by the F0 features with adaptive scales for prosody conversion between the source and target speakers. A comparison between the proposed method and the conventional methods (logarithm Gaussian, NMF) shows that our proposed model can effectively change the acoustic and the prosody for the emotional voice at the same time.

When using this method, however, some problems remain. Specifically, the proposed model must extract the parallel speech data, which can limit the process

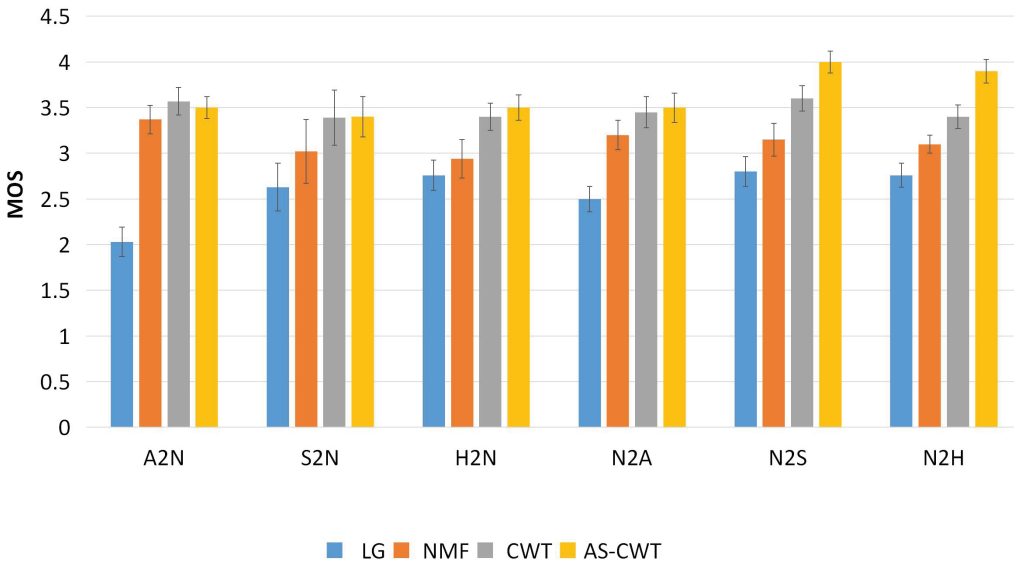

Figure 4.14: MOS evaluation of emotional voice conversion using the AS-CWT method

to a one-to-one conversion only. In future works, we will explore a many-to-many emotional VC method and use it in other applications, such as emotional voice recognition [64] or facial expression recognition [65].

Chapter 5

Emotional VC using Cross-Wavelet Transform

The related publications for this chapter is [66].

In this chapter, emotional voice conversion using XWT F0 features based on the Cross Wavelet Transform method is described. The method is viewed as combining CWT F0 features, which is discussed in Chapter 3, of different emotional voice for the emotional VC.

5.1 The Motivation and Related Work

In recent years, it has been shown that the CWT method can effectively model F0 in different temporal scales and significantly improve speech synthesis performance [17]. For this reason, Suni *et al.* [18] applied CWT to intonation modeling in a hidden Markov model (HMM) speech synthesis. Ming *et al.* [19] used CWT in F0 modeling within the non-negative matrix factorization (NMF) model for emotional VC. Chapter 4 systematically captured the F0 features of different temporal scales using adaptive scales CWT (AS-CWT), which transforms F0 features into high-dimensional CWT-F0 features containing more specifics. Conventional emotional VC tasks are focusing on converting one emotional voice to another fixed emotional voice. However, in the real world, the emotional voice is complex, with different prosody factors, and evolves over time. Thus, in order

to better represent the interactions of components in emotional voices and generate a greater variety of representations of F0 features. In this study, we want to go one step further and generate more emotional features using the cross-wavelet transform (XWT) [67]. The XWT method is a technique that characterizes the interaction between the wavelet transform of two individual time-series, which was first introduced by Hudgine *et al.* [68] and has been used in many areas such as economic analysis [69], geophysical time series [70] or electrocardiogram signals analysis [71].

5.2 Proposed Method

It is well known that prosody is influenced both at a supra-segmental level, by long-term dependencies, and at a segmental-level, by short-term dependencies. And, as has been proven in our recent studies [57, 59], the CWT can effectively model F0 in different temporal scales and significantly improve the system performance. So, in the proposed method, we adopt CWT to decompose the one-dimensional F0 features into high-dimensional CWT-F0 features, and then, use the cross-wavelet transform to combine the CWT-F0 features of different emotions. In the remaining part of this section, we will present the basic XWT theories, and illustrate how to use them in carrying out emotional VC.

5.2.1 Cross Wavelet Transform

In what follows, $L^2(\mathbb{R})$ denotes the set of square-integrable functions, which satisfy $\int_{-\infty}^{+\infty} |x(t)|^2 \, dt < \infty$. Given a time series $f_0(t) \in L^2(\mathbb{R})$, its continuous wavelet is, in respect to the mother wavelet $\psi \in L^2(\mathbb{R})$, defined as

$$\psi_{s,\tau} = s^{-1/2} \psi_0 \left(\frac{t-\tau}{s} \right) dt, s, \tau \in \mathbb{R}, s \neq 0, \tag{5.1}$$

where s is the scaling factor (which controls the width of the wavelet), and τ is the translating factor (which decides the location of the wavelet). $(s^{-1/2})$ is used for normalization, which can ensure unit variance of the wavelet and $\|\psi_{s,\tau}\|^2 = 1$.

Thus, the CWT of an input signal $f0(t)$ can be written as

$$\begin{aligned} W_{f_0;\psi}(s,\tau) &= \langle f_0(t), \psi_{s,\tau}(t)\rangle = \int_{-\infty}^{\infty} f_0(t)\,\psi_{s,\tau}(t)dt \\ &= s^{-1/2}\int_{-\infty}^{\infty} f_0(t)\,\psi_0\left(\frac{t-\tau}{s}\right)dt \end{aligned} \tag{5.2}$$

$$\psi_0(t) = \frac{2}{\sqrt{3}}\pi^{-1/4}\left(1-t^2\right)e^{-t^2/2}, \tag{5.3}$$

where $f_0(t)$ is the input signal and mother wavelet $\psi(t)$ used in our model is the Mexican hat mother wavelet [67]. The wavelet transform can give us information simultaneously on time-frequency space by mapping the original time series into the function of s and τ. Because both s and τ are real values and vary continuously, $W_{f_0;\psi}(s,\tau)$ is named a continuous wavelet transform. To be a mother wavelet of the CWT, $\psi(t)$ must fulfill admissibility conditions [72], which can be written as follows:

$$0 < C_\psi = \int_{-\infty}^{+\infty} \frac{|\hat{\psi}(f)|^2}{|f|} < +\infty, \tag{5.4}$$

where $\hat{\psi}(f)$ is the Fourier transform of the mother wavelet $\psi(t)$ and f is the Fourier frequency. It is clear that C_ψ is independent of f and determined only by the wavelet $\psi(t)$. This means that C_ψ is a constant for each given mother wavelet function. Thus, we can reconstruct $f_0(t)$ using its CWT, $W_{f_0;\psi}(s,\tau)$, by inverse function as follows:

$$f_0(t) = \frac{1}{C_\psi}\int_0^{+\infty}\left[\int_{-\infty}^{+\infty} W_{f_0;\psi}(s,\tau)\,\psi_{s,\tau}(t)d\tau\right]\frac{ds}{s^2}, s>0. \tag{5.5}$$

In this way, we can decompose F0 features $f_0(t)$ to the CWT-F0 features $W_{f_0;\psi}(s,\tau)$ and reconstruct them back to F0 features. Therefore, we have reason to believe that F0 and CWT-F0 are two different representations of the same mathematical entity. Figure 5.1 (2) shows the examples of CWT features decomposed from the F0 features of source and target emotional voices. In this figure, we show the parts of the scales that can represent the sentence, phrase, word, syllable, and

phone levels, respectively. Moreover, the energy of the examined time series is preserved by its CWT in the sense that

$$\|f_0(t)\|^2 = \frac{1}{C_\psi}\int_0^{+\infty}\left[\int_{-\infty}^{+\infty}|W_{f_0;\psi}(s,\tau)|^2\psi_{s,\tau}(t)d\tau\right]\frac{ds}{s^2}, s>0, \tag{5.6}$$

where $\|f_0(t)\|^2$ is defined as the energy of $f_0(t)$, and $|W_{f_0;\psi}(s,\tau)|^2$ is defined as a wavelet power spectrum that can interpret the degree of local variance of $f_0(t)$ scale by scale. The examples of wavelet power spectrum of the source and target CWT-F0 features are represented by the first pan and the second pan of Figure 5.1 (3).

Given two time series, $f_{0_x}(t)$ and $f_{0_y}(t)$, with their CWT features, W_x, and W_y, XWT is defined as $W_{xy;\psi}(s,\tau) = W_{x;\psi}(s,\tau)W_{y;\psi}*(s,\tau)$, where $*$ denotes complex conjugation. And their cross-wavelet power spectrum is accordingly written as

$$|W_{xy;\psi}(s,\tau)|^2 = |W_{x;\psi}(s,\tau)|^2|W_{y;\psi}*(s,\tau)|^2. \tag{5.7}$$

An example of the a cross-wavelet power spectrum is shown in the last pan of Figure 5.1 (3). The XWT of a two-time series depicts the local covariance between them at each time and frequency and shows the area in the time-frequency space, where the two signals exhibit high common power. Thus, we can do the cross-transform for different emotional voices to synthesize the new XWT-F0 features, and then reconstruct them to new synthesized F0 features. The new synthesized F0 features can be used in processing a new emtional voice or increasing the training data for emotional VC training.

5.2.2 Applying CWT and XWT in F0 features extraction

The STRAIGHT [60] is frequently used to extract features from a speech signal. Generally, the smoothing spectrum and instantaneous-frequency-based F0 are derived as excitation features for every 1 ms from the STRAIGHT. As shown in Figure 5.1, there are three steps for processing the XWT-F0 features.

