A study on rule-based emotional voice conversion with degree continuously controllable in dimensional space using the inverse three-layered model

Yawen Xue

Japan Advanced Institute of Science and Technology

Doctoral Dissertation

A study on rule-based emotional voice conversion with degree continuously controllable in dimensional space using the inverse three-layered model

Yawen Xue

Supervisor: Professor Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology

September 2018

Abstract

In terms of human-computer interaction (HCI), synthesized speech has burgeoned at a rapid rate in recent years to fulfill demands for daily speech communication. Natural sounding synthetic speech with only linguistic information is currently used in modern applications such as text to speech systems (TTS), navigation systems, robotic assistants, story teller systems and speech to speech translation systems (S2ST). Information conveyed by speech should be summarized through linguistic information, paralinguistic information as well as nonlinguistic information. Synthesized speech with only linguistic information cannot encompass all of these factors. Therefore, emotional synthesized speech that allows communication of nonlinguistic information is increasingly required. Emotions, ranging from an underlying emotional state to full-blown emotions, contribute substantially to the acoustic manifestation of the spoken language. In order to incorporate emotions to neutral TTS synthesized speech, an emotional voice conversion system which can convert the neutral speech to emotional one is necessary to cope with.

Previous prevalent methods concerned with emotional voice conversion systems mainly used statistical approach such as Gaussian mixture models, deep neural network, or neural network method. They mainly utilized some simple discrete labels such as joy, sad, anger to represent emotion. However, humans have ability to express mild emotion such as a little bit joy or very joy which is a continuum of non-extreme states. This means statistical approaches need large parallel linguistic database of affective voices with a continuum of non-extreme affective states which is a costly and impossible problem. Little attention is paid on controlling the emotional degree in a continuous scale in previous studies. A rule-based voice conversion technique which can obtain variation tendencies of acoustic features with a limited database is utilized in this research.

When modeling continuous controllable degrees of emotional synthetic speech, two primary problems are firstly needed to be considered. The first one is how to describe emotions and another problem is how to model emotion perception process of human beings.

In the literature, there are many descriptive systems for emotions. The most straightforward description is the utilization of emotion-denoting words or category labels called emotion category. Emotions in daily speech communication are highly diverse. Many human-machine dialogues need machines to express mild and non-extreme emotional states. Therefore, an emotion dimensional approach which satisfies the requirement to express a range from low-intensity to high intensity states is appropriate for representing a continuum of non-extreme emotional states for controlling the degree of emotion. In this research, two dimensions arousal (synonymous to activation and activity) and valence (synonymous to evaluation) are used for emotional speech conversion based on the database we have.

Another problem related to emotion conversion is modeling the process of expression and perception of emotion by human beings. Many researchers based their theory and research on a modified version of the Brunswik's functional lens model of perception. Huang and Akagi proposed a three-layered model for expressive speech perception with emotion (listener attributions) at the top layer, semantic primtives (proximal percepts) at the middle layer, and acoustic feature (distal indicators) at the bottom layer. They assumed that humans perceive emotion not directly from acoustic features, but semantic primitives, such as fast, bright, and so on also play important roles. The three-layered model had already been applied by many researchers in the emotion recognition area. In this research, we assume that the human production of emotion follows the opposite direction of human perception. This means the encoding process of the

speaker is the inverse process of the decoding of the listener. In that case, an inverse three-layered model is employed as the structure between emotion and acoustic feature.

The related acoustic features to each dimension are investigated as applied to emotional speech synthesis. Subjects are asked to evaluate the synthesized speech, the specific acoustic features of which, such as F0, have been replaced by the F0 from the emotional speech but leaving the other acoustic features of the neutral speech. We concluded that both the F0 trajectory and spectral sequence are important to emotion conversion. The power envelope and duration show little influence on the valence axes. In this research, we focused more on the prosody-related features such as duration, F0 and power envelope.

In order to control the emotion degree in dimensional space using the inverse three-layered model, an emotion conversion system was proposed with two inputs (positions in dimensional space and neutral speech) and two steps (rule extraction and rule application). In the first step, the rules between acoustic feature variations of neutral and emotional ones can be extracted using a fuzzy inference system. The inverse three-layered model is set as the structure between emotion dimension and acoustic features with emotion dimension as the bottom layer, semantic primitive layer in the middle and acoustic features layer at the top. The second step is to apply the rule-based voice conversion method to modify the acoustic features of neutral speech to emotional ones following the rules extracted from the first step. It is widely understood that emotion is conveyed by means of a number of prosodic parameters such as voice quality and speech rate as well as fundamental frequency. In this step, some essential prosody features such as duration, F0 contour and power envelope are parameterized by an interpolation method, Fujisaki model and target prediction model. Then the modified acoustic features are synthesized using STRAIGHT. Perceptual evaluation results in V-A space show that the synthesized speech of joyful, sad and cold angry emotion can be perceived well, including the category and the degree, although the perceived degree is decreased compared to the desired values. For hot anger emotion, since the spectral modification was not conducted, the synthesized speech of hot anger is perceived as a joyful emotion.

Commonalities and differences of human perception for perceiving emotions in speech among different languages in dimensional space have been investigated in Han et al., 2016. Results show that human perception for different languages is identical in dimensional space. According to this result, we assume that, given the same direction in dimensional space, we can convert the neutral voices in multiple languages to emotional ones with the same impression of emotion. It means that the emotion conversion system could work for other languages even if it is trained with a databases in one language. We try to convert neutral speech in two different languages, English and Chinese using an emotion conversion system trained with Japanese database. We find that all converted voices can convey the same impression as Japanese voices. On the case, we can make a conclusion that given the same direction in dimensional space, the synthesized speech among multiple language can convey the same impression of emotion. In a word, the Japanese emotion conversion system is compatible to other languages.

In conclusion, this research proposed a method for emotional voice conversion with degree continuously controllable using dimensional representation following human emotion production mechanism. Perception results show that the synthesized stimuli can be perceived with the same tendency as intended in dimension space except hot anger. Neutral voices in other languages were directly inputted into the system without training, and perception results show that the conversion system built in one language is capable for other languages without training. The emotional navigation systems, robotic assistants and S2ST system will bring an intelligent HCI and enormous promotion in human life quality. As this system enables to convert the input neutral speech from any target speaker in any language without training, it can reduce an amount of cost and make the emotional TTS applicable. And this will give a big progress in the field of emotional voice conversion. The emotional navigation systems, robotic assistants and S2ST system will bring an intelligent human computer interface HCI and enormous promotion in human life quality.

Keywords: Emotional voice conversion, rule-based speech synthesis, emotion dimension, three-layered model, Fujisaki F0 model, target prediction model.

Acknowledgments

Five years have already passed since I came to Japan and lots of things happened during the five years. I am used to the life in Japan. And I tried to speak Japanese with the surroundings. This five years is the most precious time during my life. Many people gave me lots of helps during the five year and I would like to express my thanks to them.

Firstly, I would like to express my sincere gratitude to my supervisor Professor Masato Akagi of Japan Advanced Institute of Science and Technology (JAIST) for his support, encouragement, guide, and kindness during the five years. It is Prof. Akagi who opened the window of research and taught me many knowledge related to the speech signal processing. Prof. Akagi not only set a good example for me in research but also show me how to be a good teacher. When I faced some difficulty he always smiled to me and guided me to become calm in order to solve it. And he also gave me the freedom to do research so I can have obtain the key to do research by myself not the research topic itself.

I wish to express my thanks to my co-advisor Professor Masashi Unoki of JAIST. Every time, when I gave a presentation, Professor Unoki could always checked carefully and gave me many important comments so that I can notice something that I neglected. And sometimes Prof. Unoki was on a trip when I gave a presentation or rehearsal, he still gave some suggestions although he was busy and tired. Especially when I applied the JSPS grant, Prof. Unoki gave me lots of suggestions related to my presentation. Without his help, the application will not become so smooth.

I am also grateful to Professor Jianwu Dang of JAIST and Tianjin University, China who is the advisor of my minor research at JAIST for his suggestion and comments, as well as be a member of the dissertation committee. When I talked with Prof. Dang, I could feel his energy for research which inspire me a lot.

I would like to express my thanks to Associate Professor Nose Takashi of Tohoku University for his several comments and discussion related to my research during some conference and through email. Prof. Nose taught me many knowledge related to statistical parametric speech synthesis. I will also be grateful if further corporation can be made.

I am also thankful to Professor Yoshitaka Atsuo for serving as a member of the dissertation committee and for his comments and suggestions.

I would like to express my gratitude to Junior Professor Peter Birkholz of Technology University, Dresden, Germany. Prof. Birkholz helped me lot for the three months' internship in Dresden, preparing the German visa, applying the guest house and being used to the life in Germany. During my stay in Dresden for three months, Prof. Birkholz taught me many knowledge about the articulatory speech synthesis, helped me do the experiment of Magnetic Resonance Imaging, gave me the permission to use the software, VocalTractLab and 3D image and revised the journal paper carefully.

I would like to thank Yasuhiro Hamada of JAIST who collaborated with me on building the emotional voice conversion system, shared the program of Fujisaki model with me and gave me many suggestions during my presentation.

I would like to express my thanks to Reda Elbarougy of JAIST. I sometimes faced some problems when I just started research. Every time I emailed him, he replied me kindly and patiently which helped me be familiar with this research quickly.

I would like to thank YongweiLi and YuxuanDu who gave me many information before I entered JAIST and gave me many helps either in life or in research after I entered JAIST.

I would like to express my gratitude to Xiao Han who is the tutor when I was the first year at JAIST. Xiao Han helped a lot on being familiar with the life and study in JAIST.

I want to appreciate XingfengLi, NGO Thuan Van, Rieko Kubo and all members in Acoustic Information Science Laboratory for their valuable comments, helps and encouragements during my research and life in JAIST. The discussion among us always provided me many new ideas and gave me a lot of happiness during the boring research period.

I want to express my thanks to JAIST to provide the 5D scholarship and support me for the internship in Germany. JAIST also provide me the top-rank research environment. The staff of student section at JAIST helped me revise many documents related to some application and provided me many conveniences of the life at JAIST.

I am also thankful to Japan Society For The Promotion Of Science (JSPS) to support me in the last year of my doctor period. And the Telecommunications Advancement Foundation provided me the support for attending the oversea conference.

Lastly I would like to delicate this dissertation to my father and mother for their

forever support, understanding and love of me. Without them, I can not achieve all things I have now. What's more, I am sincerely grateful to Mr. Song ZHANG for his love, company, sharing and understanding.

Table of Contents

Abstract **1**

Acknowledgments **4**

Table of Contents **7**

List of Figures **9**

List of Tables **12**

1 General introduction **14**

1.1 Research motivation 14

1.2 Research concept 16

1.2.1 Representation of emotion 16

1.2.2 Modeling the perception of emotions 17

1.2.3 Method for synthesizing emotional speech 18

1.3 Research method 22

1.4 Research novelty 23

1.5 Outline of thesis 24

2 Descriptive frameworks for emotions **27**

2.1 Categorical representation 27

2.2 Dimensional representation 28

2.3 Other descriptions 29

3 Inverse three-layered model **32**

3.1 Modified Brunswik's functional lens model for emotion perception 32

3.2 Three-layered model for emotion perception . . . 34
3.3 Inverse three-layered model for emotion production . . . 35

4 Acoustic features related to emotion dimensions 39
4.1 Acoustic features replacement procedure . . . 40
4.2 Experiment . . . 41
4.3 Results . . . 43
4.4 Discussion . . . 43

5 The emotional voice conversion system 49
5.1 Rule extraction . . . 51
5.1.1 Database . . . 51
5.1.2 Acoustic feature extraction . . . 51
5.1.3 Semantic primitives evaluation . . . 54
5.1.4 Emotion dimensions evaluation . . . 55
5.1.5 Fuzzy inference system . . . 55
5.1.6 Applying a fuzzy inference system for extracting rules . . . 57
5.1.7 System evaluation . . . 59
5.2 Rule application . . . 60
5.2.1 Fujisaki model for parameterizing F0 contour . . . 61
5.2.2 Target prediction model for parameterizing the power envelope . . . 65
5.3 Perceptual evaluation . . . 71
5.3.1 Stimuli . . . 71
5.3.2 Experiment procedure . . . 73
5.3.3 Subjective evaluation result in V-A space . . . 73
5.3.4 Subjective evaluation result of naturalness . . . 75

6 Emotion conversion for multiple languages 77
6.1 Commonalities of human perception for emotional speech among multi-languages . . . 77
6.2 Applying conversion system to other languages . . . 78
6.2.1 Listening Test . . . 80
6.2.2 Results . . . 88

6.3 Discussion 89

7 Discussion **90**

8 Conclusion **93**

8.1 Summary 93

8.2 Contribution 95

8.3 Future works 95

Bibliography **97**

Publications **112**

List of Figures

1.1 *The diagram of S2ST.* 16
1.2 *Scherers [29] modified Brunswikian lens model adopted for vocal communication of emotion.* 18
1.3 *Three-layered model [34].* 18
1.4 *The concept of unit-selection in concatenative speech synthesis [137].* 19
1.5 *The concept of statistical parametric speech synthesis [137].* 20
1.6 Organization of this dissertation. 26

2.1 *Dimensional representation.* 30
2.2 *Russels circumplex model of emotion [23].* 31

3.1 *The relationship among Brunswik's lens model, three-layered model and the inverse three-layered model.* 33
3.2 *The speech chain (speech production and speech perception) [79].* 37
3.3 *The proposed inverse three-layered model.* 38

4.1 *Procedure of acoustic feature replacement.* 41
4.2 Graphic interface of the perceptual test. 42
4.3 *Perceptual position values of original (Org) emotional utterances and synthesized (Rep) utterances on V-A space when F0 contour (F0) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)* 44
4.4 *Perceptual position values of original (Org) and synthesized (Rep) utterances on V-A space when time (TM) duration information is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)* 45

4.5 *Perceptual position values of original (Org) and synthesized (Rep) utterances on V-A space when power envelope (PW) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)* 46
4.6 *Perceptual position values of original (Org) and synthesized (Rep) utterances in V-A space when spectral sequence (SS) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)* 47

5.1 *Scheme of emotion conversion system.* 50
5.2 Structure of Fuzzy Inference System (FIS) 56
5.3 *Procedure for training ANFIS.* 57
5.4 *Applying ANFIS for estimating acoustic features (AF) and semantic primitives (SP).* 58
5.5 Mean absolute error of semantic primitives. 61
5.6 Mean absolute error of acoustic features from three- and two-layered model. 62
5.7 The procedure of modifying the neutral speech to emotional one. 63
5.8 F0 trajectory of a neutral speech (dashed) and synthesized speech (solid) [104]. 65
5.9 Procedure of reproduing power envelope 68
5.10 Speech wave of the original speech 68
5.11 Extracted power envelope 69
5.12 Target of power envelope estimated by target prediction model 70
5.13 Steplike targets of power envelope 71
5.14 Reproducing power envelope using 2nd-order critically damped model and the extracted power envelope from original speech 72
5.15 The evaluated and intended positions in V-A space. (the dashed lines are the intended position and the solid lines are the obtained position.) 74
5.16 The average and standard deviation of the evaluated results. (*1st, 2nd, 3rd stand for the intended quadrants.* Cold *and* hot *represent the intended cold anger and hot anger.* av *and* str *mean the average and standard deviation values of each quadrant.)* 75

5.17 Mean opinion scores for converted speech in each quadrant. 76

6.1 Emotional states position on Valence-Activation approach [88]. 79
6.2 Convert the source neutral speech from Japanese to other languages 79
6.3 Evaluation results (emotion category) of synthesized voices in Chinese. . . 80
6.4 Evaluation results (emotion category) of synthesized voices in English. . . . 81
6.5 Evaluation results (emotion category) of synthesized voices in Japanese. . . 82
6.6 Evaluation results (emotion degree) of synthesized voices in Chinese. . . . 83
6.7 Evaluation results (emotion degree) of synthesized voices in English. . . . 84
6.8 Evaluation results (emotion degree) of synthesized voices in Japanese . . . 84
6.9 Evaluation results (The mean opinion score) of Chinese synthesized voices. 85
6.10 Evaluation results (The mean opinion score) of English synthesized voices. 86
6.11 Evaluation results (The mean opinion score) of Japanese synthesized voices. 87

List of Tables

4.1 *Anova values of each acoustic features to valence and arousal(*$^{**}p < 0.01, ^{*}p < 0.05$*)*. 43

5.1 The content of sentences in English 52

5.2 The id number and the according expressive speech 53

5.3 Specification of speech data for Japanese database. 53

5.4 *Acoustic features used in this system* 54

Chapter 1

General introduction

1.1 Research motivation

In terms of human-computer interaction (HCI), synthesized speech has burgeoned at a rapid rate in recent years to fulfill the demand for daily speech communication. Natural sounding synthetic speech with only linguistic information is currently used in modern applications such as text to speech systems (TTS), navigation systems, robotic assistants, storyteller systems and speech to speech translation systems (S2ST).

Fujisaki [1] proposed that information conveyed by speech should be summarized in three parts:

- **Linguistic information**: which is discrete categorical information explicitly represented by the written language or uniquely inferred from context. Linguistic information can be represented either explicitly by the written language, or can be easily and uniquely inferred from context. It is discrete and categorical.

- **Paralinguistic information**: discrete and continuous information added by the speaker to modify or supplement the linguistic information, such as emphasis, or shades of meaning to what people say. For example, in one sentence "Tomorrow, my sister will have an examination." the speaker can put emphasis on "tomorrow", "my sister", or "examination" which will express different intentions of the speaker. Although the linguistic information is the same, the intention of the emphasis meaning changes a lot. Paralinguistic information can be both discrete and continuous. The intention of the speaker can be categorical such as assertion or question but

on the other hand, when the degree of the category is considered, it can also be continuous.

- **Nonlinguistic information**: information not generally controlled by the speaker, such as the speaker's emotion, gender, age, etc although it is a fact that an actor can control the emotion intentionally. These aspects are not directly to the paralinguistic information or linguistic information. Nonlinguistic information can be discrete and continuous which is the same as paralinguistic information.

Synthesized speech with only linguistic information cannot encompass all of these factors, thus resulting in machine-liked speech sounds. Affective synthesized speech that allows communication of nonlinguistic and paralinguistic information, such as affect and intent, is increasingly required [2] [3] [4]. Affect is not restricted to emotion; for instance [5] [6], there are social affective expressions, such as expression of politeness, sarcasm, irritation, flirtation, etc., which may be more or less controllable. Emotions, ranging from an underlying emotional state to full-blown emotions, contribute substantially to the acoustic manifestation of the spoken language. Incorporate the expression of emotions into the TTS synthetic speech can increase the affectiveness of the synthetic speech [7].

On the other hand, people in different countries have more and more aspiration to communicate with each other although they have different mother languages. However, it is impossible to have a conversation if a common language is not shared, that makes a challenge to design a worldwide communication environment. One possible solution to this challenge is to construct the S2ST that can convert a spoken utterance in one language to that of another language [9] [10] [11] [12] [13].

As shown in Fig. 1.1 from the speech in language A, firstly the system recognizes the linguistic information of the speech using an automatic speech recognition system (ASR). Then a translations system (TS) is applied to translate the text from language A to language B. Lastly, a TTS synthesizes the utterance in language B from the text. Using S2ST, people with different mother languages can communicate with each other.

While conventional S2ST systems only consider linguistic information. Non-linguistic information such as, emotional state conveyed by the source language is crucial to be preserved and passed in daily life. Communication without emotions will become boring and unrealistic. Therefore, in order to propose an affective S2ST system, an emotion

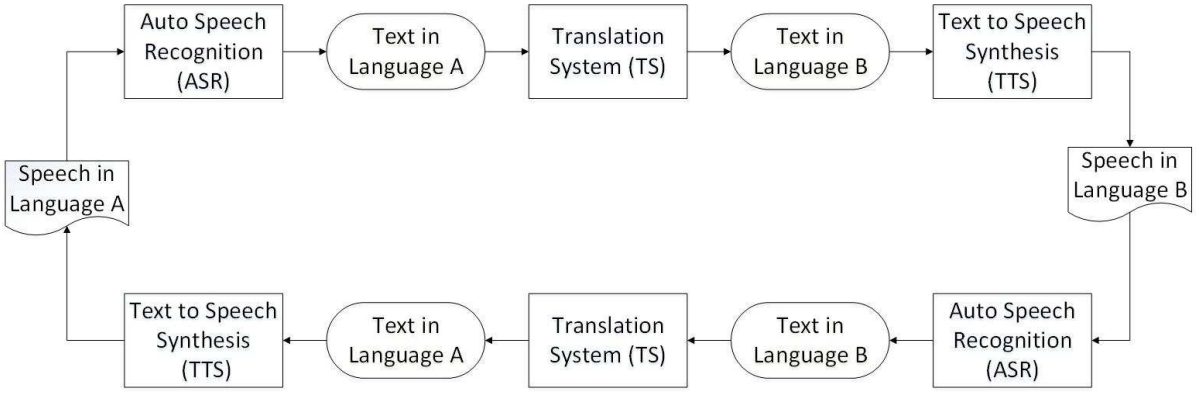

Figure 1.1: *The diagram of S2ST.*

recognition system that can recognize emotional states from different languages utterance and a system that can synthesize emotional speech in multiple languages are necessary to be explored. This research focuses on synthesizing emotional speech and would like to explore the possibility to apply the system trained with one language to others.

1.2 Research concept

When synthesis the emotion, three primary problems are to be considered; how to represent emotion; how to model the process of emotion perception and which method is suitable to use for synthesizing emotional speech.

1.2.1 Representation of emotion

In the literature, there are many descriptive systems for emotion. The most straightforward description is the utilization of emotion-denoting words or category labels [14] [15] [16] [17] [18] [19], called emotion category [20]. Also there are other less-well-known methods, prototype descriptions [21], appraisal-based descriptions [22], the circumplex model [23], physiological descriptions [24] and dimensional approaches [25] [50] [28].

Emotion in daily speech communication is highly diverse. All the above methods can not cover all types of context, on that case, we should think about the requirement of the daily application.

Many human-machine dialogues need machines to express mild and nonextreme emotional states. Therefore, an emotion dimensional approach which satisfies the requirement

to express a range from low-intensity to high-intensity states is appropriate for representing a continuum of non-extreme emotional states [25] for controlling the degree of emotion. This research conducts two dimensions following the work of Schroder [25], arousal and valence as the representation of emotion.

On that case, by moving the position in emotion dimensional space, the degree and category of emotion can be changed freely.

1.2.2 Modeling the perception of emotions

Another problem is how to model emotion by human beings. Many researchers [29] [30] [31] [32] base their theory and research on a modified version of the Brunswik's functional lens model [33] of perception. According to the model, as shown in Fig.1.2, a speaker expresses his/her emotional state using "distal indicator values", the acoustic features of the speech signal. The listener, however, perceives "distal indicator values" as "proximal percepts", that is, the subjective parameters such as pitch, voice quality, etc. Lastly, the listener uses "proximal percepts" to detect subjective attribution, that is emotion states.

Brunswik's model suggests that the process of perception of emotion is multi-layered. Huang and Akagi [34] as shown in Fig.1.3 proposed a three-layered model for expressive speech perception based on the Brunswik's model with emotion (listener attributions) at the top layer, semantic primitives (proximal percepts) at the middle layer, and acoustic feature (distal indicators) at the bottom layer. They assume that humans perceive emotion not directly from acoustic features, but semantic primitives, such as fast, bright, and so on also play important roles.

The three-layered model has already been applied by some researchers in the emotion recognition area [48] [36]. In this research, we assume that the human production of emotion follows the opposite direction of human perception. This means the encoding process of the speaker is the inverse process of the decoding of the listener. Hence, an inverse three-layered model is employed as the structure between emotion and acoustic feature following the process of human emotion perception.

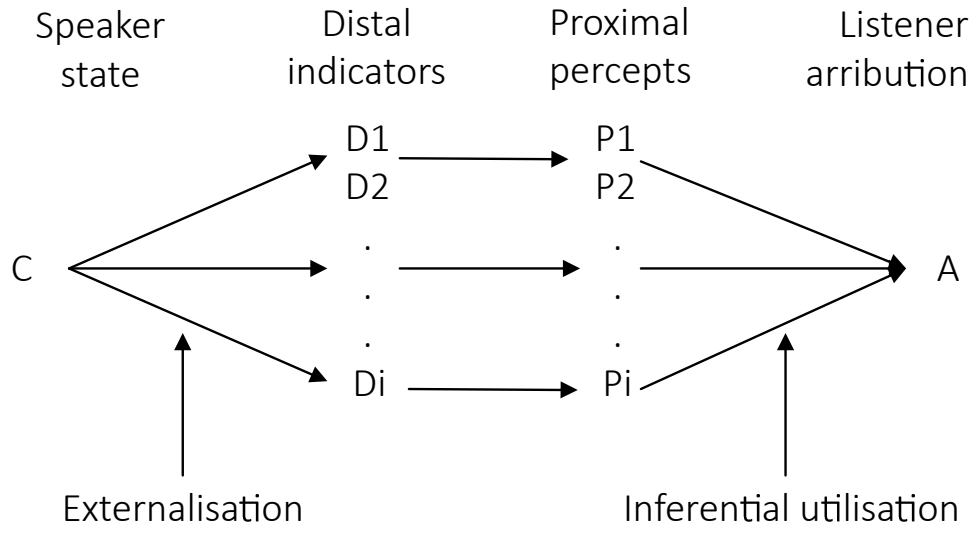

Figure 1.2: *Scherers [29] modified Brunswikian lens model adopted for vocal communication of emotion.*

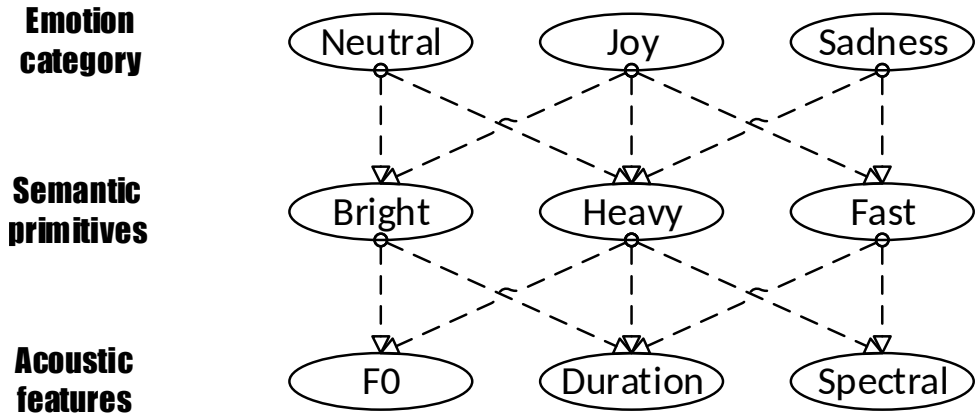

Figure 1.3: *Three-layered model [34].*

1.2.3 Method for synthesizing emotional speech

In order to incorporate emotion into neutral speech, many previous researches concentrated on emotional speech synthesis which directly synthesizes emotional speech waveform from the text, the emotional TTS. Prevalent emotional TTS mainly can be concluded as four directions: concatenative approach [38] [39] [40], [41] statistical approach [44] [45] [42], articulatory speech synthesis [46] [58] [59] and rule-based speech synthesis [56] [57].