1) To explore the perceptually-relevant information, F0 contour is transformed from the linear to logarithmic semitone scale, which is referred to as logF0. As shown in the first pan and second pan of Figure 5.1 (1), the F0 features extracted

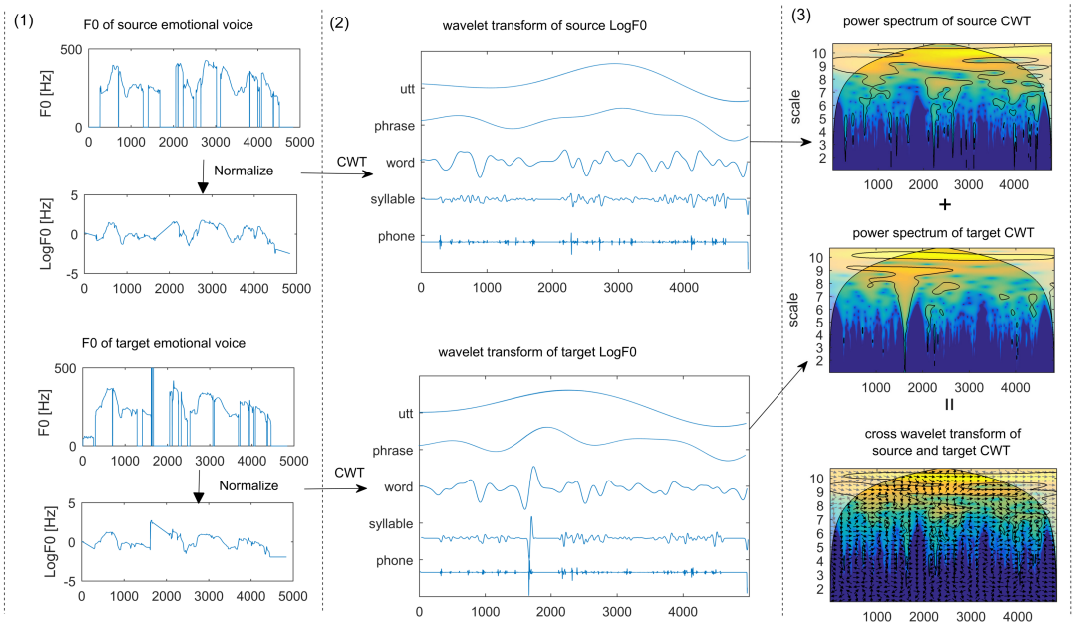

Figure 5.1: (1) Linear interpolation and log-normalized processing for source and target emotional voice. (2) The continuous wavelet transforms for normalized LogF0 contours, the top pan and bottom pan show five examples of the different level's decomposed CWT-F0 features of the source and target F0, respectively. (3) First and second pans show the continuous wavelet spectrums of the source and target CWT-F0 features and the bottom pan represents the cross-wavelet spectrums of the two CWT-F0 features. The relative phase relationship is shown as arrows in the cross-wavelet spectrum (with in-phase pointing right (→), anti-phase pointing left (←), and source emotional features leading target emotional features by 90° pointing straight down.

by STRAIGHT from the source and target emotional voice are discrete. As the wavelet method is sensitive to the gaps in the F0 contours, we must fill in the unvoiced parts in the logF0 via linear interpolation to reduce discontinuities in voice boundaries. Finally, we normalize the interpolated logF0 contour to zero mean and unit variance. Examples of interpolated LogF0 contours of source and target emotional voice are depicted in second and last pan of Figure 5.1 (1), respectively.

2) Then, we adopt CWT to decompose the continuous logF0 into 30 discrete scales, each separated by one-third of an octave. F0 is thus approximately represented by 30 separate components given by [57]

$$W_i(f_0)(t) = W_i(f_0)(2^{(i/3)+1}\tau_0, t)\left((i/3) + 2.5\right)^{-5/2}, \tag{5.8}$$

where i = 1,...,30 and τ_0 = 1 ms. In Figure 5.1 (2), we just show five examples of the decomposed CWT-F0 features which represent the utterance, phrase, word, syllable and phone levels, respectively.

3) Finally, we do the cross-wavelet transform for CWT-F0 features of the two emotional voices. First, we process the power spectrums of the source and target CWT-F0 features, and then use the cross-wavelet function to combine the two CWT-F0 features to cross-wavelet F0 features (XWT-F0). Figure 5.1 (3) shows the power spectrums of CWT-F0 features and the XWT-F0 feature. The XWT-F0 features can be used for generating the combinative emotional voice. Moreover, they can also be used for increasing the training data of the emotional VC tasks.

5.3 Experiments

In our experiments, we used a database of emotional Japanese speech constructed in a previous study [63]. In the database, there is a total of 50 different kinds of content sentences, and each sentence was read in four different emotions (angry, happy, sad and neutral) by the same speaker. Thus, there $50 \times 4 \times N$ waveforms in the database, where N is the number of speakers, the waveforms were sampled at 16kHz.

In the emotional VC experiment, we classified the three datasets into the following voice types: neutral to angry voices (N2A), neutral to sad voices (N2S), and neutral to happy voices (N2H). For each data set, 40 sentences were chosen as basic training data and 10 sentences were chosen for the VC evaluation.

Before applying the XWT-F0 features to the emotional VC, we need to calculate the correlates of different emotional voices. In other words, we do the cross-matching of each two emotional voices for the same content by the same speaker, which will result in six datasets $(N, A), (N, S), (N, H), (A, H), (A, S), (H, S)$. For each dataset, we do the cross-wavelet transform to synthesize the XWT-F0 features, and then reconstruct them to the F0 features that can be used for generating the complex emotional voice. Finally, we carry out a subjective experiment to decide the synthesized emotional voices that belong to which kind of emotion (neutral, angry, happy, sad and unknown). Here, due to some generated emotional voices may sound strange, we added the unknown option.

To evaluate the effectiveness of XWT-F0 features used as the increased training data for emotional VC, we carry out the experiments by adding the XWT-F0 features, which are selected by the subjective experiment, to the corresponding CWT-F0 features to increase the training data.

Moreover, to evaluate the relationship between different data size and the conversion methods described above, we conducted our experiment using a larger open database [73], which contains 100 sentences for each emotional voice. Among them, 80 sentences were chosen for training data and the existing 20 sentences were chosen for the testing data. In these experiments, the spectral features were converted using the same models as their CWT-F0 features training model.

5.3.1 Selecting Cross Wavelet Transform Features

To evaluate the synthesized emotional voice with the XWT-F0 features, we devised a subjective evaluation framework as shown in Figure 5.2. We first perform CWT on the neutral, angry, happy, and sad voice to decompose their F0 contours into CWT-F0 features. Then do the cross-wavelet transform for each pair of emotional CWT-F0 features, to process the XWT-F0 features. Then we

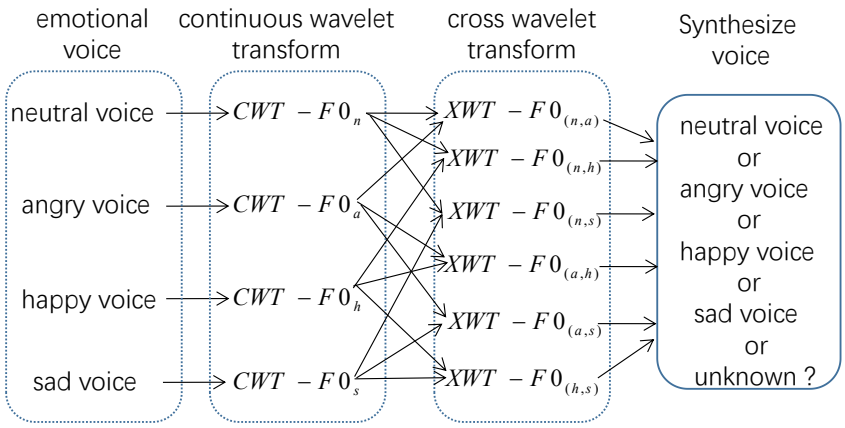

Figure 5.2: The listening experiment setup for evaluating emotional voice generated by different XWT-F0 features. The small letters in the lower right of CWT-F0 and XWT-F0 represent the emotion. For instance, (n) and (a) represent the neutral and angry. (n,a) represents the XWT-F0 features generated by CWT-F0 features of angry and neutral.

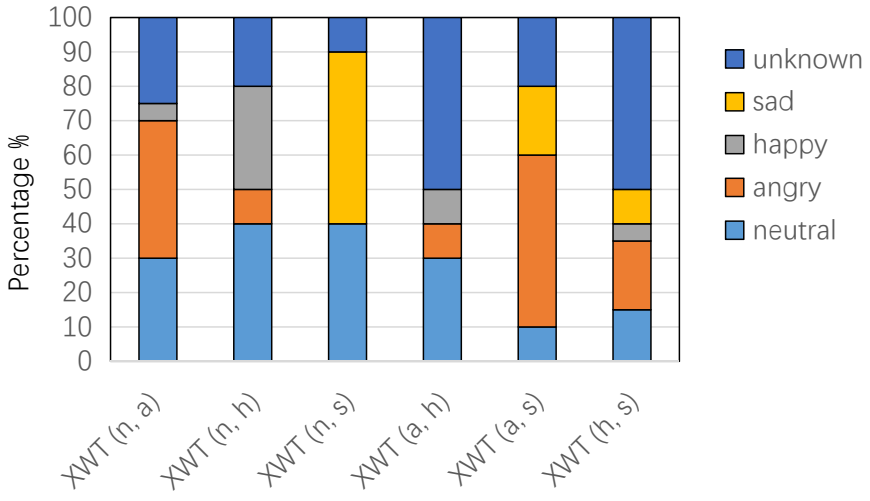

Figure 5.3: Similarity to different emotional voices (neutral, angry, happy and sad) of synthesized emotional voices with different XWT-F0 features.

reconstruct the XWT-F0 features to F0 features to synthesize the complex emotional voice. Then we ask 10 subjects to listen and choose the emotion that is close to the synthesized speeches.

Figure 5.3 shows the similarity to different emotions of the synthesized emotional voices using the different XWT-F0 features. For example, XWT (n, a) represents the synthesized emotional voice cross transformed with the neutral and angry CWT-f0 features. As the figure shows, the emotional voice generated from the neutral and sad voices XWT (n, s) got the most votes for the similarity to a sad voice. The voting results of XWT (n, a) and XWT (a, s) indicate that the angry voice can be generated by an angry and sad voice or angry and neutral voice. However, not all the generated emotional voices are good. For instance, the happy voice can be generated by XWT (n, h), but half of the subjects voted for the XWT (n, h) generated voice over the neutral voice. The XWT (h, s) and XWT (a, h) got the most votes for the unknown, which means there was strange prosody of the generated emotional voice.

In summary, depending on the emotional classification results of the emotional voice generated by the XWT-F0 (n, s), XWT-F0 (n, a), XWT-F0 (a, s) and XWT-F0 (n, h) features, we can use XWT-F0 (n, s) features as the CWT-F0 features of a sad voice, XWT-F0 (n, a) and XWT-F0 (a, s) as an angry voice, and the XWT-F0 (n, h) as a happy voice. Then, we can use both CWT-F0 features and their corresponding XWT-F0 features as the input training data, and the output is the CWT-F0 features.

5.3.2 Objective experiment

To evaluate F0 conversion, we used the root-mean-square error (RMSE) which is defined as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((F0_i^t) - (F0_i^c))^2}, \tag{5.9}$$

where $F0_i^t$ and $F0_i^c$ denote the target and the converted F0 features, respectively. A lower F0-RMSE value indicate a smaller distortion or predicting error. Unlike the RMSE evaluation function used in one previous study [19], which evaluated

Table 5.1: F0-RMSE results for different emotions. N2A, N2S and N2H represent the datasets from neutral to angry, sad and happy voice, respectively. (40) and (80) denote the number of training examples. "+" means using both the CWT-F0 features and selected XWT-F0 features for training model

	Source	LG	NNs	NNs+
N2A (40)	76.8	76.3	70.4	73.8
N2S (40)	73.7	72.0	62.3	60.7
N2H (40)	100.4	99.1	75.2	74.4
N2A (80)	77.3	73.5	66.2	65.4
N2S (80)	79.6	75.4	59.2	57.7
N2H (80)	98.4	93.2	69.2	66.4

the F0 conversion by calculating logarithmic scaled F0, we used the target F0 and converted F0 of a complete sentence including voiced and unvoiced (0 value), to calculate the RMSE values. Due to the alignment preprocessing in the parallel data, we used the unvoiced part of the source F0 in the converted F0. Considering that F0-RMSE evaluates complete sentences that contain both voiced and unvoiced F0 features instead of the voiced-logarithmic scaled F0, the RMSE values are expected to be high. For emotional voices, the unvoiced features also include some emotional information, such as the pause in a sentence, which can sometimes express nervousness or some other negative emotion. Therefore, we choose the F0 of complete sentences for evaluation as opposed to only voiced logarithmic-scaled F0.

As shown in Table 5.1, LG represents the conventional linear F0 conversion method using F0 features, directly. NNs represent the method using a neural networks training model with CWT-F0 features only. "+" represents using both the CWT-F0 features and the corresponding selected XWT-F0 features for the training model. As the results indicated, the conventional linear conversion LG has almost no effect in all the emotional VC datasets. Comparing the NNs and NNs+, we can see that adding the XWT-F0 features can improve the training effectiveness for the NNs training model.

5.3.3 Subjective experiment

We conducted a subjective emotion evaluation using a mean opinion score test to measure naturalness. The opinion score was set to a five-point scale ranging from totally unnatural (1) to completely natural (5). Here, we tested the neutral to emotional pairs (N2H, N2S, N2A). In each test, 50 utterances, converted by the five methods using both CWT-F0 and XWT-F0 features were selected, and 10 listeners were involved. The subjects listened to the speech that was converted using the five methods before being asked to assign a point value to each conversion. Figure 5.4 shows the results of the MOS test. The error bar shows the 95% confidence interval. As the figure shows, the conventional LG method shows poor performance in emotional VC pairs. The NNs+ method (CWT-F0 + XWT-F0) obtained the best score for every emotional VC.

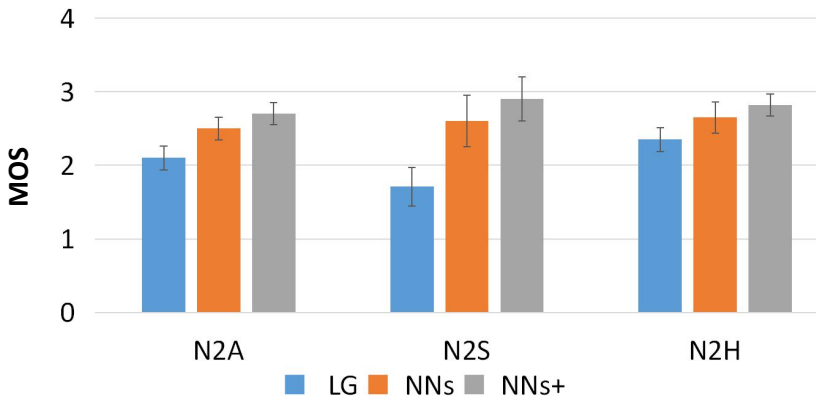

Figure 5.4: MOS evaluation of emotional voice conversion using the XWT method

Table 5.2: Results of emotion classification for recorded (original) voices and converted voice by different methods, N, A, S and H represents Neutral, Angry, Sad and Happy, respectively [%].

	(a) original voice				(b) LG				(c) NNs				(d) NNs+			
Tar./Percept	A	S	H	N	A	S	H	N	A	S	H	N	A	S	H	N
A	93	5	0	2	22	17	0	61	62	10	5	23	68	5	0	27
S	2	97	0	1	12	34	3	51	6	55	2	37	15	65	5	15
H	0	0	98	2	3	5	19	73	3	3	67	27	2	3	73	22

For the similarity evaluation, we carry out a subjective emotion classification test for neutral to emotional pairs (N2H, N2S, N2A) comparing different methods (LG, NNs, NNs+). For each test model, 30 utterances (10 for angry, 10 for sad, 10 for happy) are selected, and 10 listeners are involved. The listeners are asked to label a converted voice as Angry (A), Sad (S), Happy (H) or Neutral (N). As shown in Table 5.2 (a), when evaluating the original recorded emotional speech utterances, the classifier performed quite strongly; thus, the corpus is sufficient for recognizing emotion.

As shown in Table 5.2 (b), the conventional LG method shows poor performance in all emotional VC tasks. Comparing the results of Table 5.2 (c) and Table 5.2 (d), it is clear that NNs+ method obtains better classification results than the NNs model.

5.4 Chapter Conclusions

In this paper, we propose an effective neutral-to-emotional VC model, using the deep learning model. For the feature extraction and processing, we use CWT to systematically capture the F0 features of different temporal scales (CWT-F0 features), and then apply the cross-wavelet transforming for CWT-F0 features of different emotional voices to get the XWT-F0 features. A comparison between using the additional proposed XWT-F0 features together with CWT-F0 features for training can effectively improve the voice prosody conversion better than using CWT-F0 features only.

Chapter 6

Emotional VC using Dual Supervised GANs

The related publications for this chapter are [74, 75].

6.1 The Motivation and Related Work

To improve the emotional VC effect with a limited amount of training data, we propose dual supervised adversarial networks (dual-SANs), which combine generative adversarial networks (GANs) [22] with dual supervised learning [23] to train the Mel Cepstral Coefficients (MCC) and CWT-F0 features. The effectiveness of GANs is due to the fact that an adversarial loss forces the generated data to be indistinguishable from real data. This is particularly powerful for image generation tasks and has begun to be applied in speech synthesis [24] and VC [25]. Dual supervised learning is a technique for supervised learning tasks in dual forms, which was proposed by Xia *et al.* [23] in 2017. Prior to dual supervised learning, He *et .al* proposed the dual-learning mechanism [26] in 2016. The dual learning method uses one agent to represent the model for the primal task and another agent to represent the model for the dual task. They are then asked to teach each other through a reinforcement learning process.

In an emotional VC task, we need to convert the voice from one emotion to another, for example, from happy to neutral, sad to neutral, or their inverse conversions (these are dual tasks with labeled emotional voice data). Thus, owing

to dual supervised learning's ability to improve the training performances and GANs' ability to mitigate the over-smoothing problem caused in the low-level data space when converting the CWT-F0 features, we developed a novel dual-SANs model, which enables emotional VC to be trained from each of two sets of labeled emotional voices from two domains.

In dual-SANs architecture, the primal GAN learns to convert voices from emotion A to those of emotion B, while the dual GAN is trained for inverse conversion. The closed loop made by the primal and dual tasks allows the voice to be converted from emotion A to emotion B, and also obtain the inverse conversion from emotion B to emotion A. Moreover, the two dual tasks have intrinsic connections to each other because their source and target voice are in the same labeled emotional datasets. Using dual supervised learning can mutually benefit and regularize the training process; similar dual training models, such as disco-GAN [76], CycleGAN [77] and dual-GANs [78], have been successfully applied in image translation. Furthermore, the CycleGAN has also been used in parallel-data-free VC tasks [79] [80]. In contrast to these methods, which are most commonly applied to unsupervised tasks, our goal is to improve the emotion conversion effect with CWT-F0 features, and the proposed dual-SANs is effective in supervised tasks with insufficient emotional voice data.

6.2 Proposed Method

When dealing with training of emotional VC with the proposed dual-SANs, each set of labeled and paired 2D features (32×512) were sampled from domains of source emotional voice X and target emotional voice Y, respectively. As shown in Figure 6.1, our model contains two conversion functions $G_A : \boldsymbol{x}, \boldsymbol{z} \rightarrow \boldsymbol{y}$, and $G_B : \boldsymbol{y}, \hat{\boldsymbol{z}} \rightarrow \boldsymbol{x}$, where $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ are random noises, $\boldsymbol{z} \sim P_z(\boldsymbol{z})$ and $\hat{\boldsymbol{z}} \sim P_z(\hat{\boldsymbol{z}})$. The primal task of dual-SANs is to learn a conversion function $G_A : \boldsymbol{x}, \boldsymbol{z} \rightarrow \boldsymbol{y}$ that converts the source emotional voice $\boldsymbol{x} \in X$ to the target emotional voice $\boldsymbol{y} \in Y$, with the observed source emotional voice $\boldsymbol{x}$ and random noise vector $\boldsymbol{z}$, while the dual task is to train an inverse conversion function $G_B : \boldsymbol{y}, \hat{\boldsymbol{z}} \rightarrow \boldsymbol{x}$. To realize this, we apply two GANs. The primal GAN (GAN 1) learns the conversion function G_A, D_Y that discriminates between converted outputs ($\hat{\boldsymbol{Y}}$) of G_A and

real members of domain Y. Analogously, the dual GAN (GAN 2) learns the conversion function, G_B, and a discriminator, D_X. We apply adversarial losses to both conversion functions. For the conversion function $G_A : X \to Y$ and its discriminator D_Y, we express the objective as:

$$\begin{aligned} & L_{G_A}(G_A, D_Y, X, Y) \\ & = E_{\boldsymbol{y},\boldsymbol{x}\sim P_{data}(\boldsymbol{y},\boldsymbol{x})}[\log D_Y(\boldsymbol{y}, \boldsymbol{x})] \\ & \quad + E_{\boldsymbol{x}\sim P_{data}(\boldsymbol{x}),\boldsymbol{z}\sim P_z(\boldsymbol{z})}[\log(1 - D_Y(G_A(\boldsymbol{x}, \boldsymbol{z}), \boldsymbol{x}))], \end{aligned} \tag{6.1}$$

where $\boldsymbol{x}$ is the observed source emotional voice, $\boldsymbol{y}$ is the target emotional voice and $\boldsymbol{z}$ is a random noise vector. The goal of emotional VC is to learn a converted emotional voice distribution $P_{G_A}(\boldsymbol{x}, \boldsymbol{z})$ that matches the target emotional voice distribution $P_{data}(\boldsymbol{y})$. Eq. (6.1) enables D_Y to find the binary classifier that provides the best possible discrimination between true and converted voice and simultaneously enables the function G_A to fit the $P_{data}(\boldsymbol{y})$.