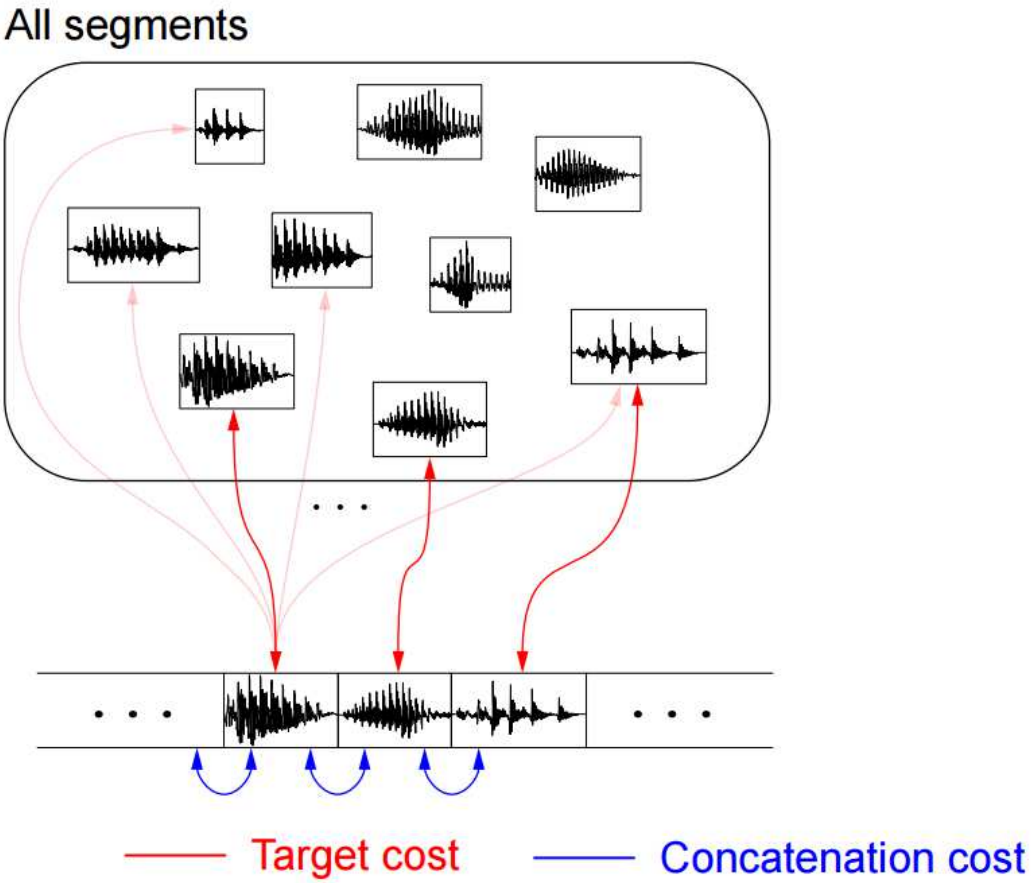

Figure 1.4: *The concept of unit-selection in concatenative speech synthesis [137].*

Concatenative speech synthesis

Concatenative speech synthesis such as unit-selection [38] [39] [40] [26] perceives segments of natural speech and then piece them together to form the desired speech output using target cost and concatenation cost as shown in Fig. 1.4. The best-synthesized speech quality can be achieved by so-called unit-selection synthesizers.

However, the quality of all concatenative synthesized voices depends much on the prerecorded database. The flexibility of modifying without a loss of quality is in limited which will lead the inconvenient for synthesizing the emotional speech.

Statistical parametric speech synthesis

Contrary to directly select the actual utterances from the speech database, statistical parametric speech synthesis has been popular over the last years. The main idea of the statistical parametric method is to generate the inherent average properties of the similar

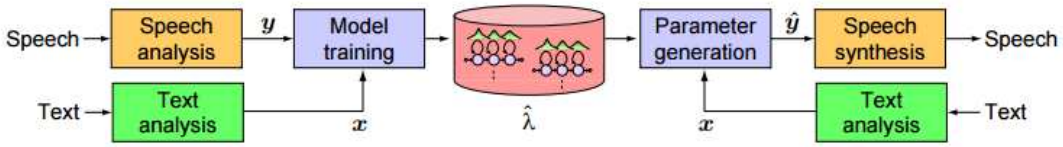

Figure 1.5: *The concept of statistical parametric speech synthesis [137].*

set of speech segments as shown in Fig. 1.5.

Statistical based synthesized speech such as HMM-based or GMM-based was reported to have the high intelligibility, the small footprint and low computation. However, the quality of synthesized speech often suffered the buzziness and over-smoothed problem which leads the low naturalness.

Articulatory speech synthesis

Neither the concatenative speech synthesizer nor the statistical parametric speech synthesizer considers the speech production mechanism of human beings. This prompt another speech synthesis direction, articulatory speech synthesis.

Articulatory synthesis directly simulates the principle of speech production based on the source-filter model. It has the potential to produce all aspects of speech production. However, speech production is a very complex process and not fully understood in every detail.

Rule-based synthesis

Rule-based synthesis or formant synthesis [56] [57] applies acoustic-domain rules to control the formant synthesizer. The generated rules are related to fundamental frequency, formant frequencies and parameters, etc. Advantages of rule-based synthesis are the flexibility, the ability to generate smooth transitions between segments and the relatively small training database.

Emotional voice conversion

The input of TTS system is text and the output is speech waveform. If the emotional speech would like to be synthesized, the TTS system needed to be trained again which

means change the neutral database to the emotional one.

As this will reduce the effectiveness, some researchers proposed voice conversion (VC) systems for emotional speech. This means, the synthesized neutral speech from conventional TTS can be directly converted to the emotional one.

Previous methods for emotional voice conversion utilized a categorical approach to express emotional states [7] for mono language.

One method is the piece-wise linear mapping using a probabilistic model, Gaussian Mixture Models (GMM) [62] [63] [64] [65]. Kawanami [14] first applied GMM for spectrum transformation to emotion voice conversion. Tao [15] tested three different methods for prosody conversion and found that GMM is suitable for a small database while a classification and regression tree model will give better results if a large context-balanced corpus can be obtained. Inanoglu [16] combined a Hidden Markov Model, GMM and F0 segment selection method for transforming F0, duration and short-term spectra in data-driven emotion conversion when large amounts of parallel data are needed. Aihara [17] improved the GMM-based emotional voice conversion for both voice quality and prosody feature conversion.

As the conversion function GMM is sometimes optimized to minimize a total error, excessively smoothed speech parameters are generated which will cause a muffled converted speech. In that case, to solve the over-smoothed problems, some methods [62] such as adding global variance [63] or modulation spectrum [64] to capture the over-smoothing effect and partly maintain the parameters of the source speech by dynamic frequency warping [65] have been proposed. However, the GMM method is a piece-wise linear mapping and human voice conversion is a non-linear one. Moreover, the GMM method is difficult to apply for a continuum representation of emotion, as it can only obtain an average in the statistical approach.

Non-linear methods such as Neutral Network (NN) [18] and Deep Neutral Network [19] are prevalently utilized. But both NN and DNN methods need large databases for training and it is a tough task to collect human responses to emotional voice.

On the other hand, a concatenative approach, like unit selection [66] [67] which selects the target syllable contours from a database using a cost function sometimes can synthesize emotional speech with good quality. But this approach also shares the problem of needing

large parallel data and does not have the ability to generate new degrees of emotional states not contained in the database.

Former studies [14] [16] [17] [18] [19] [66] [67] considered converting neutral speech to simple categories of emotions such as joy, anger and sad. Tao tried to label the emotion database using four degrees "strong," "normal," "weak," "unlike" to each emotion category [15]. However, daily social emotions conveyed by humans are mild and not purely one emotion or another, but a mixture of emotions, e.g., anger and sad and fearful; they can be described as a continuum of nonextreme states [25] [68]. So synthetic speech with simple categories of emotions is not sufficient.

1.3 Research method

The research goal of this research is to propose one method to convert neutral speech to emotional types with varying degrees following human emotional speech perception mechanism. Rule-based emotional speech synthesis concept is applied in research as it obtains variation tendencies of acoustic features with a limited database.

In this research, the voice conversion system for emotional speech is built with a single speaker. In order to control the degree of emotion, the emotion dimension is adopted to express the emotional state as a point in dimensional space so the degree can be controlled by changing the position in the emotion dimension. This research mainly focuses on prosody-related feature conversion. In the emotion conversion system as shown in Fig.5.1, two inputs (intended position in dimensional space and neutral speech) and two steps (rule extraction and rule application) are necessary. In the first step, the rules between acoustic feature variations of neutral and emotional ones can be extracted using a fuzzy inference system. The inverse three-layered model is set as the structure between emotion dimension and acoustics with emotion dimension as the bottom layer, the semantic primitive layer at the middle and acoustic layer at the top. As the emotional experience is biologically based, these rules have the potential ability to be applied to arbitrary speakers or languages.

The second step is to apply the rule-based voice conversion method to modify the acoustic features of neutral speech to emotional ones following rules extracted from the first step. It is widely understood that emotion is conveyed by means of a number of

prosodic parameters such as voice quality and speech rate as well as fundamental frequency [4] [7] [15].

In this step, some essential prosody features such as duration, F0 contour and power envelope are parameterized by an interpolation method, Fujisaki model [69] [70] and target prediction model [71]. Then the modified acoustic features are synthesized using STRAIGHT [72], a VOCODER which can decompose speech signal into parameters so as to precisely control and modify them. Fig.5.1 will be explained in detail in Section 5.

It is found that human perception of emotions in speeches in different languages is identical in the dimensional space [88] [48]. Following this finding, we assume that, given the same direction in dimensional space, the system can convert the neutral voices in multiple languages to emotional ones with the same impression of emotion. It means that even if the conversion system constructed for one language, it can also work for other languages.

In order to explore the validity of this hypothesis, we utilize the emotional speech conversion system constructed for Japanese to convert other languages such as English and Chinese.

1.4 Research novelty

Firstly, conventional emotional speech synthesis or conversion system utilize the categorical approach to represent emotion which neglects the diversity of human emotion production. This research firstly applies the dimensional approach to represent emotion for the emotion conversion system so the degree of emotion can be controlled through moving the positions in dimension space.

Secondly, the modified brunswikian's lens model has been proposed many years which indicated that the process of human emotion perception and production is multiple processes. However the present speech synthesis system only considered the relationship between acoustic and emotion. And the process of human emotion perception mechanism is firstly considered in the speech conversion area by applying the inverse three-layered model.

Lastly, in this research, multiple languages speech can be converted by using one system without training which will make great convenient for global speech communication

by using the rule-based method. And this system can control the emotion category and degree freely without the limitation of the training database and training language which is totally different from the conventional emotional speech synthesis method. And this system will make a great progress in the emotion speech synthesis field.

1.5 Outline of thesis

This dissertation is constructed in seven chapters and is structured as follows.

Chapter 2 introduces the conceptual grounds of emotion dimension. Chapter 3 reviews the three-layered model as the construction between emotion dimension and acoustic features. Chapter 4 describes the listening tests done to obtain the relation between the acoustic features and each emotion dimension. In Chapter 5, the structure of the emotional voice conversion system is explained. Chapter 5.1 illustrates the extraction of the prosody rules for emotional voice conversion using a fuzzy inference system. The prosody conversion method is explained in Chapter 5.2, and Chapter 5.3 reports the perceptual evaluation of the resulting emotion conversion system. Chapter 6 tries to extend the system building in one language to multiple languages. Discussions on controlling spectral sequence and other affect factor are made in Chapter 7. Lastly, the conclusion is made in Chapter 8.

Figure 1.6 shows the organization of this dissertation.

Chapter 2 introduces the background knowledge of emotion dimension representation methods. Firstly, the most frequently used representation approach, categorical representation is introduced. The advantage and shortcoming of this approach are explained. Then the dimensional representation which can remedy the shortcoming of categorical approach is introduced. In this research, arousal and valence, two dimensions are used for representing emotion. Lastly, some other representation methods are also shown.

Chapter 3 gives an explanation of the Brunswik's lens model which is the theory foundation of the three-layered model. Then the concept of the three-layered model is introduced. Lastly I proposed the inverse three-layered model for emotion production which following the concept of Brunswik's lens model.

Chapter 4 explored the related acoustic features to each emotion dimension from the view of speech synthesize. Four important acoustic features, F0 contour, spectral se-

quence, time duration and power envelope are replaced separately from neutral speech to emotional one. Perception results in dimension space show that F0 and spectral sequence give significant attribution to valence and arousal axis. Time duration and power envelope make importance to arousal. This study considered controlling the prosody-related acoustic features firstly.

Chapter 5 shows the structure of the emotional voice conversion system. Two inputs, position in dimension space and the neutral speech, and two parts, estimation and modification parts consist of this system.

The acoustic differences between neutral and desired emotional speech are estimated by the estimation part through the fuzzy inference system. The the differences are applied to the intended converted neutral speech. Three acoustic features, F0, power envelope and time duration are controlled by Fujisaki model, target prediction model and interpolation method.

This system is built by Fujistu database which is a one voice actress speaker Japanese database. Perception results by native Japanese show that the joy, cold anger and sad emotion can be perceived well regarding the category and degree.

Chapter 6 extended the monolingual system to multiple languages. Previous research shows that human perception for different languages is identical in dimensional space. Directions from neutral voice to other emotional states are common among languages. So this chapter tried to directly change the input neutral speech from Japanese to other languages. Perception results show that all converted voices can convey the same impression as Japanese voices.

Chapter 7 summarizes this study and emphasizes its contributions to this research field as well as other research fields.

Furthermore, future works about deepening the controlling of other acoustic features and extending the emotion field to other affective factors will be discussed.

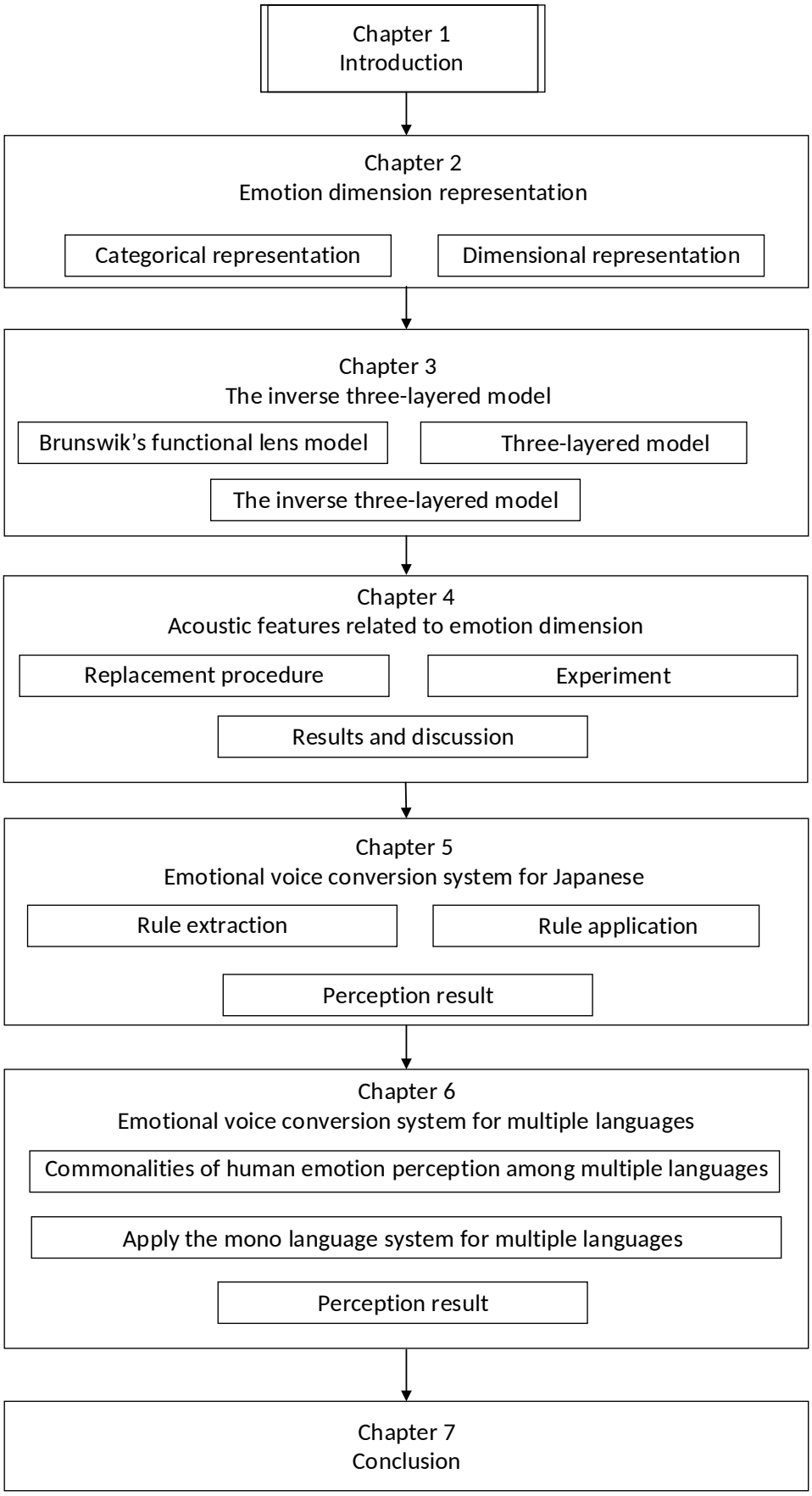

Figure 1.6: Organization of this dissertation.

Chapter 2

Descriptive frameworks for emotions

This chapter reviews the literature on frameworks for representing emotions. To synthesized emotional speech, a method related to how to represent emotions is firstly needed to be considered.

Many frameworks have been proposed already for representing emotion. Among them, categorical approach is the most common way while more and more researchers based their research on dimension representation for emotion [25] [28] [107] [108]. The categorical approach and some other less-well known methods, prototype descriptions [21] [106], appraisal-based descriptions [22] [109] [110], circumplex models [23], physiological descriptions [24] and dimensional approaches [28] [25] have already been applied to describe the emotional content. This section aims to explain these methods in detailed.

2.1 Categorical representation

The emotion category approach is the most straightforward method with simple emotion-denoting labels. It has been shown that emotion-denoting labels in human language are extremely powerful. It is reported that there are 107 emotion-denoting labels in English [73] and 235 in German [74].

However, it is difficult to apply all items when concentrating on emotion speech recognition or emotion speech synthesis. According to the research aim, some basic emotions or essential everyday emotion terms are selected. Recently, about 10 emotions are defined as the basic emotion such as happiness, sadness, anger, fear, disgust, pride, relaxed de-

pression and so on. The merit of emotion category representation is that it is the simplest and least costly method for both emotion recognition and emotion synthesis.

In the field of emotion synthesis, many previous researches have already attempted synthesizing affective speech with categorical emotion terms [7] [37]. However, many researchers [15] [25] [75] argued that discrete category representation ignores the diverse and fuzzy peculiarity of emotion and sometimes it is difficult to define a clear-cut boundary among the non-overlapping categories. Therefore, the complexity of emotional states may not be reflected well by categorical representation.

2.2 Dimensional representation

Humans tend to produce emotion with different degrees of intensity which may change during the course of the speech communication act. Most HCIs require the machine to produce human-like non-extreme emotion. Therefore, in order to build intelligent HCIs, a representation needs to satisfy the requirement that it can express mild emotions rather than full-blown ones.

The dimensional representation method which represents emotion as a point in a multi-dimensional space can scale the emotional intensity from low intensity to high intensity in a continuous way. Despite specifying emotion as an individual emotion category, dimensions used in this representation are gradual in nature and show the essential aspects of emotion concepts. Therefore, in this research, the dimensional approach is adopted as the descriptive framework which will be explained in more detail.

The history of dimensional representation can be tracked back to Wilhelm Wundt [91] who believed that the direction of feeling can be represented by two opposite terms and there are mainly three directions of the diverse feeling: pleasure and displeasure, exciting and depressing, tensing and relax. Wilhelm Wundt's proposal did not seem to be based on experimental results.

Then Harold Schlosberg proposed the based dimensions of emotion by experimental psychology [92] and came to a conclusion that it is the Pleasantness-Unpleasantness dimension that makes the great contribution to distinguish the categories. The second direction he proposed is Attention-Rejection.

Albert Meharbian and James Russell [93] firstly gave the evidence that emotions can

be represented using three dimensions through the intermodality response, synesthesia, physiological reaction and semantic differential. The names of the three dimensions are pleasure, arousal, and dominance.

Recently, Roddy Cowie et al. [94] used two dimensions to describe emotions. They interpreted emotions by a two-dimensional circular space. The two dimensions were "evaluation (from negative to positive)" and "activation (from passive to active)."

Through a variety of different methods such as semantic differential ratings and multidimensional scaling, three dimensions [50] (how active or calm, how positive or negative, how powerful or weak) are commonly utilized among researchers. The names of the three dimensions in literature have many versions (eg., pleasure, arousal, and dominance; evaluation, activity, and potency; and evaluation, activation, and power). In this paper, two dimensions, as shown in Fig.2.1, arousal (synonymous to activation and activity) and valence (synonymous to evaluation and pleasure) are used for representing emotions based on the database we have. In the valence-arousal (V-A) representation as shown in Fig.2.1, joy is positive and excited while sadness is negative and calm; thus, the position values of joy are all positive and the position values of sadness are all negative in V-A space. On the other hand, anger which is negative but excited shares the negative valence but positive arousal. According to the value of valence and arousal, anger can be divided into hot and cold anger. In psychology, hot anger corresponds to the prototypical full-blown anger emotion; milder and more subtle forms of anger expression exist, including cold anger [76].

2.3 Other descriptions

Except categorical and dimensional representation, there are also some other less-well-known methods such as prototype descriptions [21], appraisal-based descriptions [22], circumplex models [23], physiological descriptions [24]. This section will make a brief introduction to these methods.

Prototype-based [21] [95] description holds the concept that it is not easy to make a clear-cut boundary between the emotion categories. Therefore, instead of making criteria for defining emotions, prototype descriptions use the membership in an emotion group based on the similarity according to the corresponding emotion prototype. In that case,

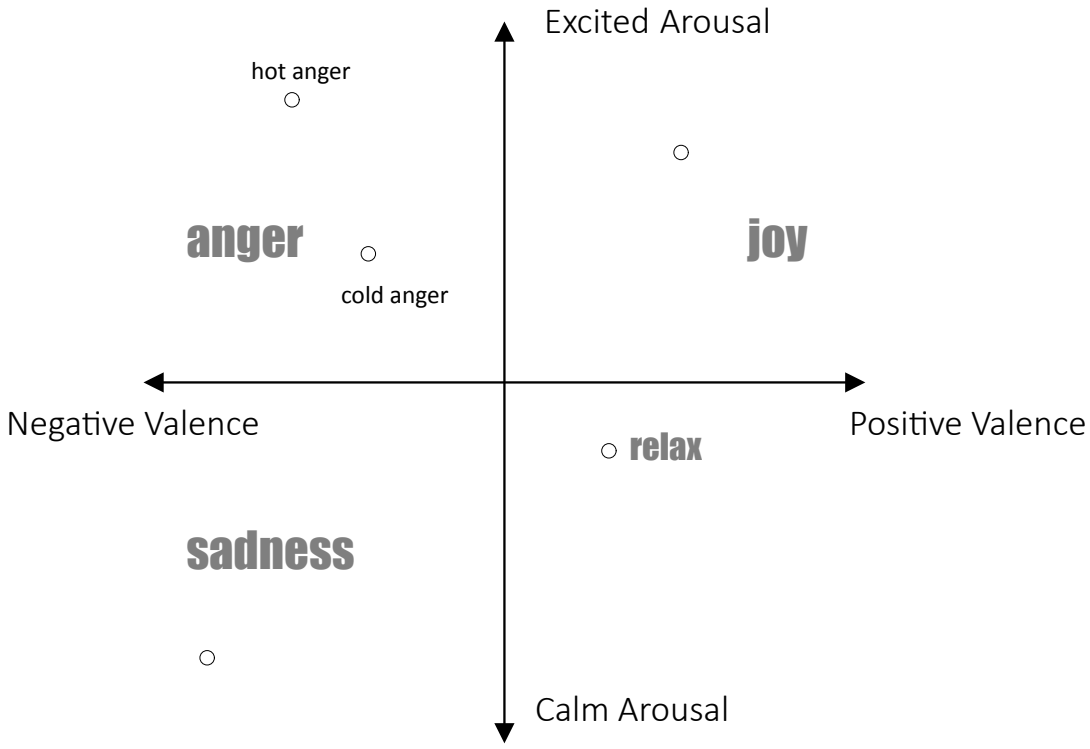

Figure 2.1: *Dimensional representation.*

an emotional state can be represented as a member of several emotion classes to different degrees. Until now, in the field of emotion and speech, there is no research-based their concept on prototype-based descriptions.