According to previous work, without $\boldsymbol{z}$, the net could still learn a mapping from $\boldsymbol{x}$ to $\boldsymbol{y}$, but would produce deterministic outputs. However, Isola *et al.* [81] reported that adding a Gaussian noise as an input to G, as done in [82], was not effective. Thus, our GAN model used in dual-SANs is similar to Pix2Pix [81], which did not explicitly provide the noise vectors $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$. We provided the noise only in the form of dropout, and applied it to several layers of our generators at both training and test phases. Furthermore, previous works on cGAN found it beneficial to replace L2 distance with L1, since it causes less blurring and it tends to generalize better [81, 83]. Therefore, in our model, we also mix the GAN objective with a traditional L1 distance loss as:

$$L_{1A}(G_A) = E_{(\boldsymbol{x},\boldsymbol{y})\sim P_{data}(\boldsymbol{x},\boldsymbol{y}),\boldsymbol{z}\sim P_z(\boldsymbol{z})}[\|\boldsymbol{y} - G_A(\boldsymbol{x}, \boldsymbol{z})\|_1],$$

where $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ represent the observed source emotional voice, target emotional voice, and a random noise vector, respectively. Our final objective is maximized and minimized with respect to D_Y and G_A, respectively.

$$G_A^* = \arg \max_{D_Y} \min_{G_A} L_{G_A}(G_A, D_Y, X, Y) + L_{1A}(G_A). \tag{6.2}$$

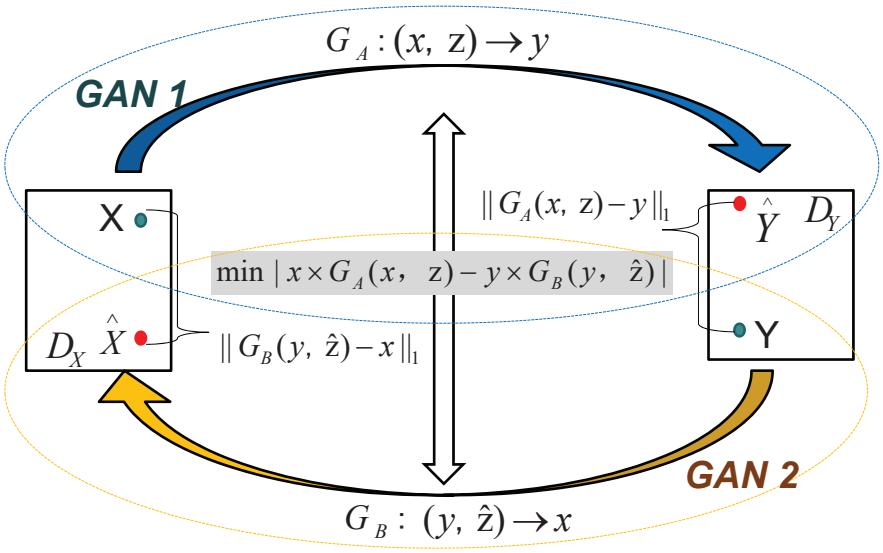

Figure 6.1: Illustration of calculating the loss of dual-SANs, $\|G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y}\|_1$ and $\|G_B(\boldsymbol{y}, \hat{\boldsymbol{z}}) - \boldsymbol{x}\|_1$ represent the distance losses for conversion function from data $\boldsymbol{x}$ to $\boldsymbol{y}$ and the dual task (data $\boldsymbol{y}$ to $\boldsymbol{x}$), respectively. $|\boldsymbol{x} \times G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y} \times G_B(\boldsymbol{y}, \hat{\boldsymbol{z}})|$ represents the dual loss.

We use the similar distance loss and adversarial loss for the inverse conversion function $G_B : Y \rightarrow X$ and its discriminator D_X as well:

$$G_B^* = \arg \max_{D_X} \min_{G_B} L_{G_B}(G_B, D_X, Y, X) + L_{1B}(G_B), \tag{6.3}$$

where

$$\begin{aligned} & L_{G_B}(G_B, D_X, Y, X) \\ & = E_{(\boldsymbol{x},\boldsymbol{y})\sim P_{data}(\boldsymbol{x},\boldsymbol{y})}[\log D_X(\boldsymbol{x}, \boldsymbol{y})] \\ & \quad + E_{\boldsymbol{y}\sim P_{data}(\boldsymbol{y}),\hat{\boldsymbol{z}}\sim P_z(\hat{\boldsymbol{z}})}[\log(1 - D_X(G_B(\boldsymbol{y}, \hat{\boldsymbol{z}}), \boldsymbol{y})], \end{aligned} \tag{6.4}$$

$$L_{1B}(G_B) = E_{(\boldsymbol{x},\boldsymbol{y})\sim P_{data}(\boldsymbol{x},\boldsymbol{y}),\hat{\boldsymbol{z}}\sim P_z(\hat{\boldsymbol{z}})}[\|\boldsymbol{x} - G_B(\boldsymbol{y}, \hat{\boldsymbol{z}})\|_1]. \tag{6.5}$$

In L_{G_B} and $L_{1B}(G_B)$, $\boldsymbol{y}$ is the input source emotional voice (used as the target emotional voice in the original conversion function), $\boldsymbol{x}$ is the target emotional voice (used as the source emotional voice in the original conversion function) and $\hat{\boldsymbol{z}}$ is the random noise vector.

In supervised dual tasks, the primal task's source emotional voice X is the dual task's target, and the target emotional voice Y is the task's source. For any

$\boldsymbol{x} \in X$, $\boldsymbol{y} \in Y$, the product of primal and dual tasks should satisfy the following equality: $|\boldsymbol{xy}| = |\boldsymbol{yx}|$. Then, we can minimize the two models' loss functions, $G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y}$ and $G_B(\boldsymbol{y}, \hat{\boldsymbol{z}}) - \boldsymbol{x}$, by using the dual loss $|\boldsymbol{x} \times G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y} \times G_B(\boldsymbol{y}, \hat{\boldsymbol{z}})|$ to guarantee their intrinsic connections. Unlike dual supervised learning [23] applied in machine translation, which uses the duality of joint probability to regularize the training process, in emotional VC, owing to the true value conversion and the duality of the paired emotional voice dataset, we used the true values of $\boldsymbol{x}$, $\boldsymbol{y}$ and their converted value $G_A(\boldsymbol{x}, \boldsymbol{z})$, $G_B(\boldsymbol{y}, \hat{\boldsymbol{z}})$ to keep the training process regularized. The purpose of the dual supervised task is to learn the two models, G_A and G_B, by minimizing their loss functions subject to the dual constraint:

$$\begin{aligned} & L_{dual}(G_A, G_B) \\ & = E_{(\boldsymbol{x},\boldsymbol{y})\sim P_{data}(\boldsymbol{x},\boldsymbol{y}),\ \boldsymbol{z}\sim P_z(\boldsymbol{z}),\ \hat{\boldsymbol{z}}\sim P_z(\hat{\boldsymbol{z}})}[\|\boldsymbol{x} \times G_A(\boldsymbol{x}, \boldsymbol{z}) \\ & \quad - \boldsymbol{y} \times G_B(\boldsymbol{y}, \hat{\boldsymbol{z}})\|_1]. \end{aligned} \tag{6.6}$$

Thus, the intrinsic connection between G_A and G_B is explicitly strengthened, which is supposed to push the learning process towards the right direction and keep a balance when learning the two dual tasks. Combining the dual loss in dual supervised learning, adversarial loss in GANs and the traditional distance loss, our full objective is:

$$\begin{aligned} & L(G_A, G_B, D_X, D_Y) \\ & = L_{G_A}(G_A, D_Y, X, Y) + \lambda_1 L_{1A}(G_A) \\ & \quad + L_{G_B}(G_B, D_X, Y, X) + \lambda_2 L_{1B}(G_B) \\ & \quad + \lambda_3 L_{dual}(G_A, G_B), \end{aligned} \tag{6.7}$$

where λ_1, λ_2 and λ_3 control the relative importance of these objectives. We aim to solve:

$$G_A^*, G_B^* = \arg \max_{D_Y, D_X} \min_{G_A, G_B} L(G_A, G_B, D_Y, D_X). \tag{6.8}$$

The details of how to train dual-SANs are indicated in Algorithm 1.

As shown in Algorithm 1, we train the discriminators n steps, and then one step for the generators. We employ batch training with arbitrary batch size m,

Algorithm 1 Training procedure of dual-SANs

Require: sequence features (32×512) dataset X and Y processed from source and target emotional voice, generators G_A and G_B with conversion parameters $\boldsymbol{\theta_{G_A}}$ and $\boldsymbol{\theta_{G_B}}$, discriminator D_Y and D_X with parameters $\boldsymbol{w_Y}$ and $\boldsymbol{w_X}$, respectively, arbitrary batch size m, and the epochs n.