In the field of emotion-cognition, some researchers represent emotions using appraisals [22]. One emotion involves some stimulus evaluation checks and will be triggered when something is perceived as an importance.

Circumplex models can represent emotions by a circular structure will have been proved by several researchers [23], [97] as shown in Fig. 2.2. The major advantage of circumplex models is that the similarity and difference between emotion categories can be explicitly shown by the distance in the circumplex models.

The essential part of an emotion is the state of the body which can be peripheral physiologically measured through skin conductance and heart rate [96]. Some researchers established the correlation between the physiological measures and vocal emotions which is called the physiology-based descriptions [24]. Scherer conducted some experiments to predict the vocal emotions through the physiological changes.

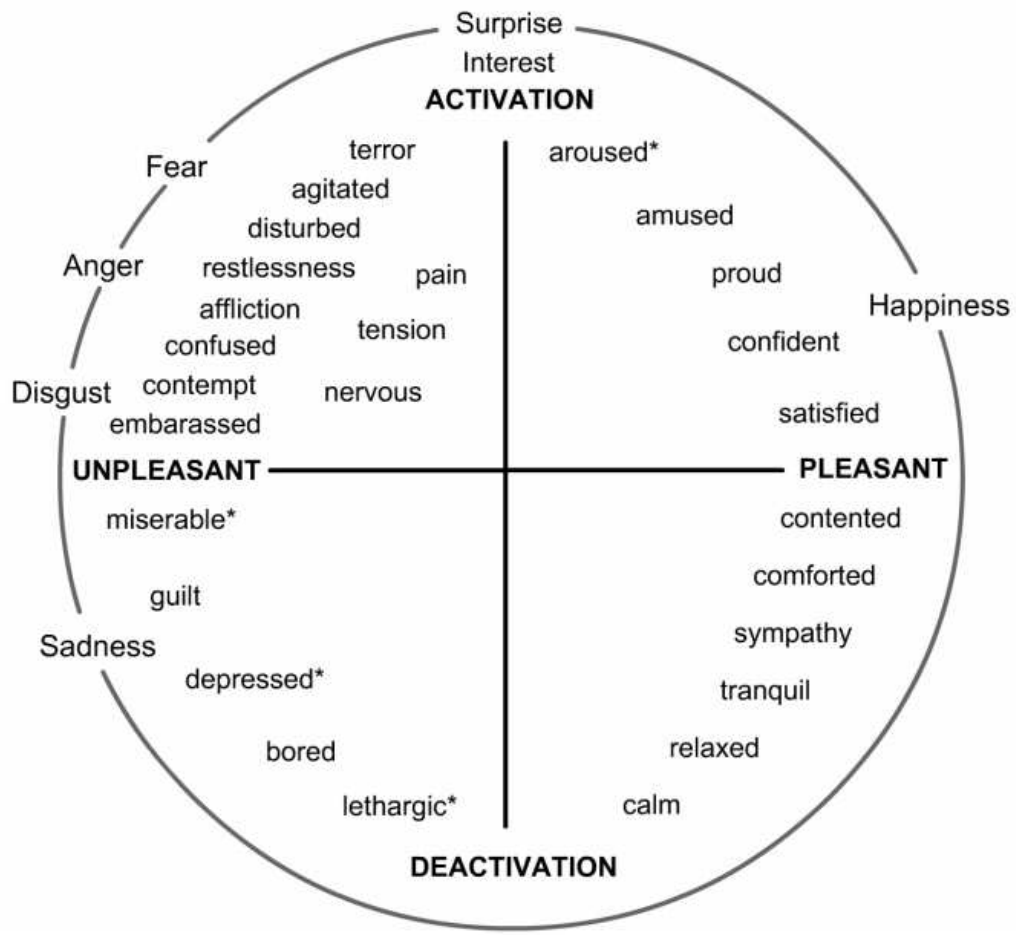

Figure 2.2: *Russels circumplex model of emotion [23].*

Chapter 3

Inverse three-layered model

Another problem addressed in this research is that the voice conversion system for emotional speech needs to follow human emotion perception and production mechanism. This section discusses methods to model the vocal communication of emotion and to apply the model to a voice conversion system.

3.1 Modified Brunswik's functional lens model for emotion perception

Several orientations have been proposed in investigating speech and emotion, such as the speaker-centered studies, listener-centered studies and the Brunswikian lens model.

The purpose of speaker-centered studies is to find the relationship between the emotional state of the certain speaker and the parameters related to speech. This orientation is widely applied to the emotion recognition system in which the computer needs to decide the emotional state of the speaker.

In the listener-centered study, the essential task is to explore the emotional meaning from the speech cues which is applied to the emotional speech synthesis field. In this field, the speech variables are modeled in order to convey the certain emotional state.

Klaus Scherer reported a graphical representation [98] that one person infers another person's emotional attributes from the external speech makers extended from the Brunswik's lens model [30]. As the three-layered model and the inverse three-layered model base their research on the Brunswik's lens model, we will explain this in detail.

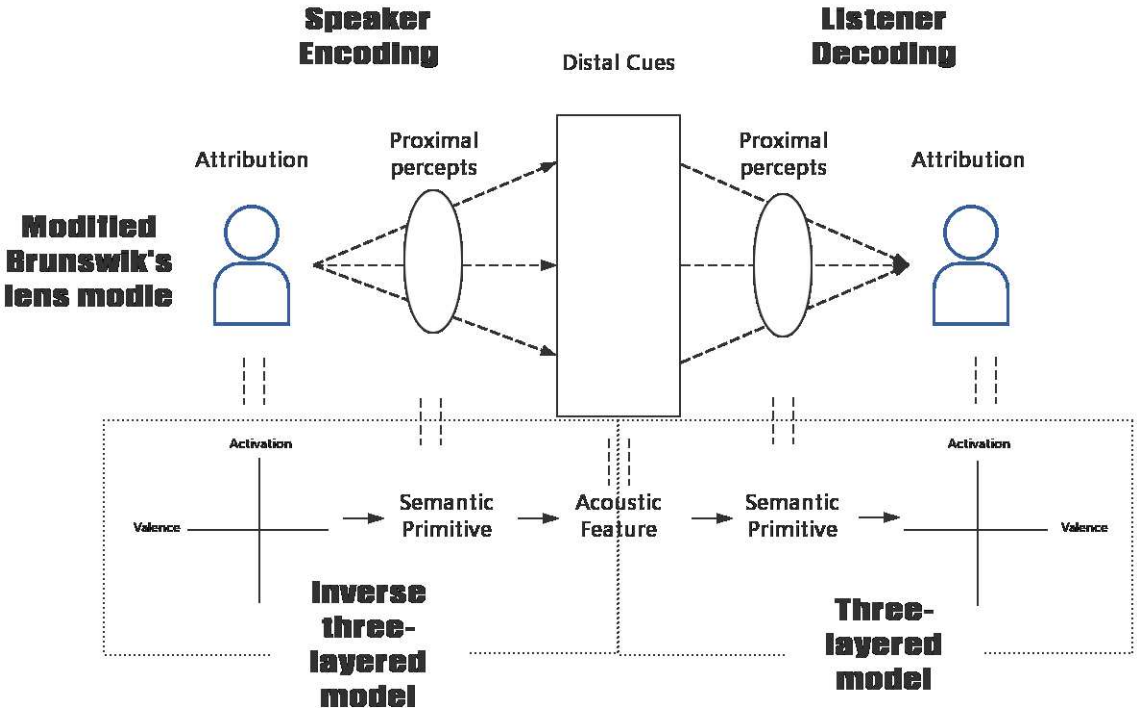

Figure 3.1: *The relationship among Brunswik's lens model, three-layered model and the inverse three-layered model.*

The Brunswik's lens model was originally presented in [99] which was used in several fields to study how observers correctly and incorrectly use objective cues to perceive physical or social reality. Hammond immediately applied this model to judgment analysis such as decision making, medical diagnosis, weather forecast and so on [100]. Then a modified Brunswik lens model and behavior mapping were used to examine the encoding and decoding of interpersonal dispositions from nonverbal cues [101]. Reynolds and Gifford explored the intelligence judgement and measurement [102] using the lens model. Juslin [77] applied this model in emotion field but only related to the emotion in music. It is Scherer who firstly introduced modified versions of the lens model applied to emotion [98].

In the modified Brunswik's functional lens model as shown in Fig. 1.2 and 3.1, emotion is encoded by means of a number of objective cues, called "distal indicator cues". In the area of speech and emotion, distal indicator cues in principle are related to the objectively measured acoustic features. A listener perceives the distal cues through the transmission channel which are internally viewed as "proximal percepts" in the first perceptual inference process. The listener uses the percepts for "attribution" to judge the speaker's state.

According to the Brunswik's lens model, we can see that the perception of emotion is not directly from "distal indicator cues", that is, acoustic to "attribution" emotion, but includes a middle procedure "proximal percepts". This means that the procedure of human emotion perception is a multiple-layered process.

3.2 Three-layered model for emotion perception

Scherer [29] [30] has highlighted several times to base theory and research in vocal emotion area on the modified Brunswik's functional lens model which illustrates the procedure of the emotion encoding by speaker and the emotion decoding by listeners.

Based on the Brunswik's lens model, Huang and Akagi [34] proposed a three-layered model for emotional speech perception who believed that the perception of human is vague. They hypothesize that humans perceive emotion not directly from acoustic features but from some descriptors where each descriptor is an adjective for describing the perceived characteristics of the speaker's voice. The combination of descriptors accounts for the decision about which emotion the speech belongs to.

In this model, emotion category is at the top layer, semantic primitives constitute the middle layer and at the bottom is the acoustic feature layer as shown in Fig.3.1. Acoustic features refer to the acoustic parameters of voice, e.g., F0, power, duration, and semantic primitive refers to the listener's label of the voice such as bright, fast or hard. The acoustic feature, semantic primitive and emotion category in the three-layered model corresponds to the "distal indicators cues", "proximal percepts" and "attribution" respectively in the Brunswik's lens model.

The relationships of the three-layered model were constructed in a topdown way. The relationship between emotion and semantic primitives was built by conducting three experiments using multidimensional scaling and applying the fuzzy inference system to the experimental results. Fuzzy inference is based on the natural language which fits well with semantic primitives.

The relationship between semantic primitives and acoustic features was built by analyzing acoustic features measured from the F0 contour, power envelope, spectral sequence, and time duration.

Then from the bottomup way, the rule-based speech morphing technique was applied

for verifying the model based on two types of experimentally-derived rules. Successful results validate the semantic primitives chosen by successfully morph emotional speech utterances.

Some researchers have utilized this model in the field of emotion recognition [48] [36]. The top layer is modified from an emotion category to an emotion dimension since human beings have the ability to perceive gradual and continuous emotion degrees, not only categorical. They found that applying a three-layered model achieves a better emotion recognition rate compared with a two-layered model with no semantic primitive layer.

3.3 Inverse three-layered model for emotion production

Speech production and speech perception are important components of the speech chain as shown in Fig. 3.2. Speech starts with a thought and intent to communicate in the brain, which activates muscular movements to produce speech sounds. A listener receives it in the auditory system, processing it for conversion to neurological signals the brain can understand. The speaker continuously monitors and controls the vocal organs by receiving his or her own speech as feedback. The speech production process starts with the semantic message in a person's mind to be transmitted to the listener via speech. The speech understanding process works actually in the reverse order.

On the other hand, Biersack and Kempe [78] explored whether vocal cues can be used to reliably infer speaker happiness. Subjects were asked to perform a simple referential communication task and to rate their current emotional state. A range of vocal cues was traced through the speech chain using path analysis. The results indicate that reported happiness of the speakers and perceived happiness of the listeners were not related. The only vocal cue that mediated between reported and perceived happiness was F1, and, for the female speakers, pitch range. In sum, they found a weak relationship between vocal cues of happiness encoded by speakers, and vocal cues used for decoding of this emotion by listeners. This supports the view that vocal cues are not just universal epiphenomena of the emotional state of the speaker [2]. This suggests that some vocal cues can mediate between experienced and perceived emotions. At the very least, their results cast doubt

on the assumption of a direct mapping of vocal cues between perception and production in the domain of emotional expression. This prompt us to consider to build a multiple process to mapping the vocal cues to emotions which support the Brunswik's functional lens model and the three-layered model.

Both Brunswik's functional lens model and the three-layered model are used to account for human emotion perception. According to Juslin who also uses the Brunswik's lens model in [77], two important conclusions can be made: firstly, speakers can communicate emotions successfully to listeners, and secondly, the cue utilization of speakers maps well to the cue utilization of listeners. This indicates that speakers and listeners share the same representation methodology (i.e., coding method) when doing vocal communication. According to this result, we assume that human production of emotion is the mirror effect of human perception of emotions which means the encoding process of the speaker is the inverse process of the decoding process of the listener.

Based on this assumption, the inverse three-layered model is applied to the structure of the voice conversion system for emotional speech. We assume that in order to express the "attribution", i.e., the emotion intended by the speakers, speakers firstly encode the attribution by means of a number of "proximal percepts", that is, semantic primitives. Then the "proximal percepts", are externally expressed by the "distal indicator cues", that is, acoustic features.

In the inverse three-layered model as shown in Fig. 3.3, at the top layer is the acoustic feature layer, the middle layer, the semantic primitives and the bottom layer, the emotion dimension representation. Valence and arousal consist of the emotion dimension layer. In semantic primitives layers, seventeen semantic primitives which were selected by three psychoacoustical experiments [34]. These were bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow.

Totally sixteen acoustic features in the inverse three-layered model are used which can be controlled in the rules application step. These are F0 related features: F0 mean value, highest F0, a rising slope of the F0 contour and rising slope of the F0 contour for the first accentual phrase; spectrum related features: first formant frequency, second formant frequency , and third formant frequency, spectral tilt and spectral balance; power envelope related features: power range, rising slope of the power for the first accentual

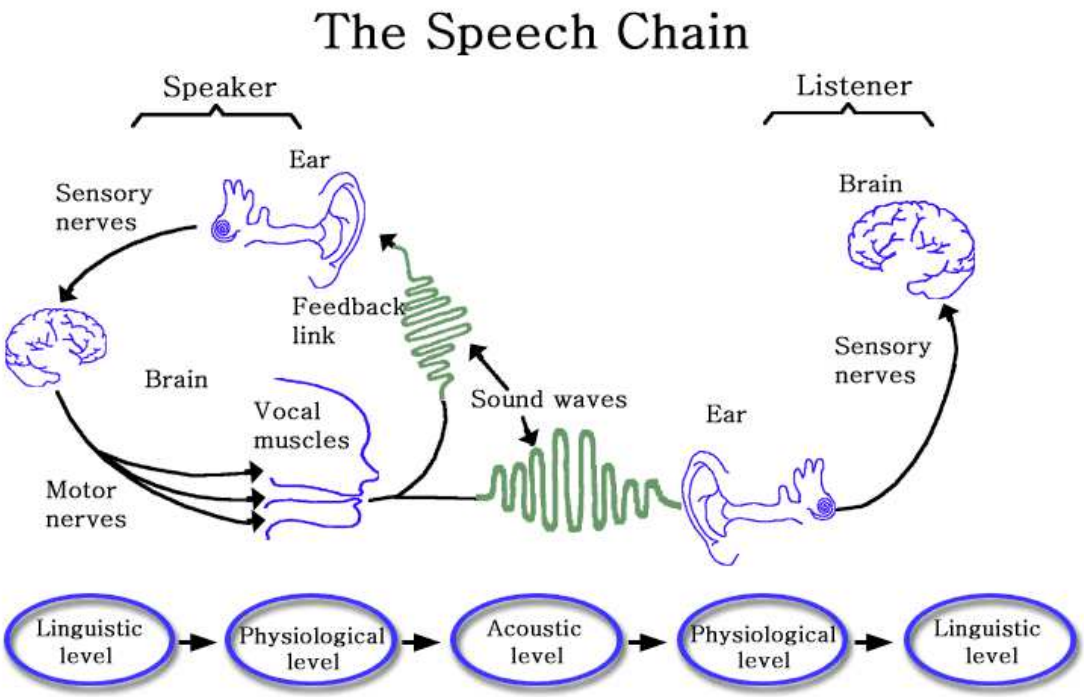

Figure 3.2: *The speech chain (speech production and speech perception) [79].*

phrase, the ratio between the average power in high-frequency portion (over 3 kHz), the average power and the mean value of power range in accentual phrase, were measured; and duration related features: total length, consonant length, the ratio between consonant length and vowel length were considered related to duration.

The inverse three-layered model is firstly proposed as the structure for extracting the rules for modifying the acoustic features of neutral speech. As the relationships among three layers are not simply the linear relationship, the inverse three-layered model is not simply changing the direction of three layers. The relationships need to be constructed again. And the fuzzy inference system is built as the connector considering the vague perception of emotion. This will be explained in Chapter 5 in detail.

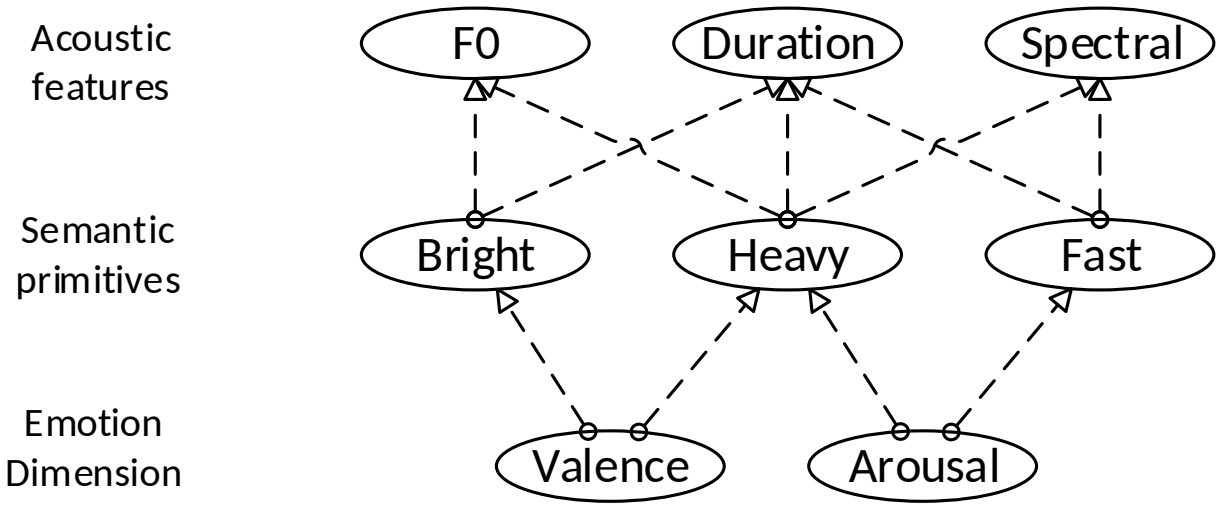

Figure 3.3: *The proposed inverse three-layered model.*

Chapter 4

Acoustic features related to emotion dimensions

For speech synthesis with different emotional styles in the V-A dimensional space, the related acoustic features to each dimension are explored in this section.

Most previous methods concentrated on related acoustic features within an emotion category [34] [80]. Previous methods such as Schröder [25] [81], applied statistical analysis such as correlation and linear regression analyses to dimension space. According to these results, almost all acoustic variables correlated with the arousal axis. Correlations with the valence axis are less numerous as well as less strong. This leads to confusion when synthesizing the speaking styles related to the valence axis. Statistical methods may make a great contribution to emotion recognition because a combination of acoustic variation may lead to one kind of emotion. However, for emotional voice conversion, even if we modify some acoustic features according to the statistically-derived rules, such as duration which show great differences between emotional and neutral speech, the synthesized speech still is not perceived as a targeted (categorical) emotion.

This section investigates the acoustic features related to each dimension as applied to emotional speech synthesis. Subjects were asked to evaluate the synthesized speech, the specific acoustic features of which, such as F0, have been replaced by the F0 contour from the emotional speech but leaving the other acoustic features of the neutral speech. The idea is that if changing only the F0 contour results in the synthesized speech being rated as similar to the original emotional speech in the arousal dimension, then this means that

the F0 contour makes a great contribution to the arousal axes. If this kind of changing makes results similar to the original neutral speech, this means that F0 contour is not related much to the arousal axes. In this paper, four types of acoustic features relating to emotion are explored: duration, F0 contour, spectral sequence and power envelope.

4.1 Acoustic features replacement procedure

Source and spectral parameters can be extracted flexibly by using the analysis/synthesis method STRAIGHT [72]. Successive refinements on the extraction procedure of source and spectral parameters enable the total system to re-synthesize high-quality speech. The literature on vocal correlates of emotion dimensions, especially with respect to speaking styles, reports the importance of prosodic parameters, such as F0 contour, spectral sequences and power envelopes [15] [25].

In order to determine which particular acoustic features of the emotional speech can be used to convert the neutral speech to emotional speech, it is necessary to keep the linguistic content constant. Thus, our research examined nine sentences with the same linguistic information but different speaking styles/ emotions. These sentences were chosen from the Fujitsu database recorded in the Fujitsu Laboratory by one professional voice actress. One of the 9 sentences is in the neutral speaking style without emotion; the remaining emotions are sadness, joy, hot anger and cold anger, with 2 utterances for each emotion type.

The procedure for replacement of F0 contours shown in Fig.4.1 is followed. Time information was first modified to keep the speech duration of the neutral and emotional speech constant; this needs to be done before modifying the F0 contour, spectral sequence and power envelope. Time modification was done first by manually segmenting the speech signal at the phoneme level for both neutral and emotional speech; then the time duration of the neutral speech is modified to that of the emotional speech, according to the ratio of the time duration of the neutral and emotional speech. Applying STRAIGHT, the first synthesized speech (neutral speech 2) can be obtained by changing only the time duration to match that of the emotional speech. Then, the F0 contour, spectral sequence, and aperiodic component (Ap) of the neutral speech 2 are extracted using STRAIGHT. At the same time, from the emotional speech, the F0 contour and spectral sequence

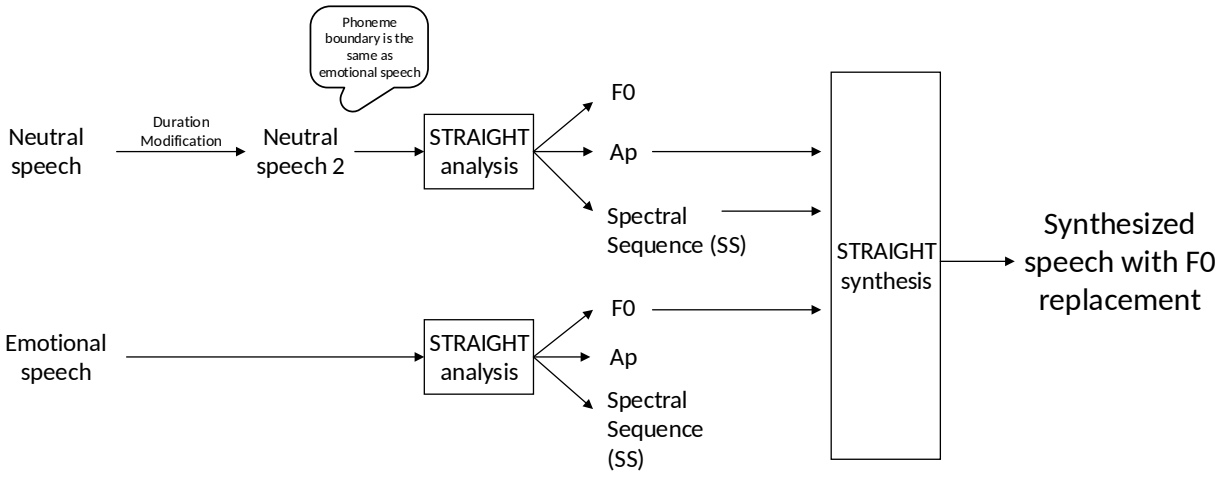

Figure 4.1: *Procedure of acoustic feature replacement.*

are also extracted using STRAIGHT. Since the time duration of the neutral speech 2 is the same as that of the emotional speech, the F0 contour of the neutral speech can be directly replaced by that from the emotional speech. The Ap and spectral sequence from neutral speech 2 and the F0 from the emotional speech are combined to be synthesized by STRAIGHT. The synthesized speech with F0 replacement is obtained lastly. By doing this, the spectral sequence and Ap information are kept, but the F0 contour is changed from neutral to emotional.

Fig.4.1 shows the procedure for replacing of F0 contour. For replacing the spectral sequence, the previous step is the same as F0 replacement. But in the last step, we use the F0 and Ap from the neutral speech, so that the spectral sequence from the emotional speech can be synthesized. This means the spectral sequence from neutral to emotional speech has been changed, but the other information is kept. For power envelope calculation, a Hilbert transform and low-pass filter are used. Synthesized speech with a different power envelope can be obtained by applying the power envelope of the emotional speech to neutral speech 2.

4.2 Experiment

The F0 contour, spectral sequence, power envelope and time duration of the emotional speech are moved one by one to the neutral speech. We obtained 32 samples of synthesized speech (8 utterances with the same linguistic information but different speaking styles,

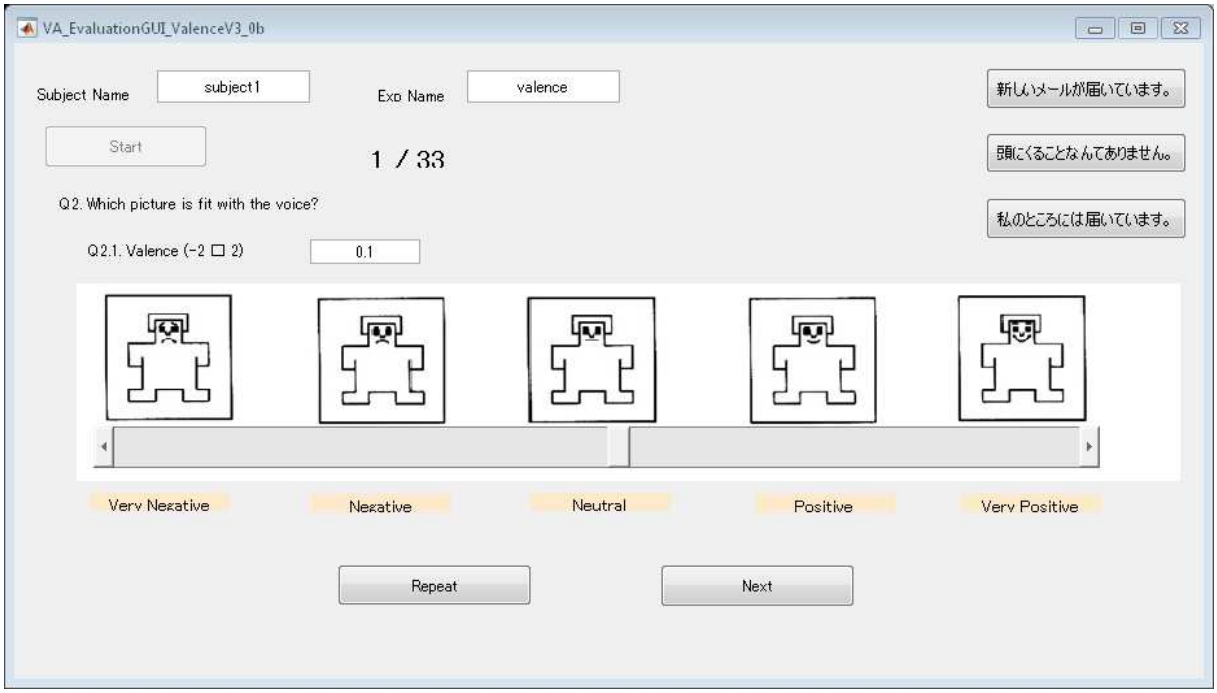

Figure 4.2: Graphic interface of the perceptual test.

4 types of acoustic features); plus 9 original utterances. Totally, there were 41 stimuli in the perception test in order to explore the influence of each acoustic feature on each emotion dimension.

In the listening test, twelve Japanese subjects with normal hearing ability were asked to evaluate the utterances in the V-A space. The stimuli were presented in an individually randomized order per subject over high-quality headphones (type HDA200, SENNHEISER).

Experiments for valence and arousal were done twice, for each dimension for a total of 4 tests. The first time served as a training test to allow the subjects to acquire an impression of all the stimuli. Valence and arousal were evaluated from -2 to 2 with a step of 0.1 (Valence: -2 [Very Negative], -1 [Negative], 0 [Neutral], 1 [Positive], 2 [Very Positive]; Arousal: -2 [Very Calm], -1 [Calm], 0 [Neutral], 1 [Excited], 2 [Very Excited]). Subjects evaluated these scales using a graphic-user interface as shown in Fig.4.2. During the listening test, subjects were allowed to listen to the stimulus as many times as they wanted.

Table 4.1: *Anova values of each acoustic features to valence and arousal(*$^{**}p < 0.01$,$^{*}p < 0.05$*).*

p-value	F0	SS	PW	TM
Valence	**	**	0.53	0.13
Arousal	**	**	0.03*	0.01*

4.3 Results

The correlation coefficients between subjects are calculated and the average results above 0.7 are chosen for the final analysis. Totally, there were 12 subjects who attended this experiment, but ten subjects were considered for the final analysis. In order to explore the influence of each acoustic feature on each emotion dimension, we assessed the original positions of the original emotional speech. The hollow points in Figs. 5.8, 4.4, 4.5 and 4.6 show the perceptual position values in V-A space of the original utterances. The neutral speech is almost at the center point which indicates neither positive nor negative, neither active nor calm. The values of joy are in the first quadrant which means positive and active. Hot anger is in the second quadrant which represents negative but active, and cold anger is in the second quadrants although the value of valence and arousal is lower than that of hot anger. For sad emotion, all points are in the third quadrant, which is negative and calm. These findings seem intuitively reasonable, which suggest that our subjects were able to understand the basic meaning of valence and arousal.

In order to investigate the influence of the emotion dimension, the four kinds of acoustic features are replaced separately; thus, the results of the listening tests for the synthesized speech are analyzed in terms of three aspects. Fig.5.8 shows the results when only the F0 contour is changed to the F0 contours of the other emotion categories but keeping the other acoustic information such as spectral sequence and power envelope. Figs 4.4, 4.5 and 4.6 show the results when only time duration, power envelope and spectral sequence are changed to those of the other emotions while holding the remaining acoustic values.

4.4 Discussion

Figs.4.4 and 4.5 show that by replacing the duration information and power envelope, the synthesized speech is still concentrated at the center point; this means that only

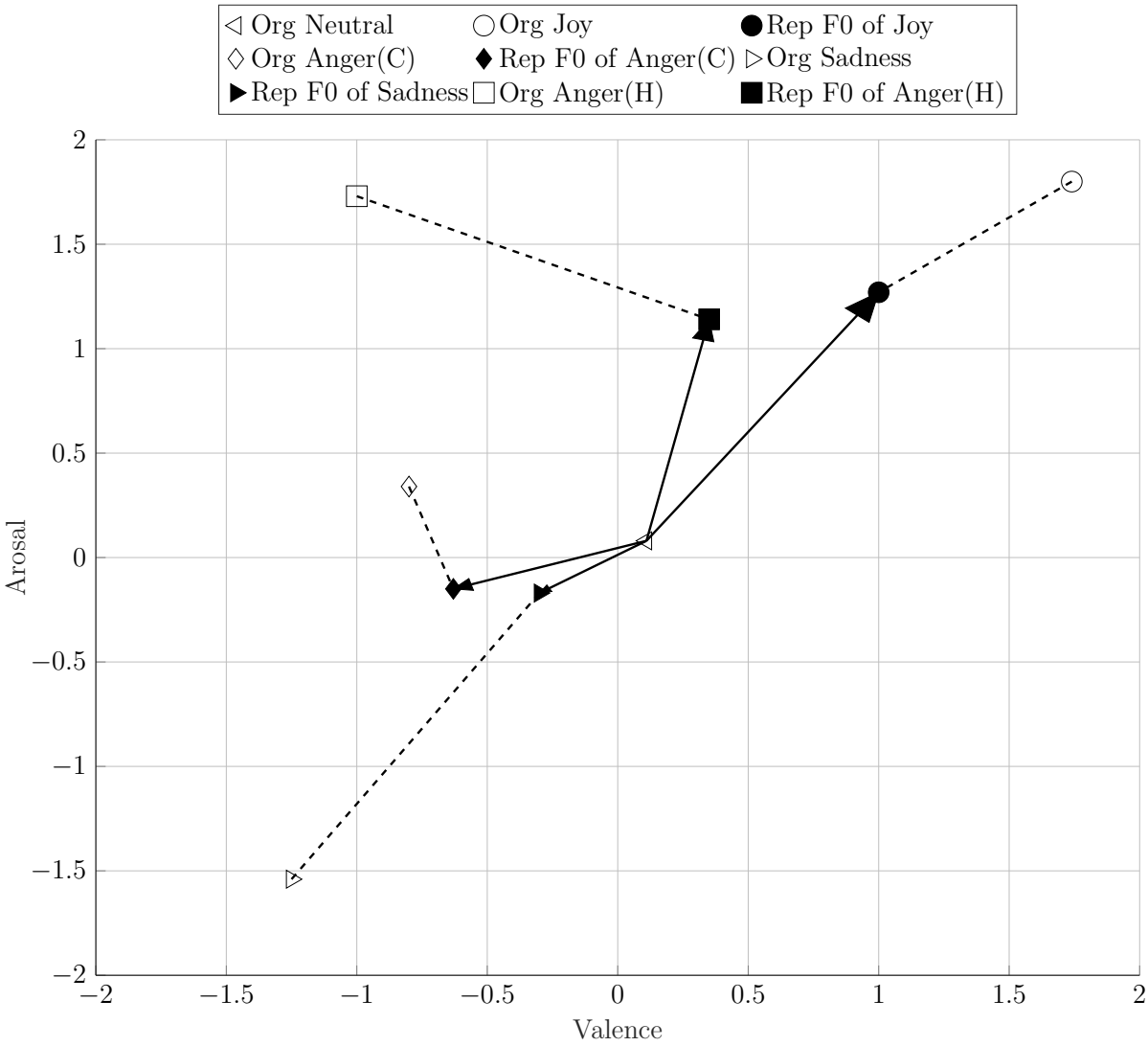

Figure 4.3: *Perceptual position values of original (Org) emotional utterances and synthesized (Rep) utterances on V-A space when F0 contour (F0) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)*

modifying the duration or power envelope does not very much change the expressiveness of a neutral utterance. Comparing the results of the original with the replaced ones, shown in Fig.5.8, we see that if only the F0 contour of the neutral speech is replaced by the F0 contours of joy and hot anger speech, the synthesized speech samples are all evaluated as joyful speech, as the evaluated position values are in the first quadrant. This is an interesting finding because most previous research proposes that F0-related acoustic features contribute greatly to the emotions of joy and anger. To a certain extent, this is true. But when converting neutral speech to emotional speech, if only F0 information is modified, it is possible to synthesize joyful speech but not angry speech. For sad speech,

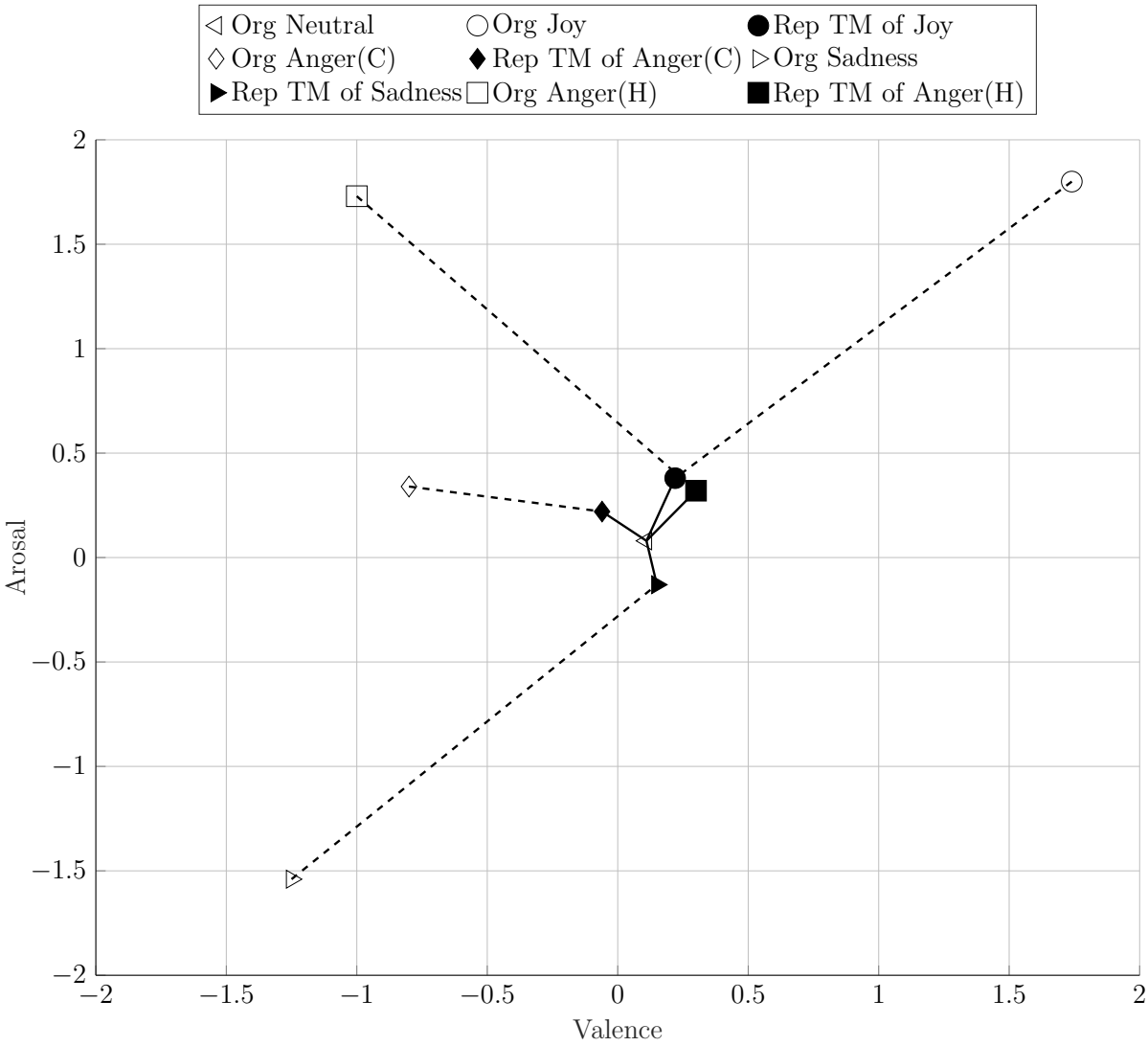

Figure 4.4: *Perceptual position values of original (Org) and synthesized (Rep) utterances on V-A space when time (TM) duration information is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)*

replacing the F0 of the neutral with that of sad results in the synthesized speech being in the third quadrant; this means sad speech can be synthesized by modifying only the F0-related acoustic features. However, the degrees of valence and arousal are reduced in joyful and sad emotions when only F0 is replaced. For cold anger emotion, by replacing only the F0, it is rated as slightly sad. Our findings show by replacing only F0, sad and joyful speech can be distinguished well, but notice that these emotions have inverse values in terms of both valence and arousal. However, joyful and angry speech cannot be differentiated; note that these differ only in the valence axis.

When replacing spectral sequences, as shown in Fig.4.6, synthesized speech was eval-

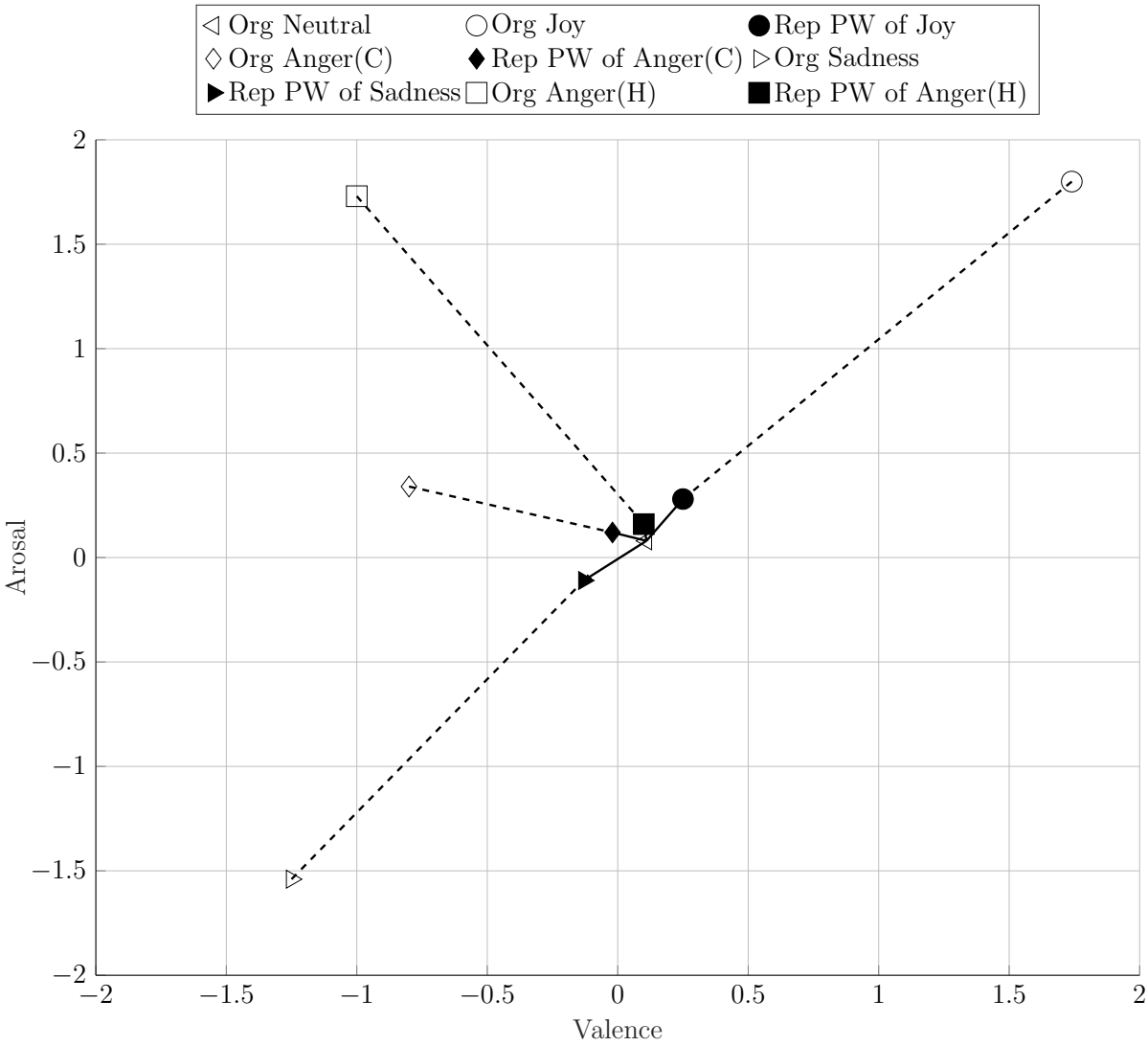

Figure 4.5: *Perceptual position values of original (Org) and synthesized (Rep) utterances on V-A space when power envelope (PW) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)*

uated as the original emotion, especially for hot and cold anger. This means that if there is a suitable method for modifying the spectral sequence of neutral speech, all emotions can be synthesized, although the degrees of valence and arousal are reduced compared to the original speech. What's more, for joyful and sad speech, by replacing only the F0, we can get closer to the original position in the V-A space than by replacing only the spectral sequence. This indicates that F0 is more related to the arousal axis than the valence axis. However, for the valence axis, the spectral sequence is more important.

The results of the ANOVA (analysis of variance) are shown in Table.4.1. From this table, we can see that F0 and spectral sequence have significant contributions to valence

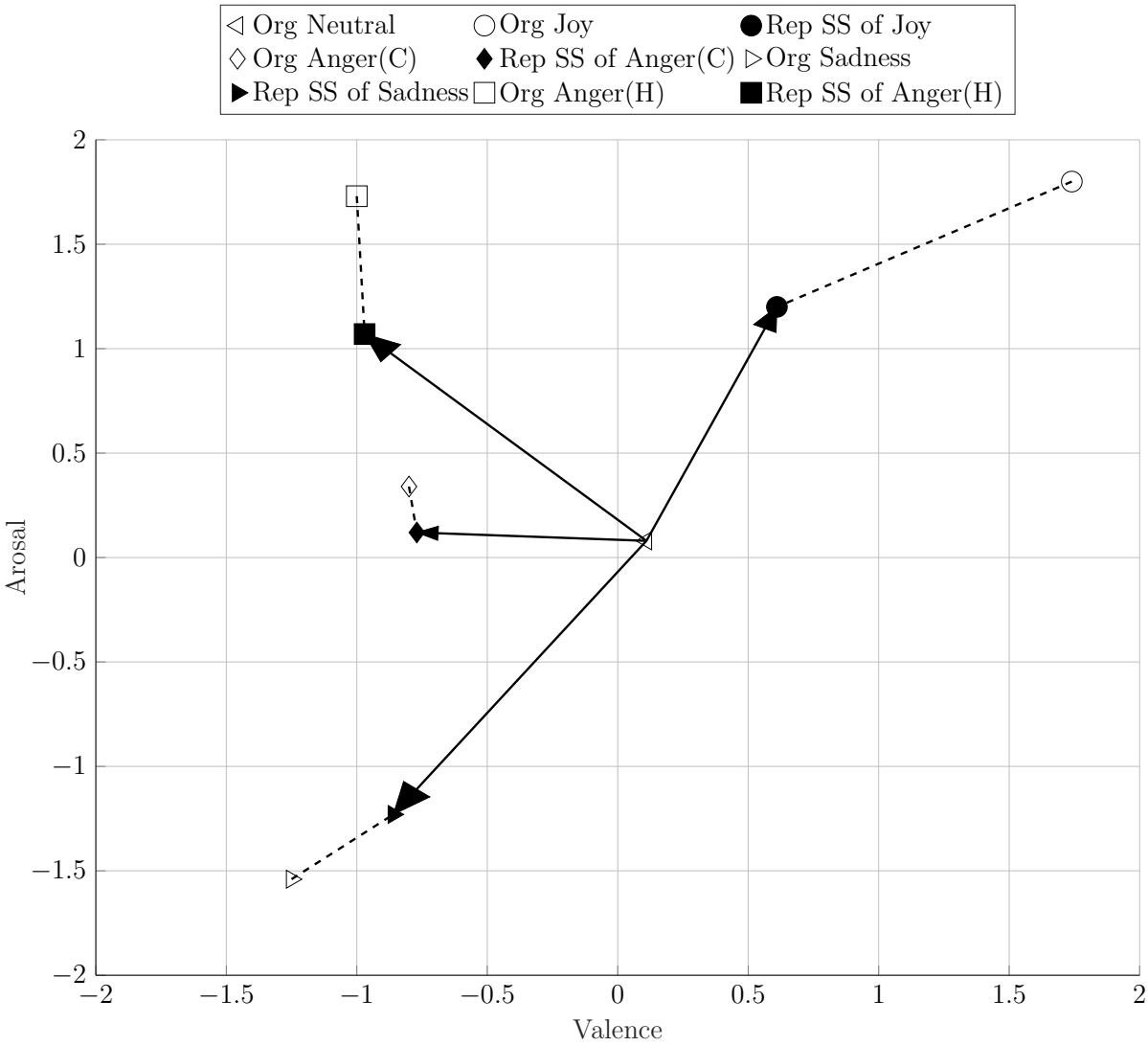

Figure 4.6: *Perceptual position values of original (Org) and synthesized (Rep) utterances in V-A space when spectral sequence (SS) is replaced from neutral to emotional speech. (Anger (C) means cold anger and anger (H) means hot anger.)*

and arousal axes ($p < 0.01$); power envelope and duration are much related to the arousal dimension ($p < 0.05$) but show no significance with the valence dimension ($p > 0.05$).

We conclude that both the F0 contour and spectral sequence are important to voice conversion for emotional speech. The power envelope and duration show little influence on the valence axes. In this paper, we focused on the prosody-related features such as duration, F0 and power envelope. The controlling of spectral sequence will be researched in the future. Since the utterances examined in this experiment are from a single speaker, and speakers have individuality differences when encoding emotion [86], future research is necessary to examine the commonalities among speakers for a better understanding of

synthesizing different speaking styles.

Chapter 5

The emotional voice conversion system

This chapter introduces the structure of the emotional voice conversion system for modifying the acoustic parameters of the neutral speech in order to convey the target emotion. Previous methods on emotion conversion systems mainly focused on applying a statistical approach, GMM, Deep neural network (DNN) or neural network (NN) [14] [16] [17] [18] [19]. As GMM often suffers from over-smoothing problems and the non-linear mapping such as DNN or NN need large databases for training suitable for categorical emotion representations. However, for a dimensional approach, it is difficult to collect a sufficiently large enough database with continuous emotional degrees. A rule-based strategy is used with a limited database in this paper to obtain tendencies of variation between emotion dimensions and semantic primitives, and then to extract rules between semantic primitives and acoustic features.

In the rule-based emotional voice conversion system, two-dimensional space, that is, valence (degree of negativity or positivity) and arousal (degree of calmness or excitedness), is used for representing the emotion; and the inverse three-layered model is used as the structure relating the acoustic features and emotion dimensions, as shown in Fig.5.1.

The emotional voice conversion system needs two inputs and two steps. Firstly, we need to know the position in the V-A space, which represents the desired emotion degree, and this step is referred to as the rule extraction step. It is this step which allows us to estimate the acoustic values of the desired emotion through the inverse three-layered model.