1: Initialize the parameters $\boldsymbol{\theta_{G_A}}$, $\boldsymbol{\theta_{G_B}}$, $\boldsymbol{w_Y}$ and $\boldsymbol{w_X}$, randomly.
2: **repeat**
3: sample 2D features $\boldsymbol{x_k} \subseteq X$, $\boldsymbol{y_k} \subseteq Y$, $k \in \{1, ..., m\}$
4: update $\boldsymbol{\theta_{G_A}}$ to minimize $\frac{1}{m}\sum_{k=1}^{m} L_{1A}(\boldsymbol{x_k}, \boldsymbol{y_k})$
5: update $\boldsymbol{\theta_{G_B}}$ to minimize $\frac{1}{m}\sum_{k=1}^{m} L_{1B}(\boldsymbol{y_k}, \boldsymbol{x_k})$
6: **for**$(i = 1; i < n + 1; i = i + 1)$ **do**
7: update $\boldsymbol{\theta_{G_A}}$, $\boldsymbol{w_Y}$ to maximize $\frac{1}{m}\sum_{k=1}^{m} L_{G_A}(\boldsymbol{x_k}, \boldsymbol{y_k})$
8: update $\boldsymbol{\theta_{G_B}}$, $\boldsymbol{w_X}$ to maximize $\frac{1}{m}\sum_{k=1}^{m} L_{G_B}(\boldsymbol{y_k}, \boldsymbol{x_k})$
9: **end for**
10: sample 2D features $\boldsymbol{x_k} \subseteq X$, $\boldsymbol{y_k} \subseteq Y$, $k \in \{1, ..., m\}$
11: update $\boldsymbol{\theta_{G_A}}$, $\boldsymbol{\theta_{G_B}}$ to minimize $\frac{1}{m}\sum_{k=1}^{m} L_{dual}(\boldsymbol{y_k}, \boldsymbol{x_k})$
12: **until** convergence

which depends on the arbitrary length of the input sentences. Throughout the training process, conversion functions G_A and G_B are optimized to learn the converted emotional voice, which cannot be distinguished from the target emotional voice by corresponding discriminators D_Y and D_X, as well as to minimize the dual loss $\|\boldsymbol{x} \times G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y} \times G_B(\boldsymbol{y}, \hat{\boldsymbol{z}}))\|$. Due to the limited training data, we assume that an appropriate regularization can benefit high complex GANs models considerably, and, in fact, the duality-oriented regularization L_{dual} indeed plays this role and consequently gives rise to higher relative improvement. In Section 5.3, we also describe an experiment comparing the full dual-SANs model against the GANs trained without dual loss ($|\boldsymbol{x} \times G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y} \times G_B(\boldsymbol{y}, \hat{\boldsymbol{z}})|$).

6.3 Experiments

We used a database of emotional Japanese speech constructed in a previous study [63]. The waveforms used were sampled at 16kHz. In the database, 50 sentences from the ATR Japanese phonetically balanced text set were used in the experiments. These 50 sentences are designed to include a minimum phone set of Japanese. All the texts were read by a professional narrator with neutral, angry, happy and sad voices. Input and output data had the same speaker, but expressing different emotions. We set the six datasets into the following: happy to neutral voice, angry to neutral voice, and sad to neutral voice, as well as their inverse conversion from neutral voice to each emotional voice. For each dataset, 35 sentences were chosen for training data and the other 15 sentences were chosen for the validation set. We conducted evaluations with a five-fold cross-validation scheme and the performance was measured using average conversion evaluation.

To evaluate the proposed method in converting different emotional datasets, we compared the results with several state-of-the-art methods listed below.

- **DBNs+LG-1 [11]:** This system proposed by Nakashika *et al.* converts spectral features using DBNs and converts the F0 features through the logarithm Gaussian (LG) method.

- **DBNs+LG-2:** This system converts spectral features by DBNs and converts the F0 features using the LG method for each **AS-CWT** scale separately.

- **DBNs+NNs [59]:** This is a previously published method that uses the DBNs to convert spectral features, while using the NNs to convert the CWT-F0 features. (In order to make a fairer comparison to other GANs models, we used the same normalization method as GANs models.)

- **GANs-1:** This method uses GANs to train MCC and basic CWT-F0 features for conversion functions without dual loss. (In order to make a fairer comparison to proposed dual-SANs, the distance loss of the GANs was set to $\|G_A(\boldsymbol{x}, \boldsymbol{z}) - \boldsymbol{y}\|$, which was used in dual-SANs.)

- **GANs-2:** This method uses the same GANs model as GANs-1 to train MCC features and the proposed AS-CWT F0 features for conversion functions.

- **CycleGAN:** In this model, we use the cycle consistency loss of CycleGAN [79] to replace the dual constraint in dual-SANs, as the regularization for the GANs training of the dual task, which has been proposed for non-parallel voice conversion tasks. (In order to make a fairer comparison to the proposed dual-SANs, the distance loss was also used in the comparative CycleGAN.)

- **dual-SANs-1:** This method uses our proposed daul-SANs to train MCC features and basic CWT-F0 features for conversion functions.

- **dual-SANs-2:** This is our proposed method that uses dual-SANs to train MCC features and the proposed AS-CWT F0 features for conversion functions.

To evaluate the relationship between different data sizes and the conversion methods described above, we conducted our experiment using the angry to the neutral dataset with a larger open database [73], which contains 100 sentences

for angry voice and neutral voice, respectively. Among them, 80 sentences were chosen for training data and the other 20 sentences were chosen for the validation set. Moreover, the evaluation was conducted with five-fold cross-validation and the last performance was measured by average conversion evaluation.

6.3.1 Training Procedure

Table 6.1 details the network architectures of the generator G_A, G_B, and discriminator D_Y, D_X. The symbols $\downarrow$ and $\uparrow$ indicate down sampling and up sampling, respectively. To upscale and downscale, we respectively used convolutions and backward convolutions with stride 2.

In the generator networks (G_A and G_B), similar to Johnson *et al.* [84], we use batch normalization (BNorm) [85] and convolutional layers are followed by Dropout [86] with a probability of 0.5 or ReLU nonlinearities [87] with the exception of the output layer. Input and output are set as 32×512 2D features. Each generator network contains two stride-2 convolutions to down sample the input, followed by two residual blocks [88], and two fractionally strided convolutions with stride $\frac{1}{2}$ to up sample.

For discriminator networks (D_X, D_Y), we use a convolutional PatchGAN classifier [89]. The patch size at which the discriminator operates is fixed at 30×30 and two stride-2 convolutions and a fully connected layer make up the discriminators D_X and D_Y. After the last layer, we apply a sigmoid activation function to obtain a probability for sample classification.

During preprocessing, we normalized the source and target, MCC and CWT-F0, features to zero-mean and unit-variance for each dimension using their respective training sets, then transform them to 32×512 2D features. When training the G_A, G_B, D_Y and D_X, we use the Adam optimizer [90] with an arbitrary batch size m depending on the length of the input sentences. For all experiments, we set $\lambda_1 = 5$, $\lambda_2 = 5$ and $\lambda_3 = 15$ in Eq. (6.7). The learning rate was set to 0.0002 for the generator G_A, 0.0002 for the generator G_B, and 0.0001 for D_Y and D_X, respectively. The momentum term was set to 0.5.

To clarify the characteristics of our framework with dual-SANs, as described above, we implemented two kinds of DBNs frameworks, the GANs (without dual

Table 6.1: Details of network architectures of G_A, G_B, D_x, D_y.

G_A (Input: 32×512 2D features processed by source features X, noise $\boldsymbol{z}$ for dropout layer, Output: 32×512 converted 2D features)
3×3 128 conv.↓, BNorm, ReLU
3×3 64 conv.↓, ReLU
residual blocks $\begin{bmatrix} 3 \times 3 & 64 & \text{conv.} & \text{ReLU} \\ 3 \times 3 & 64 & \text{conv.} & \text{ReLU} \end{bmatrix} \times 2$
3×3 64 conv.↑, Dropout, ReLU
3×3 128 conv.↑, Dropout, BNorm, ReLU
G_B (Input: 32×512 2D features processed by target features Y, noise $\hat{\boldsymbol{z}}$ for dropout layer, Output: 32×512 converted 2D features
3×3 128 conv.↓, BNorm, ReLU
3×3 64 conv.↓, ReLU
residual blocks $\begin{bmatrix} 3 \times 3 & 64 & \text{conv.} & \text{ReLU} \\ 3 \times 3 & 64 & \text{conv.} & \text{ReLU} \end{bmatrix} \times 2$
3×3 64 conv.↑, Dropout, ReLU
3×3 128 conv.↑, Dropout, BNorm, ReLU
D_Y (Input: 32×512 2D features, Output: 1 Probability)
3×3 64 conv.↓, ReLU
3×3 128 conv.↓, BNorm, ReLU
512 fully connected, BNorm, ReLU
1 fully connected, sigmoid
D_X (Input: 32×512 2D features, Output: 1 Probability)
3×3 64 conv.↓, ReLU
3×3 128 conv.↓, BNorm, ReLU
512 fully connected, BNorm, ReLU
1 fully connected, sigmoid

loss) model and CycleGAN for comparison. One DBNs model is a state-of-the art VC framework that uses DBNs and LG to convert MCC and F0 features [11]. The DBNs have four hidden layers, and the numbers of units in input, hidden, and output layers are [40, 120, 30, 30, 120, 40]. Another DBNs model is taken from our previously published work that uses the same DBNs model to train MCC features while using NNs to convert the CWT-F0 features. The units of the NNs are set to [40, 60, 60, 40]. In these DBNs models, for making a fairer comparison, we used the same normalization and de-normalization procedures as the proposed dual-SANs method. The compared GANs model without supervised dual learning has the same architecture with the generative models (G_A and G_B) used in dual-SANs, which were trained not only with an adversarial loss but also with a distance loss. The setting of the compared CycleGAN model is similar to the model used in the VC task [79].

Table 6.2: MCD and F0-RMSE results for the conversion of emotional voice to neutral voice.

	MCD [dB]			F0-RMSE [Hz]		
	A2N	S2N	H2N	A2N	S2N	H2N
Source	6.03	5.18	6.30	58.5	59.9	83.4
DBNs+LG-1	5.56	4.85	5.95	48.1	55.3	65.1
DBNs+LG-2	5.56	4.85	5.95	46.5	57.2	64.3
DBNs+NNs	5.56	4.85	5.95	28.4	33.1	38.2
GANs-1	4.55	4.67	5.21	26.3	28.2	35.1
GANs-2	4.55	4.67	5.21	22.2	25.3	30.2
CycleGAN	4.12	4.68	4.75	21.5	25.6	27.4
dual-SANs-1	**3.41**	**4.51**	**4.35**	17.3	22.3	23.5
dual-SANs-2	**3.41**	**4.51**	**4.35**	**15.2**	**18.4**	**19.3**

6.3.2 Objective Experiment

For objective evaluation, we used Mel-cepstral distortion (MCD) to measure how close the converted MCC was to the target MCC in the Mel-cepstral space, and root means square error (RMSE) to evaluate the conversion error between

Table 6.3: MCD and F0-RMSE results for the conversion of neutral voice to emotional voice.