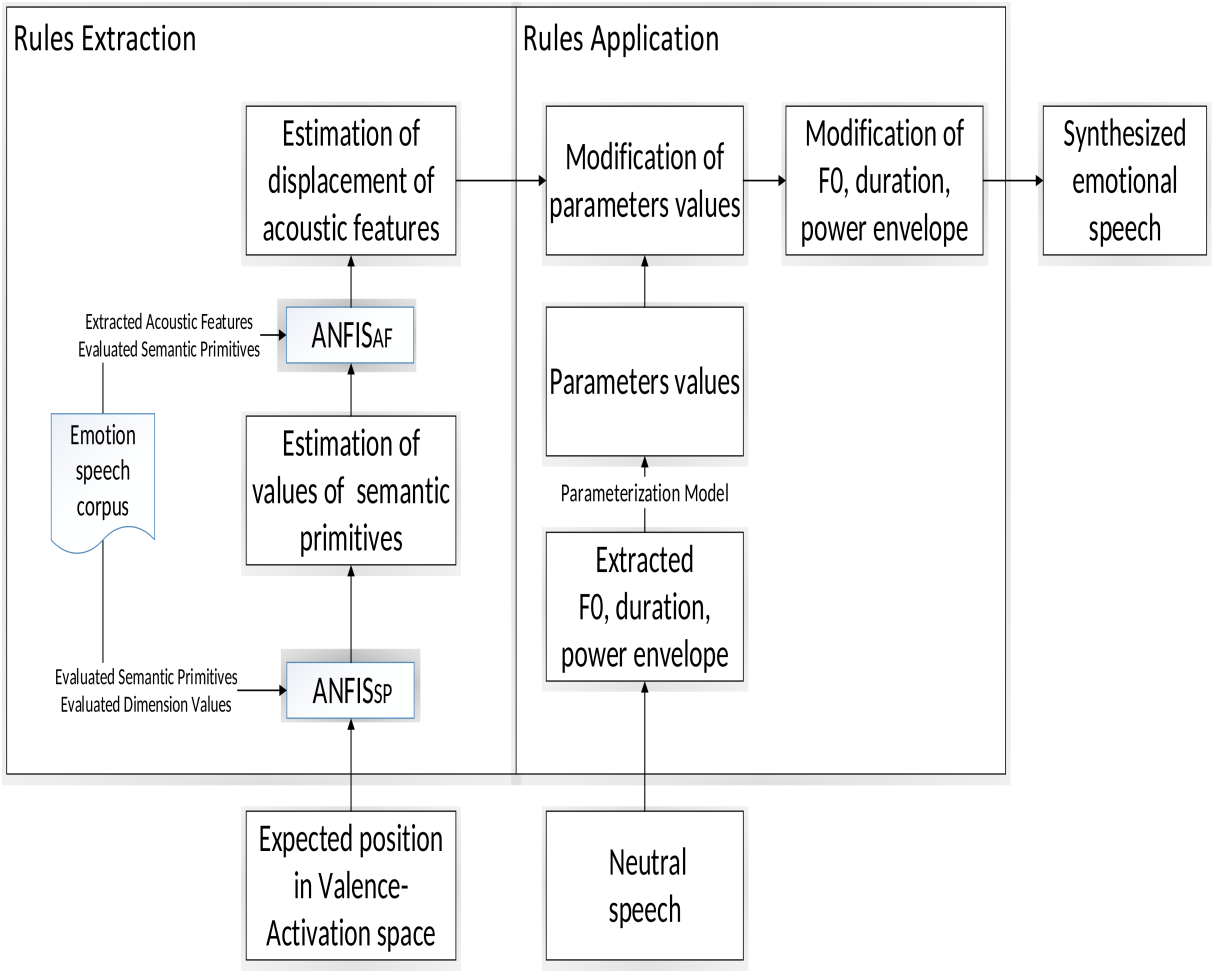

Figure 5.1: *Scheme of emotion conversion system.*

This then allows us to calculate the difference of acoustic features between the emotional and neutral speech. The details of this part along with the database is illustrated.

In the next step, the rule application step, the ratios of difference between the estimated acoustic features of the desired emotion and the acoustic features of neutral speech are applied to the extracted parameter values of the neutral speech. In order to modify the differences for the neutral speech, we concentrated on the prosody-related features, duration, F0 contour and power envelope. In order to control these features, the F0 contour and power envelope are parameterized using the Fujisaki model and target prediction model. After the modifications, applying the analysis/synthesis tool STRAIGHT to the modified acoustic features, the converted speeches with desired emotional degrees can be obtained.

5.1 Rule extraction

In this section, we illustrate how the inverse three-layered model is applied to the database to obtain the value of the various elements; how the fuzzy inference system connects the three layers to output rules relating the emotional dimensions to the semantic primitives; how the rules are extracted from semantic primitives to acoustic features and finally, how the effectiveness of the inverse three-layered model is evaluated by means of calculating mean absolute errors [82].

5.1.1 Database

We used the multi-emotional single speaker Japanese Fujitsu Database, recorded at Fujitsu Laboratories. A professional voice actress uttered 179 utterances in 5 speaking styles, joy, cold anger, hot anger, sad and neutral; 20 sentences spoken in 5 speaking styles, including one instance of neutral speech and two repetitions of each of the other speaking styles. One instance of cold anger is missing which makes a total of 179 sentences. Table 5.1 gives the translation version in English.

Each sentence has one for neutral and two for other emotional states. The total number of utterances is 179 because one cold anger utterance is missing from the database. Table **??** gives the categories of Japanese Database. It shows the id of each speech and each sentence has two versions for the emotional speech: Joy, Cold-Anger, Sadness and Hot anger. But for neutral speech, each sentence only has one version.

For Fujitsu Database, the detail of the speech data, the sampling frequency, the quantization, the number of sentences, the number of speaker and the number of utterances, are shown in Table 5.3.

5.1.2 Acoustic feature extraction

Except for duration-related features which are extracted by manual segmentation, the other acoustic features are obtained by the high-quality speech analysis-synthesis system STRAIGHT [72]. Based on the work by Huang and Akagi [34], 16 acoustic features are classified into the following subgroups.

Table 5.1: The content of sentences in English

id	Translation in English.
1	You've got a new mail.
2	There is nothing frustrating.
3	I heard that we would meet in Aoyama.
4	I brough a new car.
5	Please delete any unwanted e-mails.
6	That's an old superstition.
7	Many people sent cheers.
8	You should have reveived a letter.
9	I will think about you.
10	I have received it.
11	Thank you.
12	I am sorry.
13	I won't say thank you.
14	I'd like to travel just the two of us.
15	I felt like fainting.
16	There were our mistakes.
17	Do we need a straw mat to watch fireworks.
18	You said you would not do it again.
19	Please tell me why you don't come on time?
20	Meet me at the service area.

F0 related features: F0 mean value of average F0 (AP), highest F0 (HP), a rising slope of the F0 contour (RS) and rising slope of the F0 contour for the first accentual phrase (RS1st).

Spectrum related features: First formant frequency (F1), second formant frequency (F2), and third formant frequency (F3) were taken approximately at the midpoint of the vowels /a/, /e/, /i/, /o/, and /u/. The formant frequencies were calculated at an LPC-order of 12. Spectral tilt (SP_TL) was used to measure voice quality and was calculated using the following equation:

$$SP\_TL = A_1 - A_3 \tag{5.1}$$

where A_1 is the level in dB of the first formant, and A_3 is the level of the harmonic whose frequency is closest to the third formant. To describe acoustic consonant reduction, spectral balance (SP_SB) is adopted. It was calculated in accordance with the following

Table 5.2: The id number and the according expressive speech

UID	Expressive speech category
a001∼a020	Neutral
b001∼b020	Joy (1)
c001∼c020	Joy (2)
d001∼d020	Cold-Anger (1)
e001∼e020	Cold-Anger (2)
f001∼f020	Sadness (1)
g001∼g020	Sadness (2)
h001∼h020	Hot-Anger (1)
i001∼i020	Hot-Anger (2)

Table 5.3: Specification of speech data for Japanese database.

Item	Value
Sampling frequency	22050Hz
Quantization	16bit
Number of sentences	20 sentence
Number of emotion categories	5 category
Number of speakers	1 female speaker
Number of utterances	179 utterance

Table 5.4: *Acoustic features used in this system*

Group	Acoustic features
F0	1. average value of F0 (AP); 2. highest F0 (HP); 3. mean value of F0 in the rising slope (RS); 4. rising slope of the first accentual phrase (RS1_st);
Spectrum	5. 1st formant frequency (F1); 6. 2nd formant frequency (F2); 7. 3rd formant frequency (F3); 8. spectral tile (SP_TL); 9. spectral balance (SP_SB);
Power envelope	10. power range (PW_R); 11. rising slope of the 1st accentual phrase (PW_RS1); 12. the ratio between the average in high frequency (over 3kHz)and the total average power (PW_RHT); 13. mean value in accentual phrase (PW_RAP);
Duration	14. total length (TL); 15.consonant length (CL); 16. ratio between consonant and vowel length (RCV);

equation:

$$SP\_SB = \frac{\sum f_i \cdot E_i}{\sum E_i} \tag{5.2}$$

where f_i is the frequency in Hz, and E_i is the spectral power as a function of the frequency.

Power envelope related features: Power range (PW_R), rising slope of the power for the first accentual phrase (PW_RS1), the ratio between the average power in high-frequency portion (over 3 kHz), the average power (PW_RHT) and the mean value of power range in accentual phrase (PW_RAP) were measured.

Duration related features: Total length (TL), consonant length (CL), the ratio between consonant length and vowel length (RCV) were considered related to duration.

All the acoustic features are used for building the inverse three-layered model in the rule extraction step. In the rule application step, the prosody related features such as F0, duration and power envelope are parameterized. The conversion of spectral sequence features will be performed in the future work.

5.1.3 Semantic primitives evaluation

Based on the work by Huang and Akagi [34], 17 semantic primitives were selected to describe the perception of emotional vocalization. The 17 semantic primitives are bright,

dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow. 11 Japanese subjects were asked to give subjective values on a five-point scale ("1-Does not feel so at all", "2-Seldom feels so", "3-Feels slightly so ", "4-Feels so", "5-Feels very much so") for each semantic primitive for each 179 utterances. For each semantic primitive, the inter-rater agreement is measured by pairwise Pearson's correlation between two subjects' ratings. All subjects showed from moderate to a high-level agreement.

5.1.4 Emotion dimensions evaluation

The evaluation of emotion dimension is divided into two parts: valence and arousal [48]. The 11 Japanese subjects rated the 179 utterances on a five-point scale {-2, -1, 0, 1, 2 }. Valence was from -2 (very negative) to +2 (very positive), and arousal was from -2 (very calm) to +2 (very excited). The correlation coefficient between subjects rating for valence is about 0.9 and for arousal, about 0.85, which means subjects showed a high inter-rater agreement.

5.1.5 Fuzzy inference system

The fuzzy inference system (FIS) based on fuzzy logic [149] is considered as the training system for obtaining the rules among three layers. The reasons are as following:

1. FIS applies If-Then rules to turn human knowledge into a mathematical model and the production of emotional speech is also from the knowledge of human beings.

2. FIS can deal with arbitrary complexity using the non-linear functions. The relationships between acoustic parameters and emotion are non-linear and complicated. This is the most important reason that FIS is considered as the connector among three layers.

3. FIS is based on the natural language which fits well with the semantic primitive in three-layered model.

Contrary to the conventional fuzzy logic system which does not have a learning ability, the adaptive neuro-fuzzy inference system (ANFIS) combines the merit of fuzzy inference

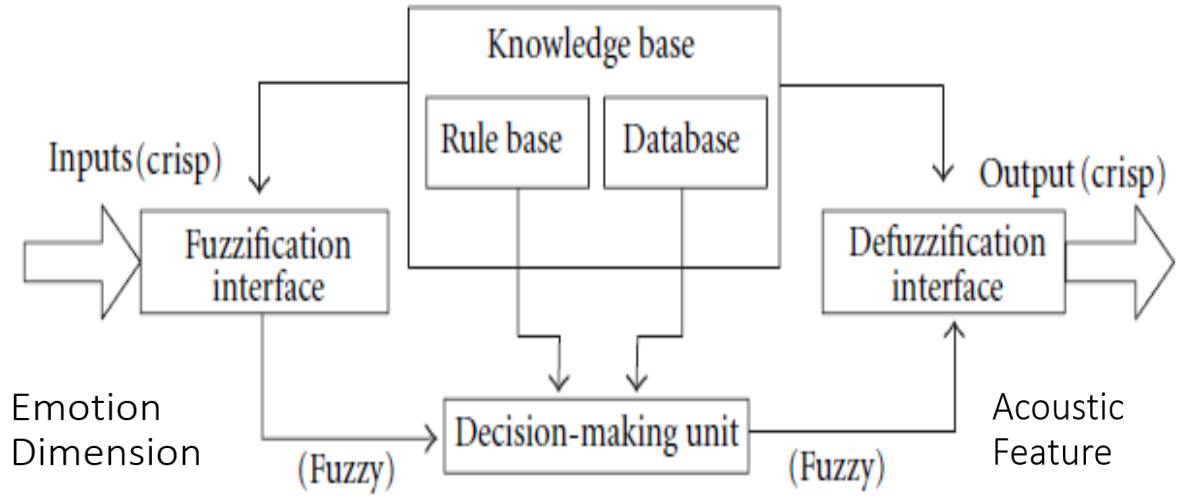

Figure 5.2: Structure of Fuzzy Inference System (FIS)

systems and neural networks as its own structure [83]. ANFIS not only has an inference ability but also a strong learning mechanism. ANFIS is considered instead of other popular methods such as DNN, or NN for two reasons. One is that ANFIS has a membership function with an interpolating method which means that the tendency of the variance in the whole V-A space can be obtained from a small database. A second reason is that fuzzy logic is based on natural language; the natural language in our system is in the form of semantic primitives (the middle layer in the three-layered model).

The structure of ANFIS is shown in Figure **??** which is consisted of five functional components as follows.

1. A rule base including the fuzzy if-then rules.
2. A database containing the membership functions of the fuzzy sets.
3. A decision-making unit which is acted as the inference engine.
4. A fuzzification interface which transforms discrete inputs into linguistic parameters.
5. A defuzzification interface changing fuzzy outputs to discrete output.

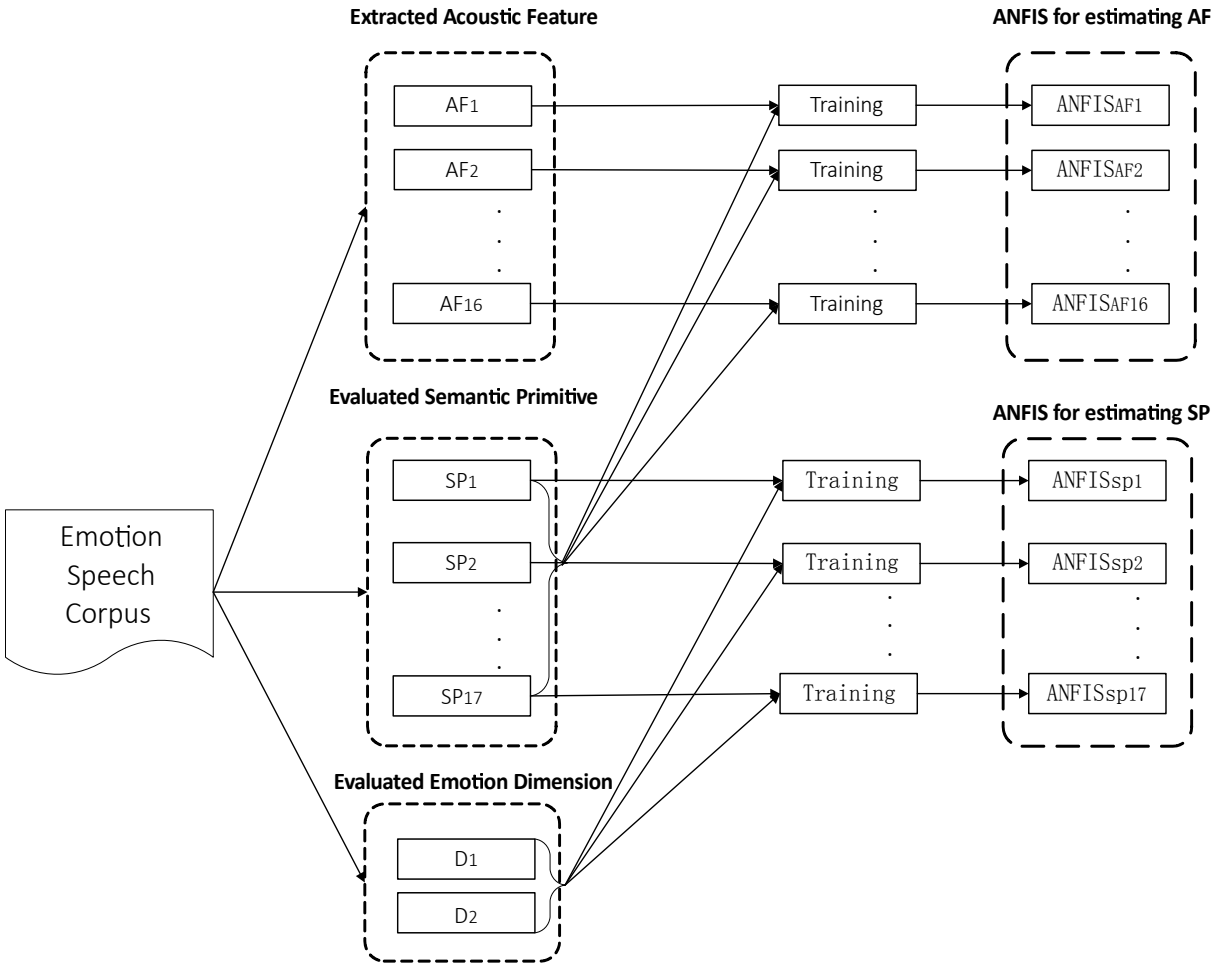

Figure 5.3: *Procedure for training ANFIS.*

5.1.6 Applying a fuzzy inference system for extracting rules

Fig.5.3 shows the flowchart for training the ANFIS to extract rules. Firstly, from the emotion speech corpus as introduced in Sections 6.2, 6.3, 6.4, 16 acoustic features ($AF_1, ..., AF_{16}$) are extracted by STRAIGHT; the 17 semantic primitives values ($SP_1, SP_2, ..., SP_{17}$) and the two emotion dimensions (D_1, D_2) are evaluated by subjects' ratings. To avoid any emotion dependency, all acoustic features are normalized by the mean value of neutral speech. For the ANFIS, all input and output need to range from 0 to 1. We then normalized the acoustic features, semantic primitives and emotion dimensions using the range and minimum value of each parameter using the following Eq.5.3.

$$\tilde{f}_{(i,m)} = \frac{\hat{f}_{(i,m)} - fmin_m}{fran_m} \tag{5.3}$$

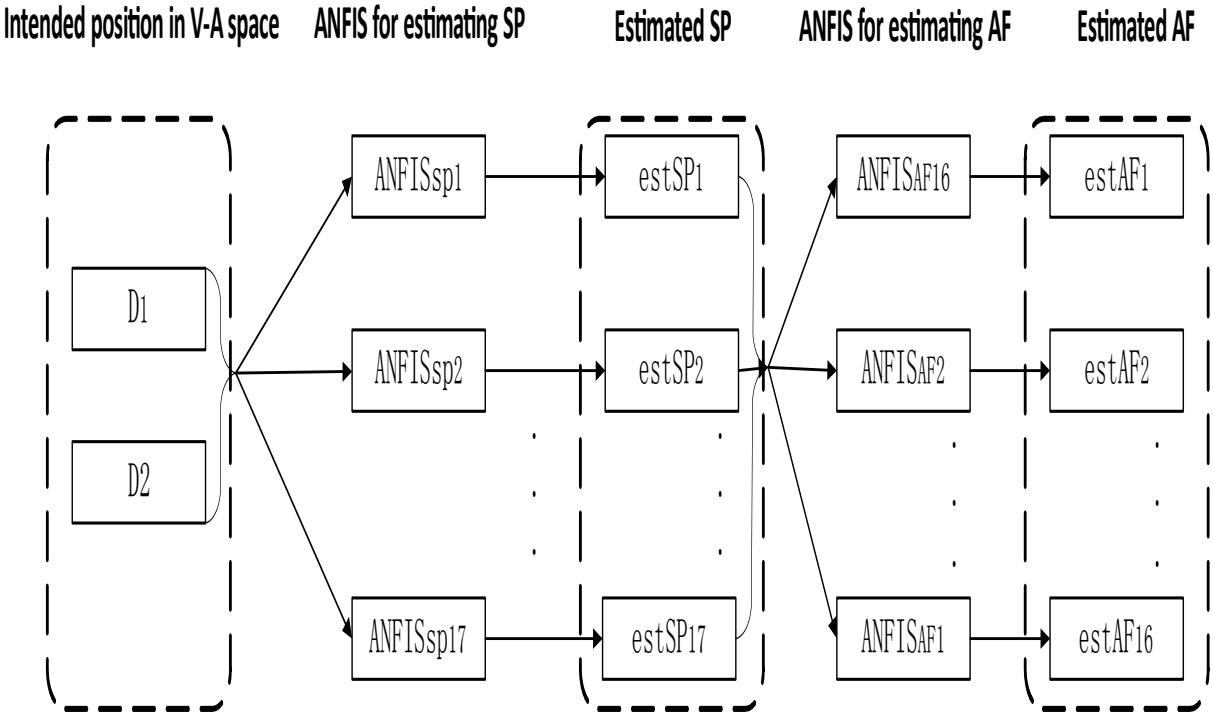

Figure 5.4: *Applying ANFIS for estimating acoustic features (AF) and semantic primitives (SP).*

where m is the number of acoustic features ($m = 1, \ldots, 16$) and i is the number of utterances in the database ($i = 1, \ldots, 179$). $\hat{f}_{(i,m)}$ is the normalized value of the neutral speech. $fmin_m$ and $fran_m$ is the minimum value and range of the mth acoustic features. For semantic primitives and emotion dimensions, the normalized part in $[0, 1]$ are the same as the acoustic features.

ANFIS is a system with multi-inputs and a single-output. In the training phase as shown in Fig.5.3, from the bottom to the middle layer, for each semantic primitive (SP_1, SP_2,..., SP_{17}), we train the appropriate ANFIS ($ANFIS_{SP1}$, ... , $ANFIS_{SP17}$) whose input is the same, that is, the evaluated value of valence and arousal in the emotion dimension (D_1, D_2). From the middle to the top layer, 17 semantic primitives (SP_1, SP_2, ... , SP_{17}) are the input of each ANFIS ($ANFIS_{AF1}$, ... , $ANFIS_{AF16}$), whose outputs are the acoustic features (AF_1, AF_2, ... , AF_{16}).

After the training step, 17 semantic primitives and 16 acoustic features are used in this system to generate 17 ANFISs for estimating semantic primitives and 16 ANFISs for estimating acoustic features. When given the intended position in the V-A space to each of the 17 ANFIS ($ANFIS_{SP1}$, ... , $ANFIS_{SP17}$) for estimating SP , the estimated semantic

primitive ($estSP_1$, $estSP_2$, ... , $estSP_{17}$) is obtained. Applying the 17 estimated semantic primitives ($estSP_1$, $estSP_2$, ... , $estSP_{17}$) as the input to each ANFIS ($ANFIS_{AF1}$, $ANFIS_{AF2}$, ... , $ANFIS_{AF16}$) for estimating acoustic features, the estimated acoustic features ($estAF_1$, $estAF_2$, ... , $estAF_{16}$) are acquired as shown in Fig.5.4.

In the estimation step, the neutral position and the intended position in V-A are given separately to the ANFIS to obtain the acoustic value of neutral speech and the intended speech. We then use the estimated acoustic feature of the intended position in V-A space to divide the estimated AF of the neutral position in V-A space. In this system, we assume that (0,0), the center point in V-A space, is the neutral position. The ratio differences, i.e., the rules between intended and neutral acoustic features, are calculated using the following equation:

$$rule_n = estAF_n / \overline{estAF_n} \tag{5.4}$$

where $estAF_n$ shows the estimated nth acoustic feature value from the ANFIS of the intended emotional state in V-A space and $\overline{estAF_n}$ shows the estimated nth acoustic feature value from the ANFIS of the neutral speech (0,0) in V-A space. Then $rule_n$ represents the rule for the nth acoustic features($n = 1, 2, ..., 16$) which is applied for modifying the neutral speech in the next step.

5.1.7 System evaluation

All data sets are divided into training data (90%) and testing data (10%). ANFIS is first trained using the training data and then validated using the testing data. By giving the value of arousal and valence to the $ANFIS_{sp1}, ANFIS_{sp2}, ..., ANFIS_{sp17}$, firstly, the estimated semantic primitives, $estSP_1, estSP_2, ..., estSP_{17}$ can be obtained and then we input the estimated semantic primitives to $ANFIS_{AF1}, ANFIS_{AF2}, ..., ANFIS_{AF16}$, and after that, the estimated acoustic features $estAF_1, estAF_2..., estAF_{16}$ can be obtained. The accuracy of the estimated acoustic features and estimated semantic primitives are evaluated by mean absolute error (MAE); this can measure the distance between the estimated values by the proposed system and the annotated values from listeners evaluations.

$$MAE = \frac{\sum_{i=1}^{N} |x_i - y_i|}{N} \tag{5.5}$$

where $x_i (i = 1, 2, ..., N)$ is the sequence of estimated values of one semantic primitive or one acoustic feature. $y_i (i = 1, 2, ..., N)$ is the sequence of annotated values by listeners for the corresponding semantic primitive and acoustic feature. N is the number of utterances in the database.