	MCD [dB]			F0-RMSE [Hz]		
	N2A	N2S	N2H	N2A	N2S	N2H
Source	6.03	5.18	6.30	58.5	59.9	83.4
DBNs+LG-1	5.76	4.94	5.59	48.8	54.7	65.3
DBNs+LG-2	5.76	4.94	5.59	46.7	57.9	64.2
DBNs+NNs	5.76	4.94	5.59	32.6	35.8	42.9
GANs-1	4.79	4.66	4.52	28.3	25.3	34.7
GANs-2	4.79	4.66	4.52	30.2	22.4	29.2
CycleGAN	4.52	**4.41**	4.45	22.2	24.5	27.8
dual-SANs-1	**3.84**	4.48	**3.59**	16.3	20.9	21.4
dual-SANs-2	**3.84**	4.48	**3.59**	**13.7**	**16.3**	**17.2**

Table 6.4: MCD and F0-RMSE results for the conversion of angry voice to neutral voice with different data size.

Data size	MCD [dB]			F0-RMSE [Hz]			
(training)	DBN	GANs	dual-SANs	LG	NNs	GANs	dual-SANs
80	4.99	4.12	3.37	44.2	25.6	20.7	14.9
35	5.56	4.55	3.41	46.5	28.4	22.2	15.2

target F0 and converted F0. MCD is defined below.

$$MCD = (10/\ln 10)\sqrt{2\sum_{i=1}^{24}(mc_i^t - mc_i^c)^2}. \tag{6.9}$$

In Eq. (6.9), mc_i^t and mc_i^c represent the target and the converted mel-cepstral, respectively.RMSE of interpolated F0s between the converted and corresponding target emotional voice is calculated as

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((F0_i^t) - (F0_i^c))^2}, \tag{6.10}$$

where $F0_i^t$ and $F0_i^c$ denote the target and the converted interpolated F0 features, respectively. A lower MCD and F0-RMSE value indicate a smaller distortion or predicting error.

The average MCD and F0-RMSE results from emotional to neutral pairs are shown in Table 6.2. Furthermore, the results of their inverse conversion, from a neutral to emotional voice are shown in Table 6.3. In the two tables, A2N, S2N, and H2N denote angry-to-neutral conversion, sad-to-neutral conversion, and happy-to-neutral conversion, respectively. Their inverse conversion, N2A, N2S, and N2H represent neutral-to-angry conversion, neutral-to-sad conversion, and neutral-to-happy conversion, respectively.

As shown in Table 6.2 and Table 6.3, comparing the MCD values of the source voice and converted voice by using deep learning models (DBNs, GANs, CycleGAN, and dual-SANs), all deep learning models decreased the value of MCD, showing the significant ability in the spectral conversion. While comparing the DBNs+LG and DBNs+NNs, MCD value is the same, proving that the conversion of F0 by different models does not have an impact on the spectral feature conversion. For reference, we also performed the experiments using GANs without dual supervised learning (GANs-1 and GANs-2) and GANs with cycle-consistency loss (CycleGAN). The MCD results are shown in Table 6.2 and Table 6.3, GANs-1 and GANs-2 obtained slightly better results than the DBNs models, but the results were worse than those for the CycleGAN and dual-SANs. Comparing dual-SANs with CycleGAN, the MCD decreases or increases slightly for S2N and N2S, but

decreases significantly for A2N, H2N, and their inverse conversion. This is because CycleGAN is effective for transferring sophisticated local texture appearance between image domains, but it has difficulty with objects that have both related appearance and shape changes. For the angry and happy voice conversion, their spectrum shape is more different from the neutral voice than the sad voice. This shows that although cycle-consistency is effective for the training of GANs when converting the sad voice to neutral voice and their inverse conversion, the effect of cycle-consistency is no more than the duality constraint proposed in our dual-SANs when dealing with the parallel data for the emotional VC task, in which the emotional features' shapes change a lot.

The average F0-RMSE results of the emotional to neutral voice are shown in the right part of Table 6.2. The conventional linear conversion, LG, can only affect the conversion of happy to neutral voices, but only slightly affects the other conversions. Comparing the method LG-1 and LG-2, we can see that using the LG method for each CWT scaled feature separately can only decrease the RMSE value slightly. This indicates that the linear LG method is insufficient to convert the CWT-F0 features. Comparing GANs-1 with GANs-2, and dual-SANs-1 with dual-SANs-2, we found that using the CWT-F0 features with the arbitrary scales can increase the dimensions of the F0 features that were well trained using deep learning models. Comparing the RMSE results of NNs, GANs, CycleGAN and dual-SANs, these methods can all affect the conversion of all emotional voice dat asets. In addition, the GANs, CycleGAN, and proposed dual-SANs methods can obtain the better results compared with NNs in F0 conversion.

As shown in Table 6.3, the proposed dual-SANs method can also obtain good results in spectral and F0 conversion when dealing with the inverse conversion from neutral to emotional voice, as a whole. Moreover, comparing Table 6.2 and Table 6.3, the results of emotional voice to neutral voice and their inverse conversion, we can see that using adversarial models (GANs, CycleGAN, dual-SANs) with AS-CWT F0 features can effectively convert F0 features from neutral to emotional voices and obtain better results than the conversion from emotional to neutral voices. This is likely due to the fact that AS-CWT models decompose the one-dimensional F0 features into several scales containing further details of differ-

ent temporal scales, which modeled and captured F0 features more appropriately than original F0 features, as well as alleviating estimation errors.

Table 6.4 shows the results of MCD and RMSE when converting an angry voice to a neutral voice using dual-SANs, GANs and DBNs models trained with different size of training data. 35 and 80 represent using 35 and 80 sentences for training, respectively. The MCD and RMSE of GANs and DBNs increase significantly when using the smaller training data size, which means the conversion effect getting worse. On the other hand, the MCD and RMSE of dual-SANs increased only a little when using a smaller data size, which shows the strength of dual-SANs when using limited training data.

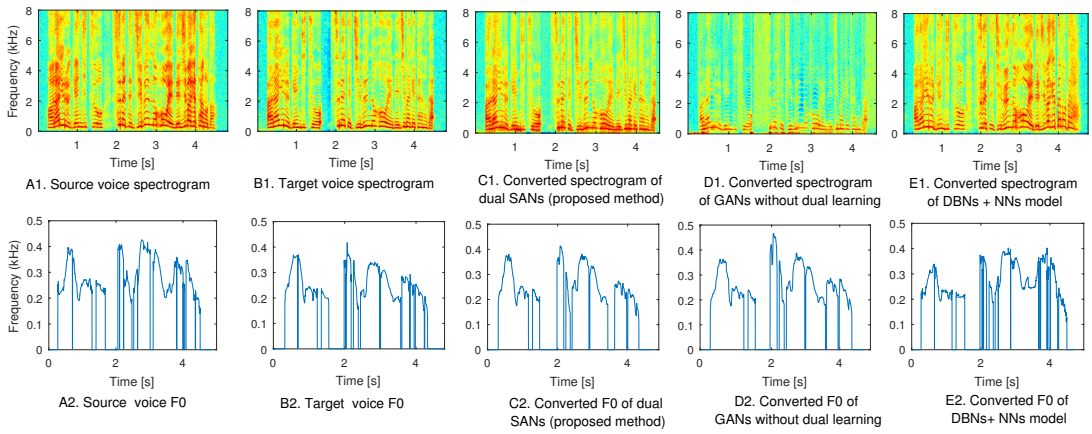

Figure 6.2: Spectrograms and F0 of the source, target and converted emotional voices by different methods.

Figure 6.2 shows examples of source, target, and converted voice spectrograms and F0 features by dual-SANs, GANs without dual supervised learning and DBNs models. As discussed above, the MCD and RMSE values obtained using GANs without dual supervised learning are slightly better than DBNs models. However, as shown in Figure 6.2 (D1), when converting the emotional voice by GANs without dual supervised learning, there is significant loss in the converted spectrogram, especially in the high frequency part. We predict this is because, unlike normal object image generation, the dissimilarities between the source and target emotional voice are very small. Therefore, without dual supervised learning, sometimes, it is hard to obtain a good result due to the problem of insufficient

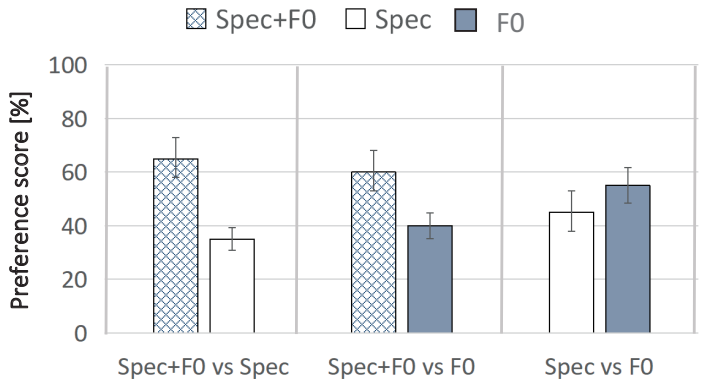

Figure 6.3: Results of XAB test for similarity evaluation of the conversion from angry to neutral voice by dual-SANs with different features.

data. As shown in Figure 6.2 (E1) and Figure 6.2 (E2), the converted spectrogram and F0 by DBNs do not lose the information in high frequency part as GANs, however, there is apparent over-smoothing. As shown in Figure 6.2 (C1) and Figure 6.2 (C2), we can clearly see that adding the dual supervised learning to the GANs can obtain converted spectrograms and F0 which are most similar to the target do not suffer from over-smoothing. This is due to the adversarial loss that does not require explicit density estimation; the dual supervised learning can keep the converted spectrogram feature regularized.

6.3.3 Subjective Experiment

For the subjective experiment, naturalness and similarity were used as evaluation metrics.