Figs 5.5 displays the MAE results of semantic primitives between the training data and testing data from the three-layered model. Fig.5.6 shows the MAE of acoustic features from three-layered model and two-layered model. The two-layered model utilized the same methodology of applying the emotion dimension for representing emotion but without considering using the semantic primitive layers. The MAE of 15 semantic primitives are all below 10%, and for the fast and slow semantic primitives, the MAE is somewhat higher, near 10% which means that the estimation accuracy of semantic primitives is very high. From Fig.5.6, comparing the results from the two-layered and three-layered model, it is found that among 16 acoustic features, the MAE values of 10 acoustic features from the three-layered model are lower than the two-layered model which means that the three-layered model can provide higher estimation accuracy than the two-layered model. Among the 16 acoustic features, all are below 20% and only the MAE of PW_RAP is higher than 15% using the three-layered model.

5.2 Rule application

For the emotional voice conversion system, the acoustic parameters of neutral speech need to be modified in order to synthesize the emotional speech. The ratios, rules of the relationships between acoustic features between neutral and intended emotion, are calculated by ANFIS through the inverse three-layered model. In this section, the modification method based on the extracted rules is explained.

As shown in Fig.5.7, first, the phoneme boundaries of the vowels and consonants are extracted manually from the neutral speech. Then the ratios between the neutral and target emotional speech of the acoustic features TL, CL, RCV are used to modify the

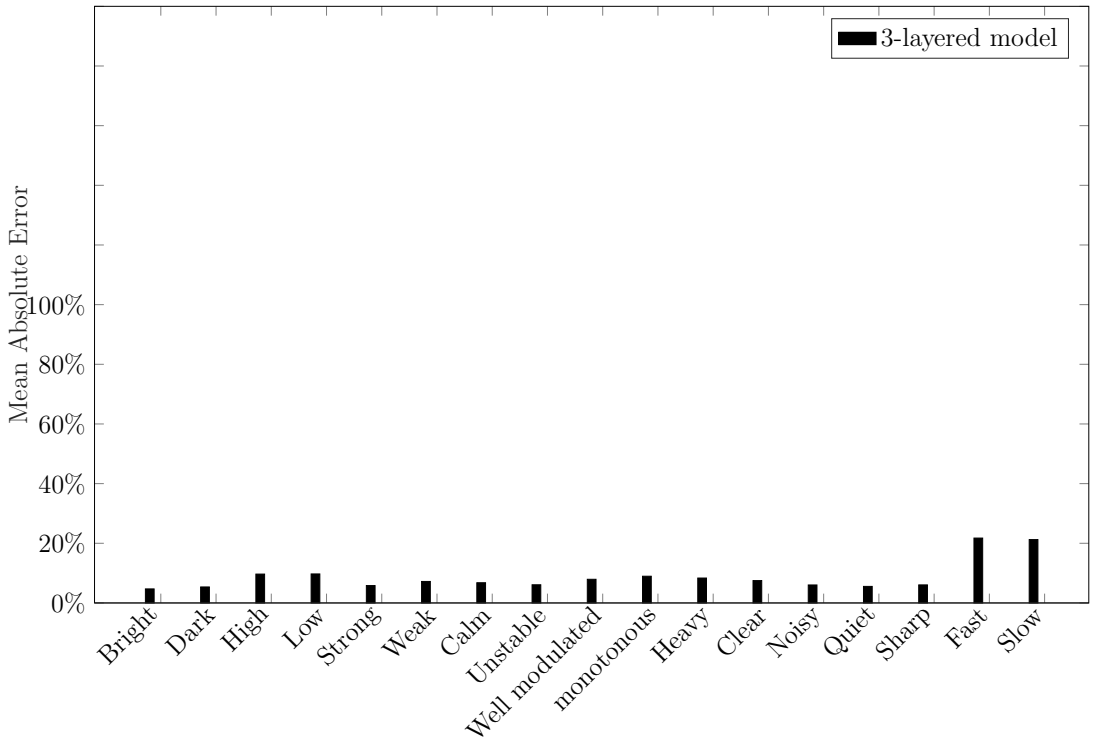

Figure 5.5: Mean absolute error of semantic primitives.

phoneme boundaries. The F0 contour is extracted by STRAIGHT at the same time and interpolated using the duration information. The F0 contour is parameterized by a modified version of the Fujisaki model to modify the F0 contour. After F0 modification, STRAIGHT is applied to obtain the modified speech. Lastly, the power envelope modification is done by using the target prediction model. After the power envelope modification, the final converted emotional speech can be acquired.

5.2.1 Fujisaki model for parameterizing F0 contour

Previous work separately modified the F0 related acoustic features, such as average F0 (AP), highest F0 (HP), the mean value of F0 in the rising slope (RS) and rising slope of the first accentual phrase (RS_1st). In our case, separately modifying the acoustic features is not suitable, because modifying one acoustic feature such as RS may influence other acoustic features such as AP and HP and there is no appropriate order for modification. We parameterized the F0 contour to control the entire contour using only a limited set of parameters.

The Fujisaki model [1], a mathematical model represented by the sum of phrase com-

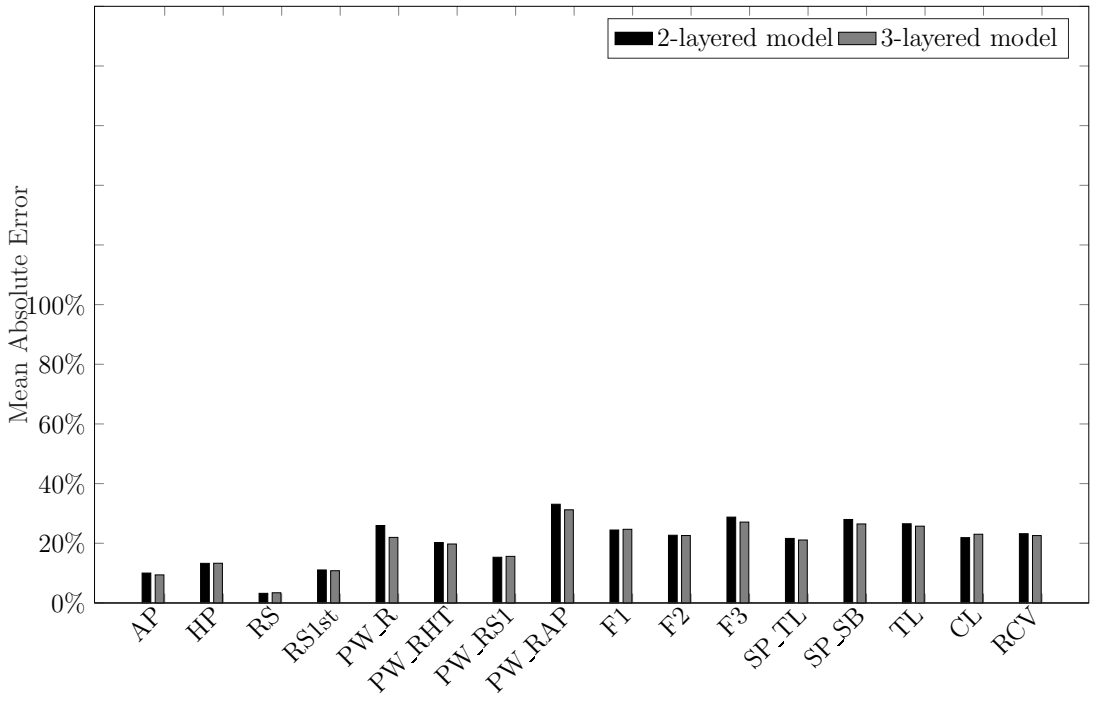

Figure 5.6: Mean absolute error of acoustic features from three- and two-layered model.

ponents, accentual components, and the baseline Fb, is adopted to parameterize the F0 contour. The F0 contour can be expressed as follows.

$$\ln F_0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp_i(t - T_{0i}) + \sum_{j=1}^{J} Aa_j\{Ga_j(t - T_{1j}) - Ga_j(t - T_{2j})\} \quad (5.6)$$

$$Gp_i(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (5.7)$$

$$Ga_j(t) \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (5.8)$$

where $G_{p(t)}$ represents the impulse response function of the phrase control mechanism,

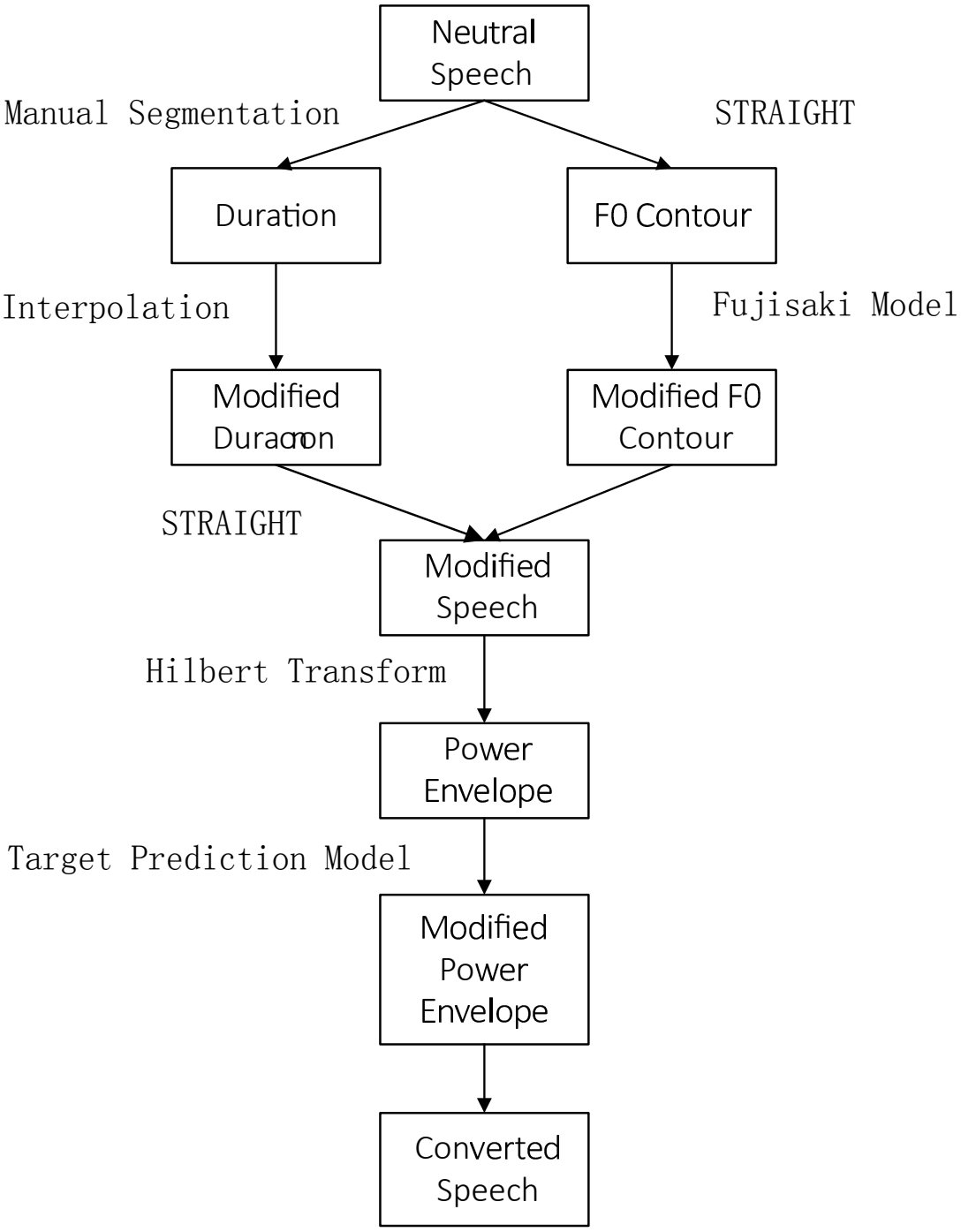

Figure 5.7: The procedure of modifying the neutral speech to emotional one.

and $G_{a(t)}$ represents the step response function of the accent control mechanism. The symbols in these equations forecast

F_b: baseline value of fundamental frequency,

I: number of phrase commands,

J: number of accent commands,

A_{pi}: magnitude of the ith phrase command,

A_{aj}: amplitude of the jth accent command,

T_{0i}: timing of the ith phrase command,

T_{1j}: onset of the jth accent command,

T_{2j}: end of the jth accent command,

α: natural angular frequency of the phrase control mechanism,

β: natural angular frequency of the accent control mechanism,

γ: relative ceiling level of accent components.

Many researchers utilize the Fujisaki model; the work of Mixdorff [84] is adopted in this paper where $\alpha = 1.0/s$ and $\beta = 20/s$. By using Mixdorff's method the parameters ($T0$, $T1$, $T2$, Ap, Aa, and Fb) in the Fujisaki model are extracted. We then modify the parameters to obtain a modified F0 contour using Equations 5.6, 5.7 and 5.8. We can extract the AP, HP, RS, and RS_{1st} of the modified F0 contour. The root mean square error (RMSE) between the desired acoustic features and extracted one from the modified F0 contour is calculated using the following equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(AF_i - \hat{AF_i})^2}{N}} \tag{5.9}$$

where AF_1, ... ,AF_n is the desired acoustic feature value which is denormalized after being estimated from ANFIS. And the $\hat{AF_1}$, ... ,$\hat{AF_n}$ is the extracted value from the modified F0 contour. The F0 contour with the smallest RMSE is selected as the final F0 contour.

By controlling fundamental frequency, neutral speech was converted into emotional speech related to a position on the V-A space which is shown in Figure 5.8. The results showed the fundamental frequency was able to control using Fujisaki model with appropriate values compared to an estimated values gotten from a position of the V-A space in three-layered model.

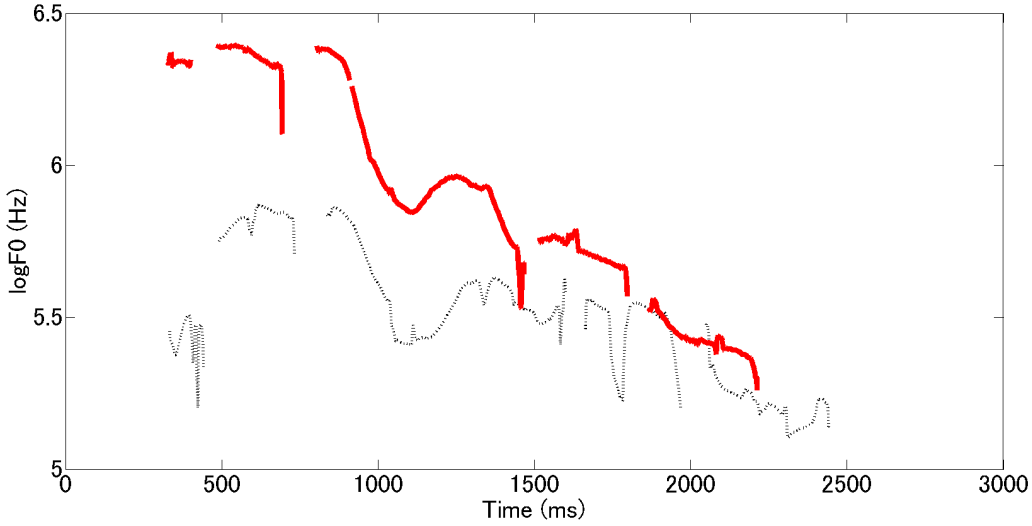

Figure 5.8: F0 trajectory of a neutral speech (dashed) and synthesized speech (solid) [104].

5.2.2 Target prediction model for parameterizing the power envelope

In order to parameterize the power envelope target, a prediction model which predicts the stable power target in short-term intervals is used to estimate the targets of the power envelope [85]. We then change the targets to a stepwise function by using the segmentation information, the starting and ending points of each phoneme. By modifying the magnitude of the stepwise targets of the power envelope, a modified power envelope is reproduced by a 2nd-order critically damped model. The procedure of reproducing power envelope is shown in Fig. 5.9.

The neutral speech signal in Fig. 5.10 is represented as $y(t)$ The power envelope of the neutral speech signal is firstly extracted by

$$e_y(t) = \mathbf{LPF}\left[|y(t) + j\text{Hilbert}\left[y(t)\right]|^2\right] \tag{5.10}$$

where LPF[·] is a low-pass filtering and Hilbert[·] is the Hilbert transform. Then we used Eq.5.11 to change the power envelope in the log power envelope domain. Fig. 5.11 shows the extracted log power envelope.

$$\log e_y(t) = 10\log_{10}(e_y(t)) \tag{5.11}$$

Then the power envelope is approximated by a 2nd-order critically damped system which can estimate the target power envelope using short-term power sequences without being given the onset positions of the power transition.

A 2nd-order critically damped model is generally represented as follows

$$\left(\Delta^2 - 2\lambda\Delta + \lambda^2\right) y_n = \lambda^2 b \tag{5.12}$$

where Δ is a differenctial operator in time, λ is a reciprocal time constant, time $n = 0$ is the onset position of the transition and b is a target to which y_n converges in the past if $\lambda > 0$ and $n \leq 0$, or in the future if $\lambda < 0$ and $n \geq 0$. The solution of Eq.5.12 is

$$y_n = (a + cn) \exp(\lambda n) + b \tag{5.13}$$

where a and c are constants obtained from the boundary condition. Previous methods that estimated the parameters of 2nd-order critically damped models have predicted all parameters directly by using Eq.5.13 and the following measure,

$$e(n_0 \text{ or } n_1, \lambda) = \sum_{n=n_0}^{n_1} \left| y_n^i - y_n \right|^2, \quad n_0 < n_1 \tag{5.14}$$

where y_n^i is an unknown input sequence. For these methods, a long-term sequence sufficient to start at the onset position of the transition $n_0 = 0$ when $\lambda < 0$ or $n_1 = 0$ when $\lambda > 0$ is essentially required. Then, non-linear optimization under two values, n_0 and λ or n_1 and λ is needed. However, the purpose of our target prediction model is to estimate b only.

Divide Eq.5.12 such that;

$$(\Delta - \lambda) \{(\Delta - \lambda) y_n\} = \lambda^2 b \tag{5.15}$$

and assume that

$$x_n = (\Delta - \lambda) y_n \tag{5.16}$$

$$(\Delta - \lambda) x_n = \lambda^2 b \tag{5.17}$$

By substituting Eq.5.13 into Eq.5.16,

$$x_n = c \exp(\lambda n) - \lambda b \tag{5.18}$$

and Equation 5.18 is a first-order equation.

Assuming that

$$c_m = c \exp(\lambda m) \tag{5.19}$$

at time $n = m$, the neighborhood x_{m+t} of x_m is represented by

$$x_{m+t} = c_m \exp(\lambda t) - \lambda b \tag{5.20}$$

Thus, if the measure

$$\begin{aligned} e(\lambda) &= \sum_{t=n_0}^{n_1} \left|(\Delta - \lambda) y^i_{m+t} - x_{m+t}\right|^2 \\ &= \sum_{t=n_0}^{n_1} \left|x^i_{m+t} - x_{m+t}\right|^2 \end{aligned}$$

can be used, non-linear optimization under only λ is needed and it does not require any knowledge of the onset position of the transition estimating the target b, because x_{m+t} is an exponential function. In this prediction, if $\lambda \geq 0$, it is the backward prediction (target in the past). If $\lambda < 0$, it is the forward prediction (target in the future). We use forward prediction (target in the future) to reproduce the power envelope.

In Fig.5.11, the blue line shows the estimated target of the power envelope using the target prediction model.

The onset point T_{1j} and ending point T_{2j} of each phoneme was segmented manually. After obtaining the estimated power envelope, we calculated the average value, Au_j of the jth step in each period of one phoneme which consisted of the stepwise function shown in Fig.5.13, black line. These are the inputs of the Eq.5.21 that follow the accent mechanism of the Fujisaki model. The stepwise input signals to the power control mechanism are defined by their amplitude Au_j, onset time T_{1j} and offset time T_{2j} using Eq.

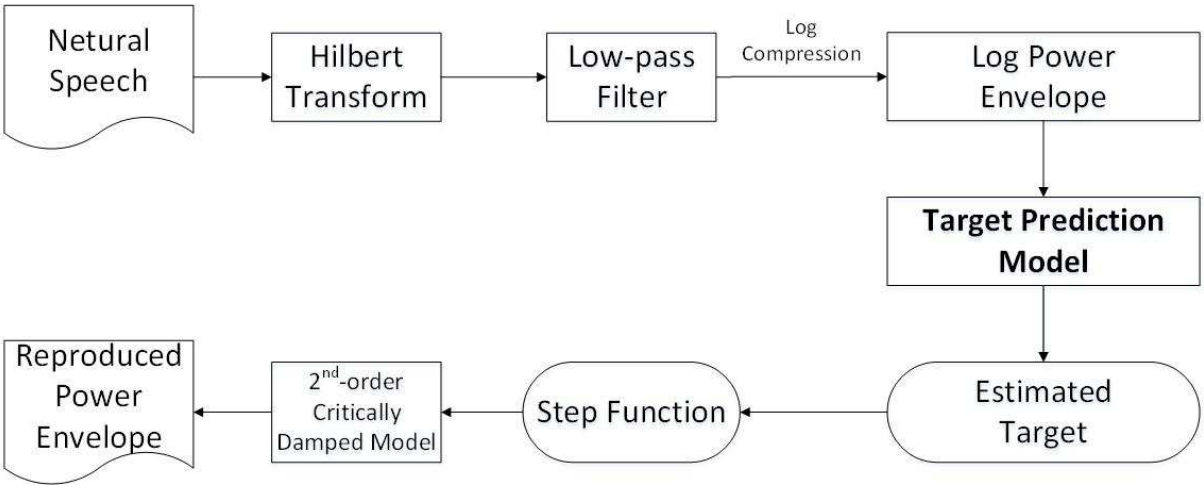

Figure 5.9: Procedure of reproduing power envelope

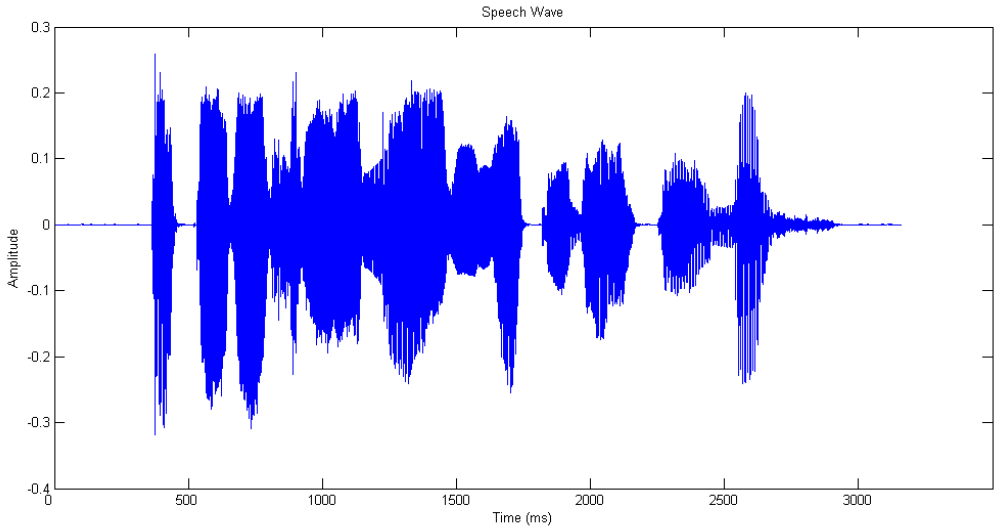

Figure 5.10: Speech wave of the original speech

$$\log e_y(t) = \sum_{j=1}^{J} Au_j[Gu(t - T1_j) - Gu(t - T2j)] \tag{5.21}$$

where $\log e_y(t)$ is the reproduced power envelope. And the step-response $Gu_{(t)}$ is calculated using the following equation

$$Gu_j(t) = 1 - (1 + \delta t)\exp(-\delta t) \quad t \geq 0 \tag{5.22}$$

The symbols in these equations forecast

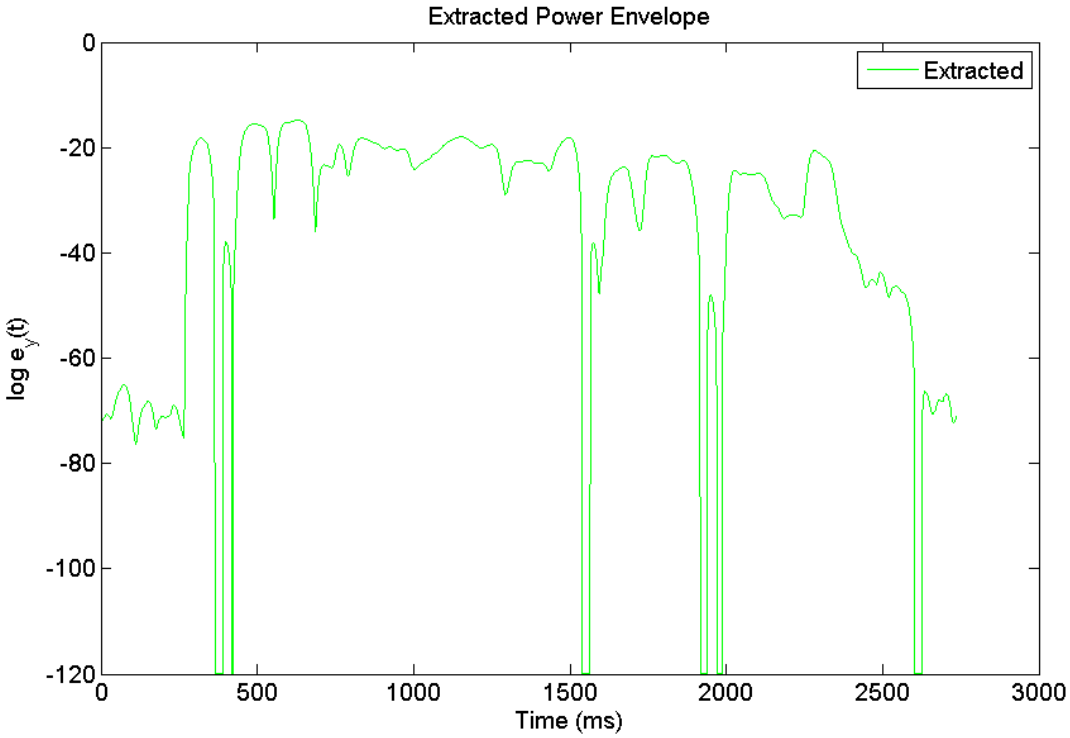

Figure 5.11: Extracted power envelope

- Au_j: amplitude of the jth step, Au_j is the average value of b in each segmentation,
- $T1_j$: onset of the jth step,
- $T2_j$: offset of the jth step,
- δ: time constant.

  δ is the absolute value of the sum of the negative parts of λ as we use a forward prediction, $\lambda < 0$ (target in the future), to reproduce the power envelope.

In Fig.**??**, the reproduced power envelope and extracted log power envelope are shown. Signal/Error Ratio (SER) in Eq.5.23 and Mean Absolute Error (MAE) in Eq.5.24 are used to evaluate the difference between the extracted and reproduced power envelope. As the voiced signal is more important than the unvoiced parts in this research, SER is calculated only during the voiced part.

$$SER = 10\log_{10}\frac{\sum_{i=1}^{N}(x_i)^2}{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{5.23}$$

Figure 5.12: Target of power envelope estimated by target prediction model

$$MAE = \frac{\sum_{i=1}^{N} |x_i - y_i|}{N} \tag{5.24}$$

where x_i is the extracted power envelope and y_i is the reproducing power envelope. N is the number of bits in the voiced part.

The value of SER is 18.01dB and the MAE is about 1.82dB which means that the reproduced power envelope is almost the same as the original extracted power envelope. Therefore, we can conclude that this method works well for parameterizing the power envelope. After this, we modified the power envelope by controlling A_{aj} to fit the estimated acoustic features.

In Fig. 5.11, the blue line shows the estimated target of power envelope using the target prediction model.

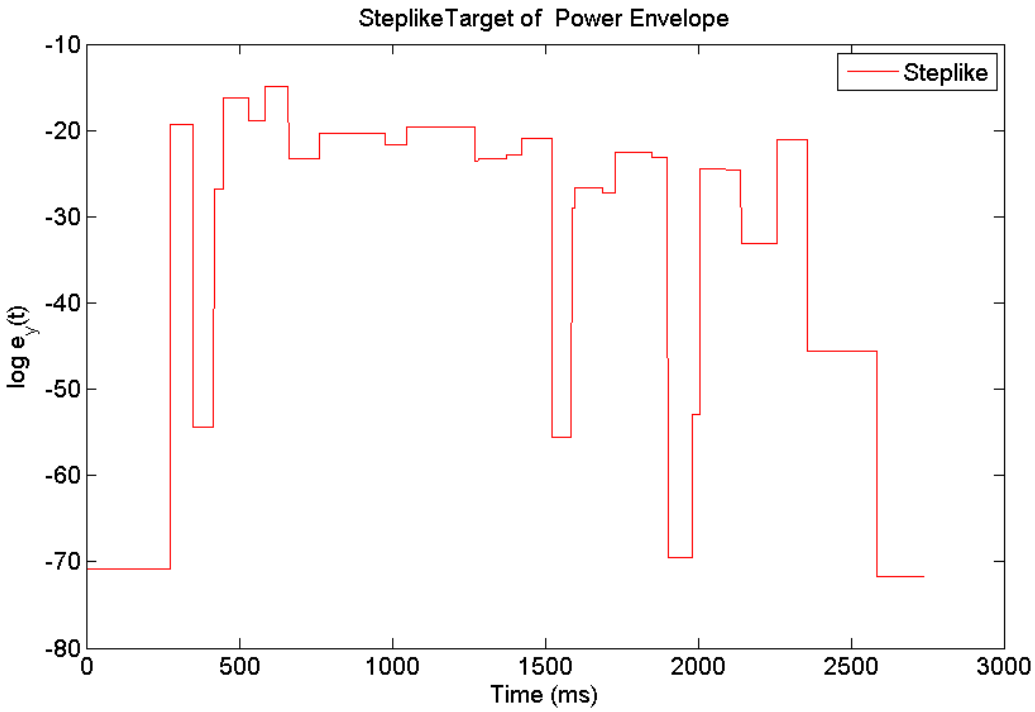

Figure 5.13: Steplike targets of power envelope

5.3 Perceptual evaluation

The voice conversion system for emotional speech aims to control the degree of emotion in dimensional space. We hypothesize that the system can convert any utterance from any speaker by a given point in dimensional space using a limited database. This is the procedure for the synthesized utterances for the evaluation phase, so there is no reference to the desired position of the corresponding emotional utterance in the corpus. Hence, the objective measures such as Mel-cepstral distortion or mean squared error between converted and target are not suitable. The inputs of the conversion system are the intended position value in V-A space and the neutral speech. We utilized the distance between the intended position and the evaluated position obtained from the perception experiment to evaluate the category and the degree of emotion.

5.3.1 Stimuli

In the following subjective evaluation experiments, the inputs of the system for emotional voice conversion are three different neutral statements spoken by the single speaker from

Figure 5.14: Reproducing power envelope using 2nd-order critically damped model and the extracted power envelope from original speech

the Japanese Fujitsu database. The English meaning of the three statements are the following:

1. You have new mail.

2. Nothing new has come to mind.

3. I am already home.

The input positions to the system in V-A space are shown in Fig.5.15 with solid points. The range of valence and arousal is from -2 to 2 in increments of 0.1. The position values among the three utterances are the same. In the 1st and 3rd quadrants, there are 3 positions. Since there are two kinds of anger emotion, hot anger and cold, there are two positions for each in the 2nd quadrant. One position in the V-A space represents one synthesized utterance with different degrees of emotion. Including the neutral original speech, there are 11 stimuli for each utterance with a total number of 33 synthesized speech utterances.

5.3.2 Experiment procedure

16 Japanese subjects (7 females and 9 males) with normal hearing, average age about 23.3 years old, participated in the experiment. Subjects listened to the stimuli in a random order presented through an audio interface (FIREFACE UCX, Syntax Japan) and headphones (HDA200, SENNHEISER) in a soundproof room. The original sound pressure level was about 64 dB. The subjects evaluated the stimuli with regard to three aspects, valence, arousal and naturalness. Each aspect was evaluated as a separate test in order to avoid the conceptual confusion between valence and arousal, with at least a 3 hour time interval between tests. Subjects evaluated these scales using a graphic user interface as shown in Fig.4.2. The ranges, scale steps, and other rules are the same as explained in Section 4.

5.3.3 Subjective evaluation result in V-A space

Analysis of the evaluated results mainly focuses on two parts: perception of the emotion category and the degree of emotion. The evaluated results (perceived positions) analyzed in terms of emotion category in the valence and arousal spaces are shown in Fig.5.16. The oval is calculated using average and standard deviation of valence and arousal values. The central point of each oval is the mean value of each emotion. The radius of the oval shows the standard deviation related to valence and arousal of each emotion. Fig.5.16 shows that the mean value of evaluated joy, cold anger and sadness can be obtained in the intended quadrant and the standard deviation is acceptable for each emotion; this means that the category of emotion can be perceived well by subjects for joyful, cold anger, and sad emotional speech. But for hot anger, the intended position is the second quadrant while the evaluated line of hot anger is in the first quadrant, so subjects perceived synthesized hot anger as a joyful emotion. The reason for this misunderstanding is that, as we mentioned in Section 4, only by replacing the spectral sequence of hot anger can the neutral speech be perceived as hot anger. For now, our modification method only controls for duration, F0, and power envelope. Therefore, hot anger emotion cannot be well obtained.

The degree of emotion perception is shown in Fig.5.15. As the input positions of the three different linguistic utterances are the same, the average evaluated values for

each position among the three utterances are calculated. In Fig.5.15, the solid circles represent the intended position and the hollow circles are the positions evaluated from the perceptual experiment in V-A space. The dashed lines show the distance of the two pairs: intended and evaluated. From Fig.5.15, we can see that the tendencies of the degrees of valence and arousal for the intended and evaluated emotions are the same, except for cold anger. Moreover, we note that the degree of the synthesized speech is more mild than intended. This phenomenon is in line with the results reported in Section 4. It is found that if only the F0 or spectral sequence of neutral speech is replaced by those from the emotional speech, the degree of perceived emotion is decreased.

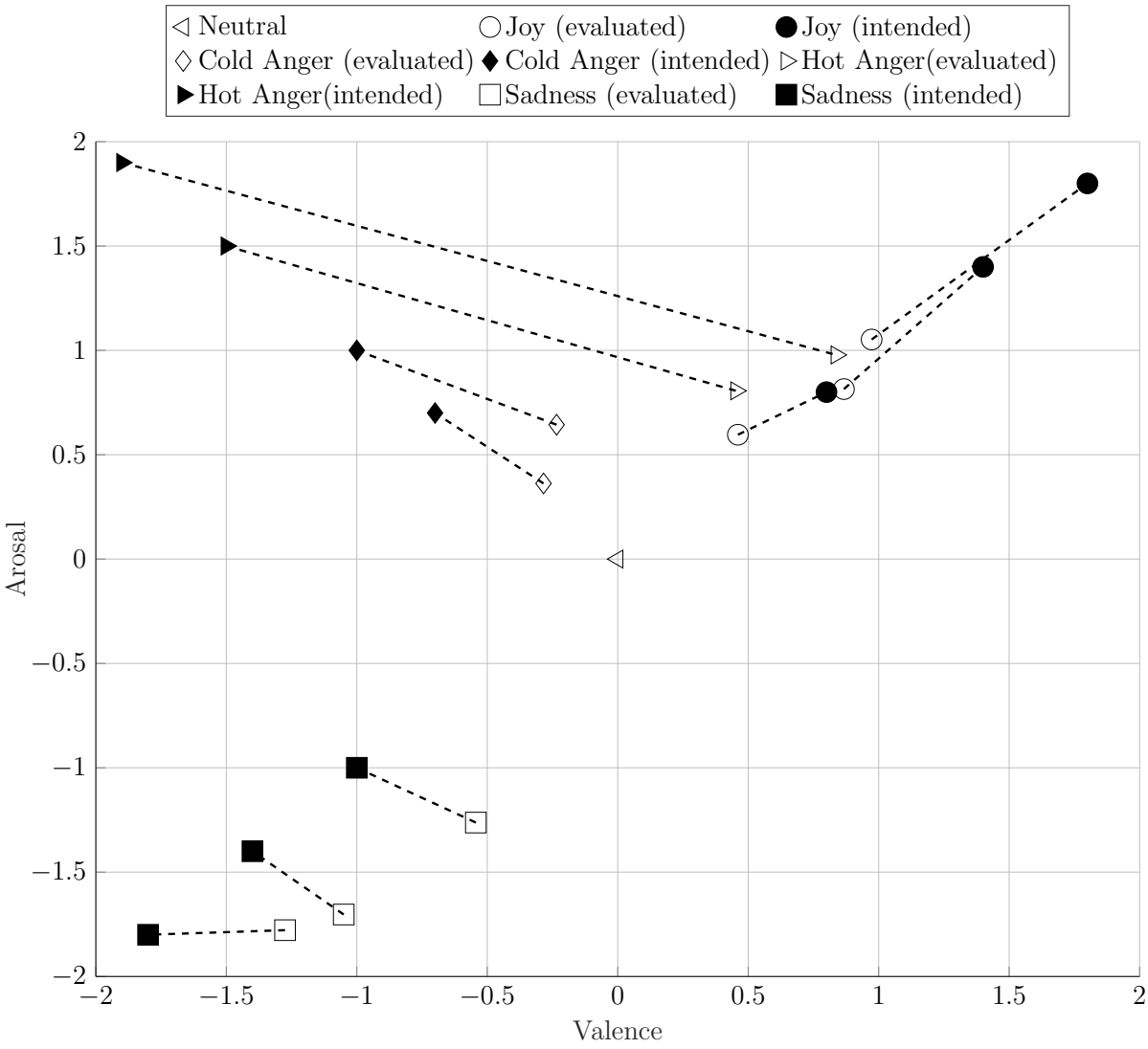

Figure 5.15: The evaluated and intended positions in V-A space. (the dashed lines are the intended position and the solid lines are the obtained position.)

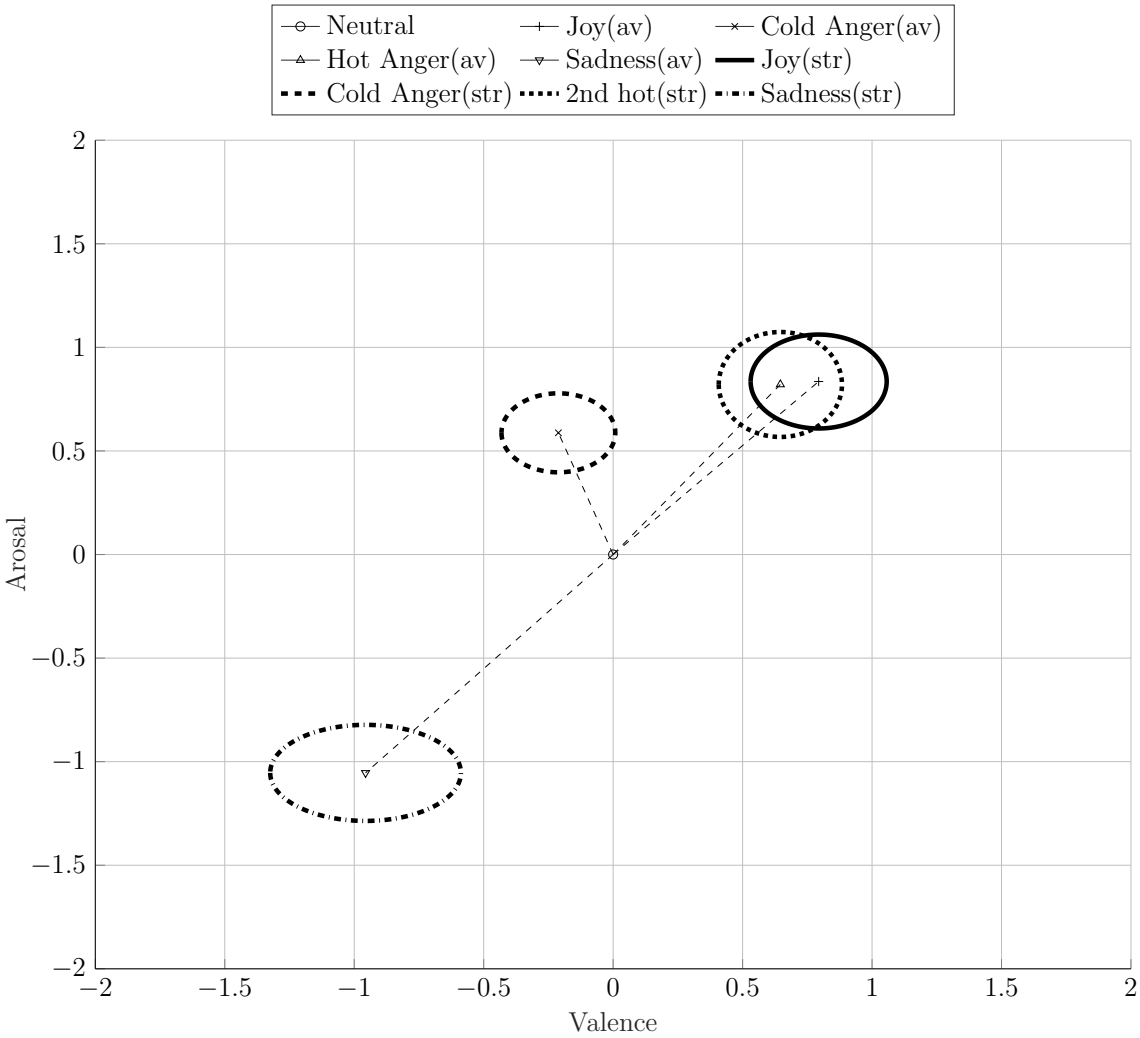

Figure 5.16: The average and standard deviation of the evaluated results. (*1st, 2nd, 3rd stand for the intended quadrants.* Cold *and* hot *represent the intended cold anger and hot anger.* av *and* str *mean the average and standard deviation values of each quadrant.)*

5.3.4 Subjective evaluation result of naturalness

The naturalness quality of the converted utterances was rated on a 1-to-5 scale [1-bad, 2-poor, 3-fair, 4-good, 5-excellent] using the neutral sentence as the reference. The Mean Opinion Score (MOS) is shown in Fig.5.17. The MOS of each emotion is calculated separately. From these results, we see that all naturalness scores are fair, i.e., above 2.5. Joyful speech was rated best (MOS about 3.38), with cold anger as a second (MOS about 3.1). The MOS of hot anger and sad are about 2.98 and 2.27. The reason that the quality of sadness is the lowest is because that the duration of sad speech is long but the pauses between phrases were not markedly obvious. We treated the ratio of modification to voice and unvoiced part the same. Therefore, the synthesized speech seemed machine-like. More

precise control of duration ratios between voiced and unvoiced periods is needed in order to improve the quality of sadness; this is a topic that will be researched in the future.

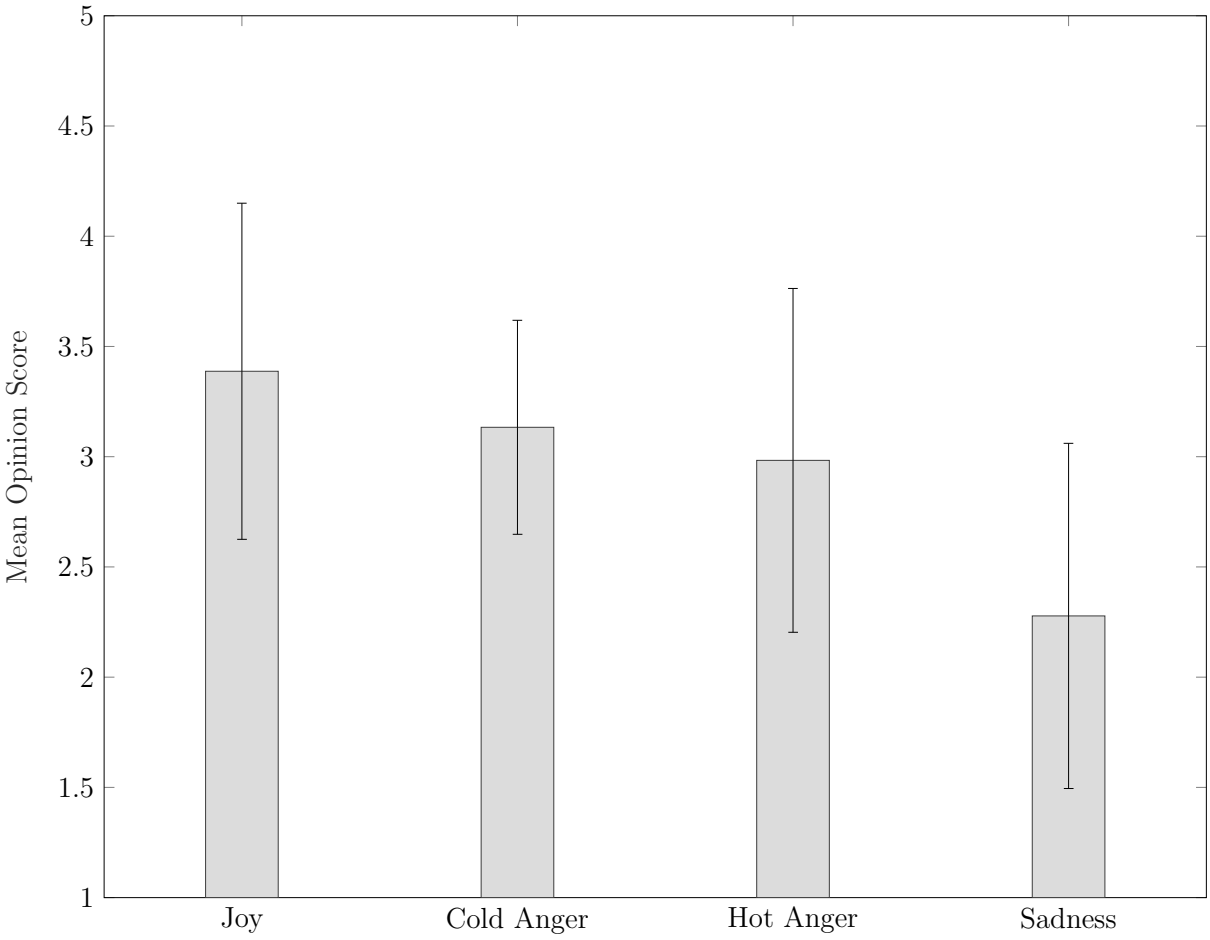

Figure 5.17: Mean opinion scores for converted speech in each quadrant.

Chapter 6

Emotion conversion for multiple languages

As the collection of speech for multiple emotions in multiple languages costs huge, if one system can be applied to multiple languages, it will improve the efficiency. Therefore, this chapter aims to explore the possibility for apply the mono trained emotional voice conversion system to other languages without training.

6.1 Commonalities of human perception for emotional speech among multi-languages

Many previous researches have focused on the perception differences of emotional speech among different languages and among different mother-languages listeners. Huang [105] tested the adjective that Japanese and Mandarin listeners used to describe the Japanese emotional speech. Results show that 60% adjectives they chose are common. Subjects from three different countries were invited to evaluate the emotion categories of the emotional speech database without linguistic information. Principal component analysis results revealed that some common factors are shared among human being for perceiving emotion.

These previous researches based their research on emotion category representation. Han in 2015 analyzed [88], five emotional speech databases in five different languages, Japanese, German, English, Vietnamese and Chinese in valence-arousal space. Four emo-

tional states, happy, angry, neutral, and sad, were selected from the five databases. Thirty subjects from three different countries, Japan, China and Vietnam, evaluated the three databases in terms of valence and arousal. The five databases are CASIA database, IEMOCAP database, Berlin database, Fujitsu database and VNU database in Chinese, American English, German, Japanese and Vietnamese respectively.

Han compared the experimental results in three points of view:

1. the position of neutral state.

2. the direction of emotional states.

3. the degree of emotional states.

Results show that same positions of neutral states can be achieved by different native language groups. And the direction from neutral to other emotional state is also identical among them. However, the degree is different. This suggests that even with different language and cultures, human beings can have the same feeling on neutral speech.

Some emotion recognition system based their researches on these findings. Li [49] proposed a multilingual speech emotion recognition system based on a three-layer model who believed that human emotion recognition can be constructed using one system.

Based on this result, we hypothesize that, given the same direction in V-A space from neutral voice to other emotional states, the emotion conversion system can also convert other languages with the same impression of emotion.

In order to confirm this hypothesis, we apply the emotion conversion system build for Japanese to two other languages, English and Chinese, without training using the two languages. In the following section, the outline of the emotion conversion system and the procedure for applying the emotion conversion system to other languages are illustrated.

6.2 Applying conversion system to other languages

As we investigate whether, given the same direction from neutral speech to other emotional states, the converted speech can give the same impression among multiple languages, two different neutral speech as well as Japanese one are given as input to the emotion conversion system.

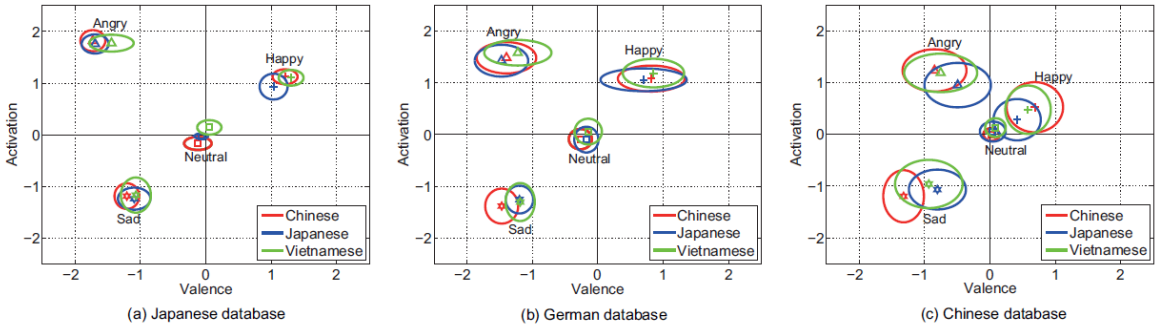

Figure 6.1: Emotional states position on Valence-Activation approach [88].

One utterance is spoken in Chinese by a professional female voice actor which is selected from the Chinese emotional corpus developed by Institute of Automation, Chinese Academy of Sciences (CASIA). The content is in Chinese means "He ends an meaningless love" in English. Another utterance is spoken in English by a female US English speaker from English CMU ARCTIC database. The content is "LORD BUT I'M GLAD TO SEE YOU AGAIN PHIL".