6.3.3.1 Similarity

First, we conducted a similarity evaluation by carrying out a XAB test to explore the contributions of the spectral and F0 features for emotional VC. As shown in Figure 6.3, "Spec" represents only the spectrum being used for conversion. In "F0", only F0 is used for conversion. And, in "Spec+F0", both the spectrum and F0 are converted using our dual-SANs. We did the XAB test for each of them, in which, 20 sentences were evaluated by 10 listeners. Each subject

listened to the voice of the target emotion, and then to the voice converted by the two methods. Listeners were asked to judge which sample sounded more similar to the target emotion. The error bars show 95% confidence intervals.

The results show that converting the spectrum or F0 only is not as effective as converting both. Comparing the results of using the spectrum only and F0 only, in our proposed dual-SANs, the F0 contributes slightly more than the spectrum in the emotion conversion.

Table 6.5: Results of classification for recorded (original) voices and converted voice by different methods [%].

	(a) original recorded voice				(b) DBNs+LG				(c) DBNs+NNs				(d) GANs				(e) dual-SANs			
Tar./Percept	Ang	Sad	Hap	Neu	Ang	Sad	Hap	Neu	Ang	Sad	Hap	Neu	Ang	Sad	Hap	Neu	Ang	Sad	Hap	Neu
Angry	90	2	0	8	27	5	0	68	55	14	11	20	33	21	2	44	71	4	1	24
Sad	2	97	0	1	3	33	2	62	5	55	1	39	39	45	15	1	12	68	3	17
Happy	0	0	98	2	3	5	18	74	4	3	64	29	3	2	47	48	2	3	80	15
Neutral	0	0	5	95	24	30	27	19	10	16	21	53	18	22	23	37	12	17	11	60

We then carried out a subjective emotion classification test, for the emotional to neutral pairs (H2N, S2N, and A2N) and their inverse conversion (N2H, N2S, and N2A) comparing different methods (DBNs+LG, DBNs+NNs, GANs, Dual-SANs). For each test model, 60 utterances (10 for angry, 10 for sad, 10 for happy and 30 for neutral) are selected, and 10 listeners are involved. The listeners are asked to label a converted voice as Angry, Sad, Happy or Neutral. As shown in Table 6.5 (a), when evaluating the original recorded emotional speech utterances, the classifier performed quite well, hence the corpus can be used in the emotion classification test. The classification results for the converted voices of DBNs+LG, DBNs+NNs, GANs and dual-SANs are shown in (b), (c), (d) and (e) of Table 6.5, respectively.

As shown in Table 6.5 (b), the conventional DBNs+LG method shows poor performance in all emotional VC tasks, especially for the conversion of the emotional voice to neutral voice. Therefore the F0 features converted by the conventional logarithm Gaussian method are not enough for emotional VC under the normal VC framework.

Comparing the results of Table 6.5 (c) with Table 6.5 (d), it is clear that DBNs+NNs method obtains better classification results than the GANs model without dual supervised learning, although the GANs model yielded a slightly

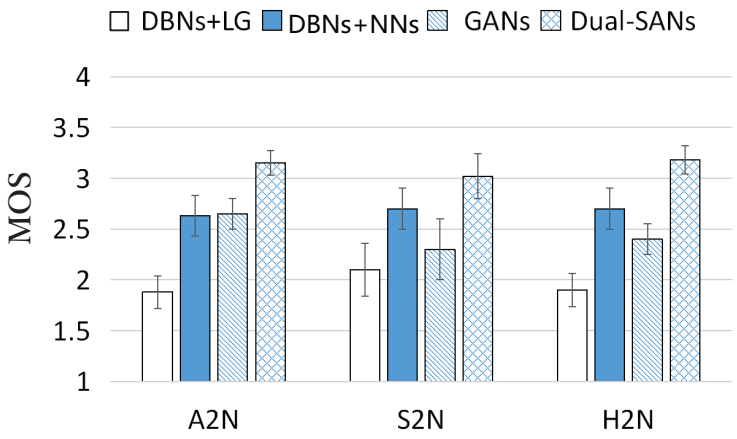

Figure 6.4: MOS of the naturalness evaluation for the conversion of emotional voice to neutral voice using the dual-SANs model

better result than the DBNs+NNs method in the objective experiment. This result confirms that the GANs model training without the dual loss may lead to instability and de-regularized processing of some converted samples.

With reference to the results of dual-SANs shown in Table 6.5 (e), the proposed method yielded about 70% classification accuracy on average, which has indicated a better result than the other models. Especially for the conversion of neutral voice to emotional voice, the average classification accuracy of the converted emotional voice is about 75%, which is 15% higher than the converted neutral voice.

6.3.3.2 Naturalness

In line with most previous work in the VC field, to measure naturalness, we also conducted a MOS test for naturalness evaluation. The scale ranged from totally unnatural (1) to completely natural (5). The evaluations of the naturalness of voice converted by different models are shown in Figure 6.4 and Figure 6.5. In this test, a higher value indicates a better result, where the error bar shows the 95% confidence interval. From these results, we can see that all naturalness scores are above or near 2, which means that the naturalness is not bad. Comparing the Figure 6.4 and Figure 6.5, we can see that the naturalness of the conversion from a neutral voice to emotional voice is slightly better than their inverse conversion

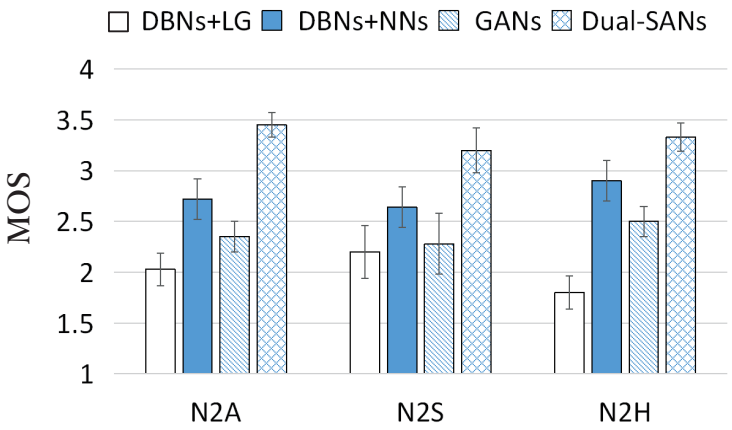

Figure 6.5: MOS of the naturalness evaluation for the conversion of neutral voice to emotional voice using the dual-SANs model

when using our proposed dual-SANs method. Comparing the results of different models, we can see that using the deep learning models for F0 training (NNs, GANs, dual-SANs) are all better than the linear conversion (LG) approach, and, our proposed dual-SANs model can obtain better naturalness than other models overall.

6.4 Chapter Conclusions

This study proposed an emotional VC method using dual supervised adversarial networks (dual-SANs) with MCC and CWT-F0 features. In order to obtain better training results, we segment MCC features and AS-CWT features to sequences with computable size, and then separately trained them with the proposed dual-SANs. A comparison between the proposed method and the conventionally used methods shows that our proposed model can effectively change the prosody of the emotional voice due to GANs' ability to mitigate the over-smoothing problem caused in the low-level data space. We also compared our proposed method to the model using GANs without dual supervised learning, and the results show that utilizing the intrinsic connection between two dual tasks can strengthen and regularize the training process in the emotional VC.

Chapter 7

Conclusions

The thesis describes a new framework for emotional voice conversion based on the recently-fashionable deep neural network methods.

In Chapter 4, we proposed a method using DBNs to train the MFCC features to construct the mapping relationship of the spectral envelopes, while using NNs to train the AS-CWT F0 features. Such features are conducted by the F0 features with arbitrary scales for prosody conversion between the source and target speakers.

In Chapter 5, to increase the training data for the emotional VC, we carry out the cross-wavelet transforming for CWT-F0 features of different emotional voices to get the XWT-F0 features. The MCD and RMSE results show that our proposed model can effectively change the acoustic and the prosody for the emotional voice at the same time in an objective experiment, and the MOS results also represent the better effect in the subjective experiment.

In Chapter 6, we propose a novel training model using the duality of the emotional VC to improve the robustness of the GANs models. A comparison between the proposed method and the conventionally used methods shows that our proposed model can effectively change the prosody of the emotional voice due to GANs' ability to mitigate the over-smoothing problem caused in the low-level data space. We also compared our proposed method to the model using GANs without dual supervised learning, and the results show that utilizing the intrinsic connection between two dual tasks can strengthen and regularize the training process in the emotional VC.

When using our proposed emotional VC methods, however, some problems remain. Specifically, the proposed model must extract the parallel speech data, which can limit the process to a one-to-one conversion only. In future works, we will explore a many-to-many emotional VC method and use it in other applications, such as emotional voice recognition [64] or facial expression recognition [65].

References

[1] Shinya Mori, Tsuyoshi Moriyama, and Shinji Ozawa. Emotional speech synthesis using subspace constraints in prosody. *in ICME*, pages 1093–1096, 2006. 1

[2] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo 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(1), 2014. 1, 22, 27

[3] Jelena Krivokapić. Rhythm and convergence between speakers of American and Indian English. *Laboratory Phonology*, 4(1):39–65, 2013. 1

[4] Tuomo Raitio, Lauri Juvela, Antti Suni, Martti Vainio, and Paavo Alku. Phase perception of the glottal excitation of vocoded speech. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. 1

[5] Zhi-Wei Shuang, Raimo Bakis, Slava Shechtman, Dan Chazan, and Yong Qin. Frequency warping based on mapping formant parameters. In *Ninth International Conference on Spoken Language Processing*, 2006. 1

[6] Daniel Erro and Asunción Moreno. Weighted frequency warping for voice conversion. *in Interspeech*, pages 1965–1968, 2007. 1

[7] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235, 2007. 1

[8] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj. Voice conversion using partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(5):912–921, 2010. 1

[9] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Exemplar-based voice conversion in noisy environment. *in Spoken Language Technology Workshop (SLT)*, pages 313–317, 2012. 1, 22, 27

[10] Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W. Black, and Kishore Prahallad. Voice conversion using artificial neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3893–3896. IEEE, 2009. 2

[11] Toru Nakashika, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Voice conversion in high-order eigen space using deep belief nets. *in Interspeech*, pages 369–372, 2013. 2, 17, 20, 42, 69, 73

[12] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, 1986. 2

[13] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006. 2

[14] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):580–587, 2015. 2

[15] Kun Liu, Jianping Zhang, and Yonghong Yan. High quality voice conversion through phoneme-based linear mapping functions with straight for Mandarin. *in Fuzzy Systems and Knowledge Discovery*, 4:410–414, 2007. 2, 30

[16] Manuel Sam Ribeiro and Robert AJ Clark. A multi-level representation of F0 using the continuous wavelet transform and the discrete cosine transform. *in ICASSP*, pages 4909–4913, 2015. 2, 30

[17] Martti Vainio, Antti Suni, Daniel Aalto, et al. Continuous wavelet transform for analysis of speech prosody. *in TRASP 2013-Tools and Resources for the Analysys of Speech Prosody*, 2013. 2, 30, 51

[18] Antti Santeri Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, Martti Vainio, et al. Wavelets for intonation modeling in HMM speech synthesis. In *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013. 2, 3, 14, 17, 30, 34, 51

[19] Huaiping Ming, Dongyan Huang, Minghui Dong, Haizhou Li, Lei Xie, and Shaofei Zhang. Fundamental frequency modeling using wavelets for emotional voice conversion. *in Affective Computing and Intelligent Interaction (ACII)*, pages 804–809, 2015. 2, 3, 17, 30, 43, 51, 59

[20] Huaiping Ming, Dongyan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li. Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion. *in Interspeech*, 2016. 2, 3

[21] Pierre Goupillaud, Alex Grossmann, and Jean Morlet. Cycle-octave and related transforms in seismic signal analysis. *Geoexploration*, 23(1):85–102, 1984. 2

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3, 63

[23] Yingce Xia, Tao Qin, Wei Chen, Jiang Bian, Nenghai Yu, and Tie-Yan Liu. Dual supervised learning. *in ICML*, 2017. 3, 63, 67

[24] Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, Yusuke Ijima, Kaoru Hiramatsu, and Kunio Kashino. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2017)*, pages 4910–4914, 2017. 3, 63

[25] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. *Proc. Interspeech 2017*, pages 1283–1287, 2017. 3, 63

[26] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016. 3, 63

[27] Bruce P Bogert, Michael JR Healy, and John W Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Proceedings of the symposium on time series analysis*, volume 15, pages 209–243. chapter, 1963. 10

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

[29] Ben Milner and Xu Shao. Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model. In *Proc. Interspeech*, pages 2421–2424, 2002. 12

[30] Zbynek Tychtl and Josef Psutka. Speech production based on the mel-frequency cepstral coefficients. In *EUROSPEECH*, volume 99, pages 2335–2338, 1999. 12

[31] Tenkasi Ramabadran, Jeff Meunier, Mark Jasiuk, and Bill Kushner. Enhancing distributed speech recognition with back-end speech reconstruction. In *INTERSPEECH*, pages 1859–1862, 2001. 12

[32] Satoshi Imai. Cepstral analysis synthesis on the mel frequency scale. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, volume 8, pages 93–96. IEEE, 1983. 12

[33] Ingrid Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE transactions on information theory*, 36(5):961–1005, 1990. 14

[34] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, 27(3):187–207, 1999. 14

[35] Hideki Kawahara, Masanori Morise, Toru Takahashi, Ryuichi Nisimura, Toshio Irino, and Hideki Banno. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3933–3936. IEEE, 2008. 14

[36] Arthur P Dempster, Nan M Laird, Donald B Rubin, et al. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society*, 39(1):1–38, 1977. 22

[37] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2000. 22

[38] Mikkel Schmidt and Rasmus Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Interspeech*, 2006. 22

[39] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):550–563, 2010. 22

[40] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1066–1074, 2007. 22

[41] Nasser Mohammadiha, Jalal Taghia, and Arne Leijon. Single channel speech enhancement using bayesian nmf with recursive temporal updates of prior distributions. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4561–4564. IEEE, 2012. 22

[42] Jort F Gemmeke, Tuomas Virtanen, and Antti Hurmalainen. Exemplar-based sparse representations for noise robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2067–2080, 2011. 22

[43] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180. IEEE, 2003. 22

[44] Arshia Cont, Shlomo Dubnov, David Wessel, et al. Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints. In *Proceedings of Digital Audio Effects Conference (DAFx)*, 2007. 22

[45] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. Probabilistic spectrum envelope: Categorized audio-features representation for nmf-based sound decomposition. In *INTERSPEECH*, pages 1765–1768, 2011. 22

[46] Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. Constrained spectrum generation using a probabilistic spectrum envelope for mixed music analysis. In *ISMIR*, pages 181–184, 2011. 22

[47] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li. Exemplar-based voice conversion using non-negative spectrogram deconvolution. In *the 8th ISCA Speech Synthesis Workshop*, 2013. 22, 27

[48] Ryo Aihara, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Exemplar-based individuality-preserving voice conversion for articulation disorders in noisy environments. In *INTERSPEECH*, pages 3637–3641, 2013. 22, 27

[49] 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*, 2014. 22, 27

[50] Ryoichi Takashima, Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki. Noise-robust voice conversion based on spectral mapping on sparse space. In *SSW8*, pages 71–75, 2013. 22, 27

[51] Peter Földiak. Forming sparse representations by local anti-hebbian learning. *Biological cybernetics*, 64(2):165–170, 1990. 22

[52] Julian Eggert and Edgar Korner. Sparse coding and nmf. In *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, volume 4, pages 2529–2533. IEEE, 2004. 22

[53] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research*, 5:1457–1469, 2004. 22

[54] Patrik O Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565. IEEE, 2002. 22

[55] Weixiang Liu, Nanning Zheng, and Xiaofeng Lu. Non-negative matrix factorization for visual coding. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 3, pages III–293. IEEE, 2003. 22

[56] Elina Helander, Hanna Silén, Tuomas Virtanen, and Moncef Gabbouj. Voice conversion using dynamic kernel partial least squares regression. *IEEE transactions on audio, speech, and language processing*, 20(3):806–817, 2012. 27

[57] Zhaojie Luo, Jinhui Chen, Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki. Emotional voice conversion using neural networks with different temporal scales of F0 based on wavelet transform. In *9th ISCA Speech Synthesis Workshop*, pages 140–145. 29, 31, 34, 42, 52, 56

[58] Zhaojie Luo, Tetsuya Takiguchi, and Yasuo Ariki. Emotional voice conversion using deep neural networks with mcc and f0 features. In *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*, pages 1–5. IEEE. 29

[59] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. Emotional voice conversion using neural networks with arbitrary scales f0 based on wavelet transform. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1), 2017. 29, 52, 70

[60] Hideki Kawahara. STRAIGHT, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006. 29, 30, 54

[61] Takako Toda et al. Interlanguage phonology: Acquisition of timing control and perceptual categorization of durational contrast in japanese. *The Australian National University*, 2013. 33

[62] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008. 34

[63] Hiromichi Kawanami, Yohei Iwami, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. GMM-based voice conversion applied to emotional speech synthesis. *In: Proc. European Conf. on Speech Communication and Technology (Eurospeech03)*, pages 2401–2404, 2003. 40, 56, 69

[64] Dimitrios Ververidis and Constantine Kotropoulos. Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181, 2006. 49, 84

[65] Jinhui Chen, Zhaojie Luo, Tetsuya Takiguchi, and Yasuo Ariki. Multithreading cascade of surf for facial expression recognition. *EURASIP Journal on Image and Video Processing*, 2016(1):37, 2016. 49, 84

[66] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. Neutral-to-emotional voice conversion with cross-wavelet transform f0 using generative adversarial networks. *APSIPA Transactions on Signal and Information Processing*, 8, 2019. 51

[67] Christopher Torrence and Gilbert P Compo. A practical guide to wavelet analysis. *Bulletin of the American Meteorological society*, 79(1):61–78, 1998. 52, 53

[68] Lonnie Hudgins, Carl A Friehe, and Meinhard E Mayer. Wavelet transforms and atmopsheric turbulence. *Physical Review Letters*, 71(20):3279, 1993. 52

[69] Xiao Jing Cai, Shuairu Tian, Nannan Yuan, and Shigeyuki Hamori. Interdependence between oil and east asian stock markets: Evidence from wavelet coherence analysis. *Journal of International Financial Markets, Institutions and Money*, 48:206–223, 2017. 52

[70] Aslak Grinsted, John C Moore, and Svetlana Jevrejeva. Application of the cross wavelet transform and wavelet coherence to geophysical time series. *Nonlinear processes in geophysics*, 11(5/6):561–566, 2004. 52

[71] Swati Banerjee and Madhuchhanda Mitra. Application of cross wavelet transform for ecg pattern analysis and classification. *IEEE transactions on instrumentation and measurement*, 63(2):326–333, 2014. 52

[72] Ingrid Daubechies. *Ten lectures on wavelets*, volume 61. Siam, 1992. 53

[73] Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354*, 2017. 57, 70

[74] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. Emotional voice conversion with adaptive scales f0 based on wavelet transform using limited amount of emotional data. In *INTERSPEECH*, pages 3399–3403, 2017. 63

[75] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform f0 features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1535–1548, 2019. 63

[76] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jungkwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017. 64

[77] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2017. 64

[78] Zili Yi, Hao Zhang, Ping Tan Gong, et al. DualGAN: unsupervised dual learning for image-to-image translation. *arXiv preprint arXiv:1704.02510*, 2017. 64

[79] Takuhiro Kaneko and Hirokazu Kameoka. Parallel-data-free voice conversion using cycle-consistent adversarial networks. *arXiv preprint arXiv:1711.11293*, 2017. 64, 70, 73

[80] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. *arXiv preprint arXiv:1804.00425*, 2018. 64

[81] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 65

[82] Olof Mogren. C-rnn-gan: Continuous recurrent neural networks with adversarial training. *arXiv preprint arXiv:1611.09904*, 2016. 65

[83] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 65

[84] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 71

[85] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015. 71

[86] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 71

[87] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, 2013. 71

[88] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 71

[89] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016. 71

[90] D Kinga and J Ba Adam. A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 71

Acknowledgements

First, I would like to thank my supervisors, Professor Tetsuya Takiguchi, Professor Ryoichi Takashima and Emeritus Professor Yasuo Ariki 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 Ohta, 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. LUOKBUPhDThesis2020,
           title={{Emotional Voice Conversion using
                   Deep Learning Models}},
           journal={Doctoral Thesis},
           author={Zhaojie Luo},
           institution={Kobe University},
           month={Mar.},
           year={2020}
          }
```

Doctor Thesis, Kobe University
"Emotional Voice Conversion using Deep Learning Models", 120 pages
Submitted on January, 21, 2020.
The date of publication is printed in the cover of repository version published in Kobe University Repository Kernel.

©Zhaojie Luo
All Rights Reserved, 2020.