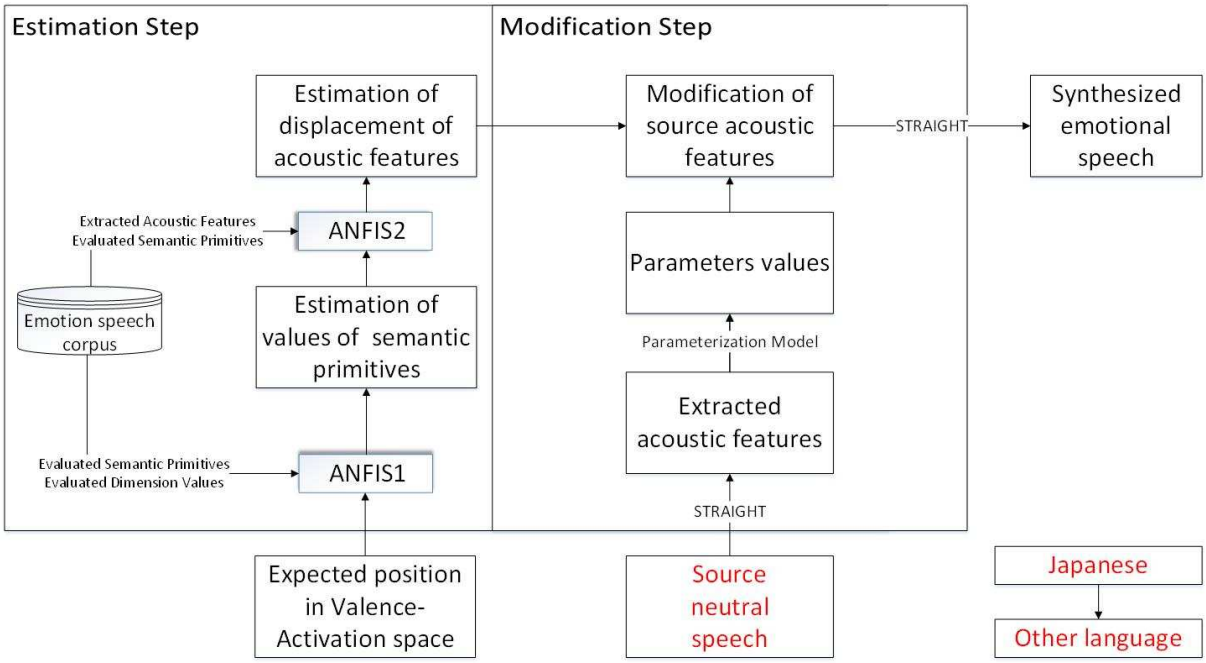

Figure 6.2: Convert the source neutral speech from Japanese to other languages

This emotion conversion system built for Japanese shown in Fig. 6.2 is applied to Chinese and English sentences without changing parameter values in the rules extraction and rules application steps. The inputted positions are the same among three languages,

that we choose the 8 positions with either large or small value of valence or activation to represent the intended and intensity of emotion; Four in the 1st quadrant that represent joy and four in the 3rd quadrant which represent sad emotion. The input position is shown as dashed line in Fig. **??**.

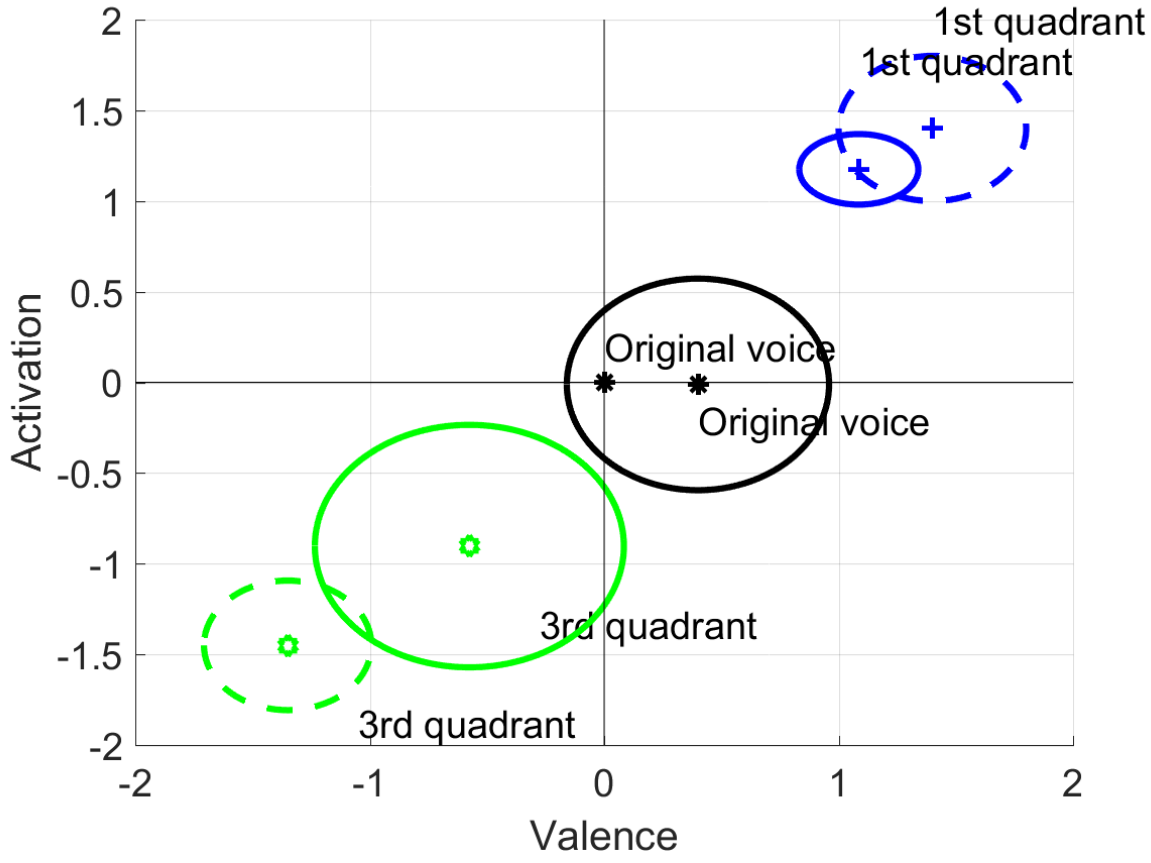

Figure 6.3: Evaluation results (emotion category) of synthesized voices in Chinese.

6.2.1 Listening Test

To verify whether the synthesized voices care well perceived by humans, we carried out subjective listening tests to let subjects evaluate the synthesized speech in the V-A space.

Subjects and Stimuli

In the listening test, 9 subjects (two Vietnamese female, one Vietnamese male, three Japanese male, one Chinese female and two Chinese male, mean 26 years old) with normal

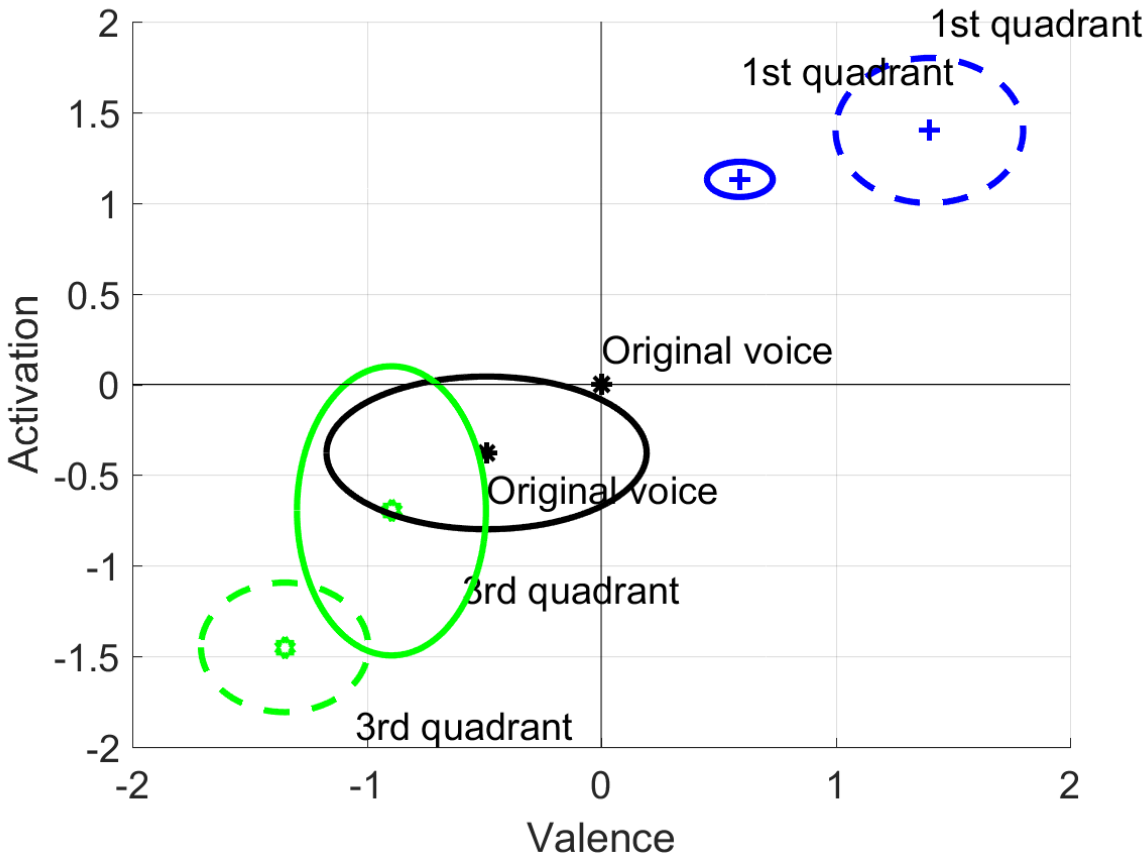

Figure 6.4: Evaluation results (emotion category) of synthesized voices in English.

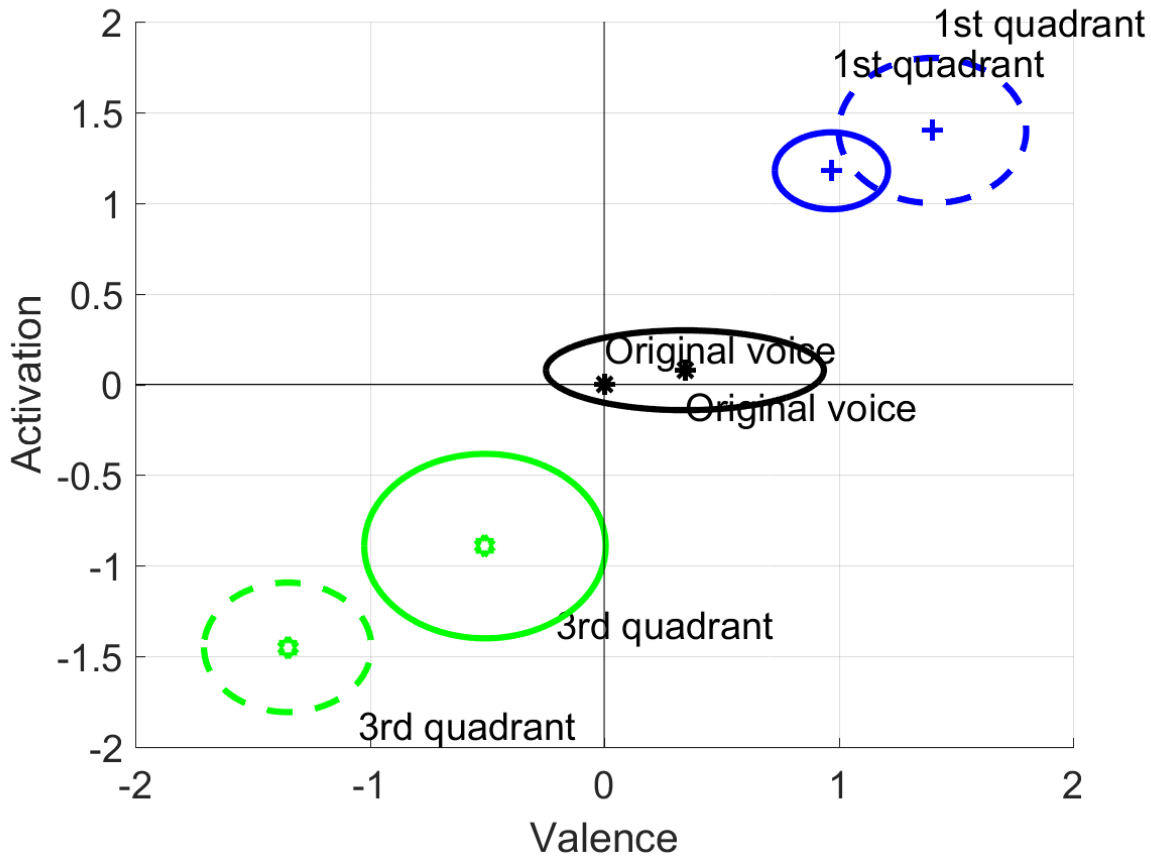

Figure 6.5: Evaluation results (emotion category) of synthesized voices in Japanese.

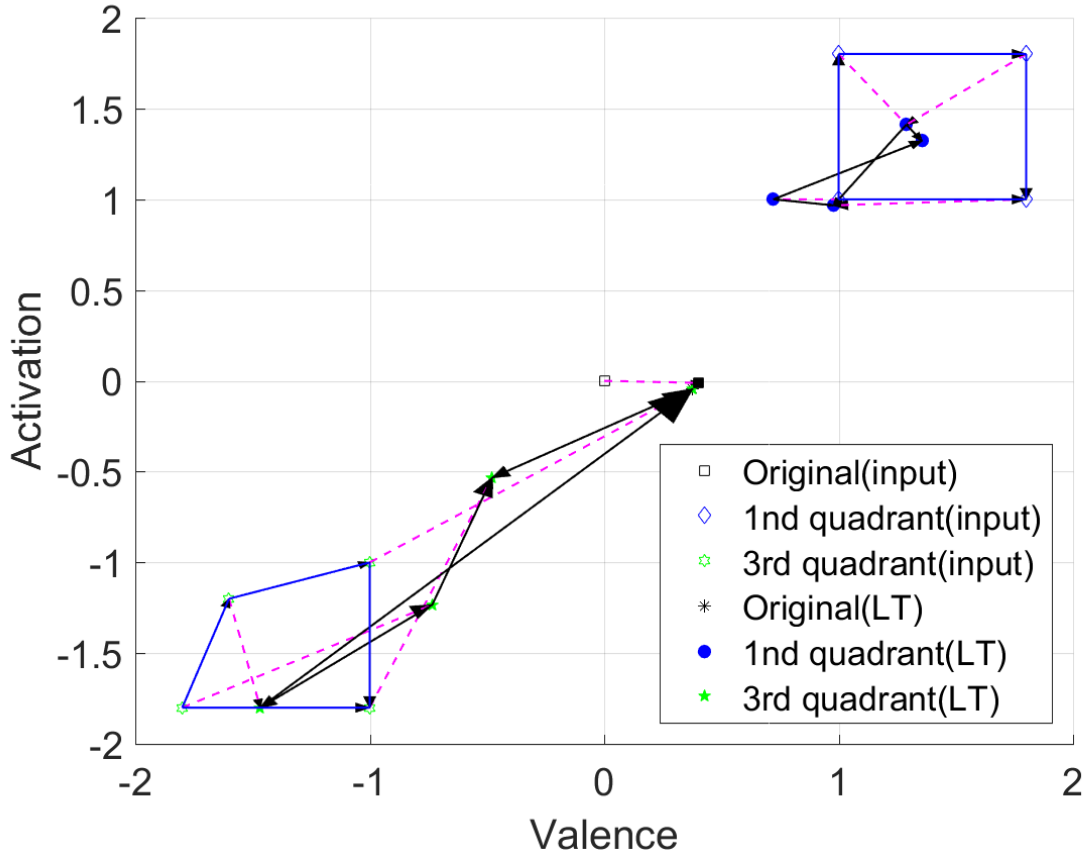

Figure 6.6: Evaluation results (emotion degree) of synthesized voices in Chinese.

hearing ability gave evaluation scores on three aspects: activation, valence and naturalness. 27 stimuli are presented to the subjects. The 27 stimuli contain three languages and each language has 9 stimuli. Among the 9 stimuli, four voices of joy, four voices of sad and one voice of neutral. The neutral speech is the original speech given as inputs of the system and the 8 stimuli for joy and sad are prepared with either largest or smallest values of valence or activation.

Procedure

Subjects were asked to listen to the stimuli presented through an audio interface (FIREFACE UCX, Syntax Japan) and headphones (HDA200, SENNHEISER) in a soundproof room. The original sound pressure level was 64 dB.

For valence and activation, subjects listened to all stimuli twice. This was done so that they could acquire an impression of the whole stimulus the first time and then evaluate one dimension from -2 to 2. Valence and activation needed to be done separately

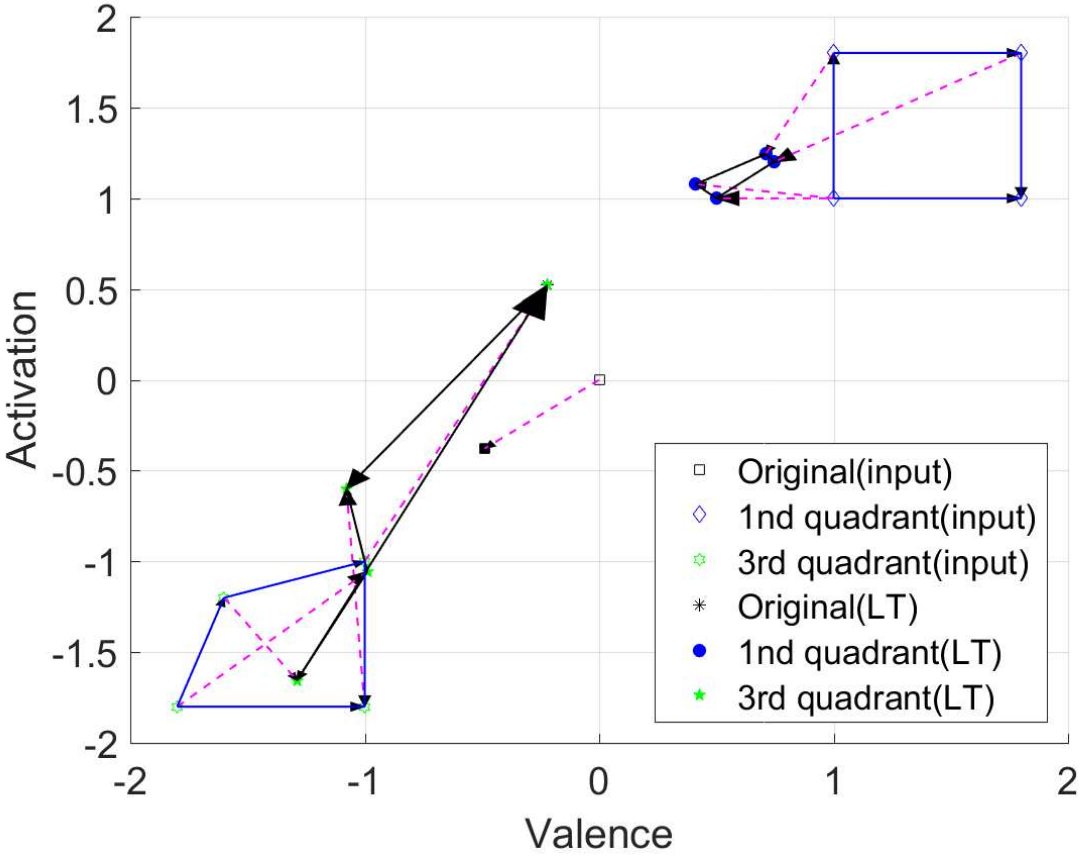

Figure 6.7: Evaluation results (emotion degree) of synthesized voices in English.

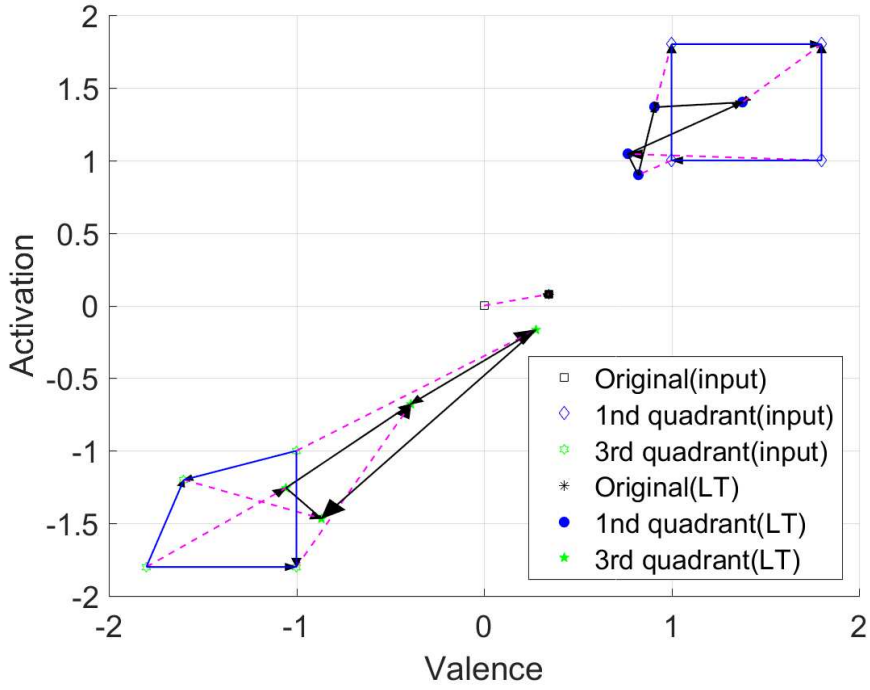

Figure 6.8: Evaluation results (emotion degree) of synthesized voices in Japanese

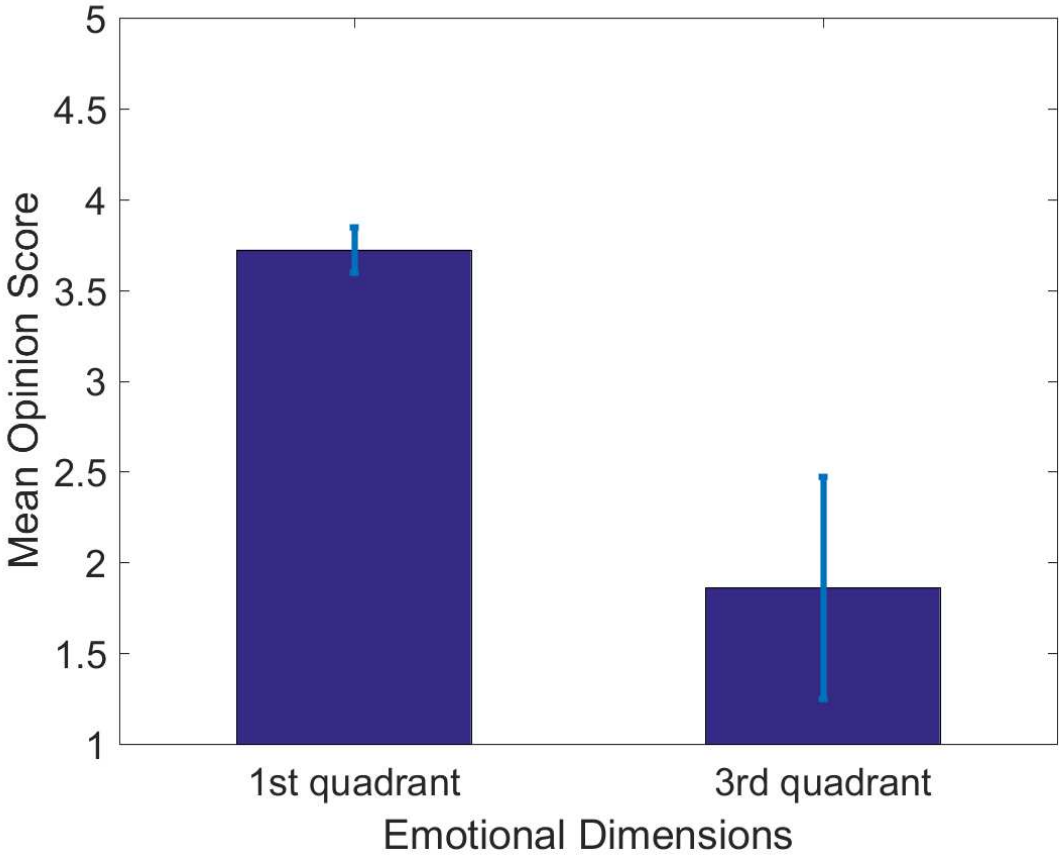

Figure 6.9: Evaluation results (The mean opinion score) of Chinese synthesized voices.

in order to avoid conceptual confusion. Valence and activation were evaluated using 40 scales (Valence: Left [Very Negative], Right [Very Positive]; Activation: Left [Very Calm], Right [Very Excited]: range $-2 \sim 2$ in increments of 0.1). Subjects evaluated these scales using a graphic user interface. During the listening test, subjects could listen to the stimulus as many times as they wanted.

For naturalness, all synthesized voices were presented once before subjects gave evaluations. The scale of evaluations was divided into five levels from bad to excellent ($1 \sim 5$). Subjects gave evaluations according to original speech spoken by a human whose naturalness is excellent.

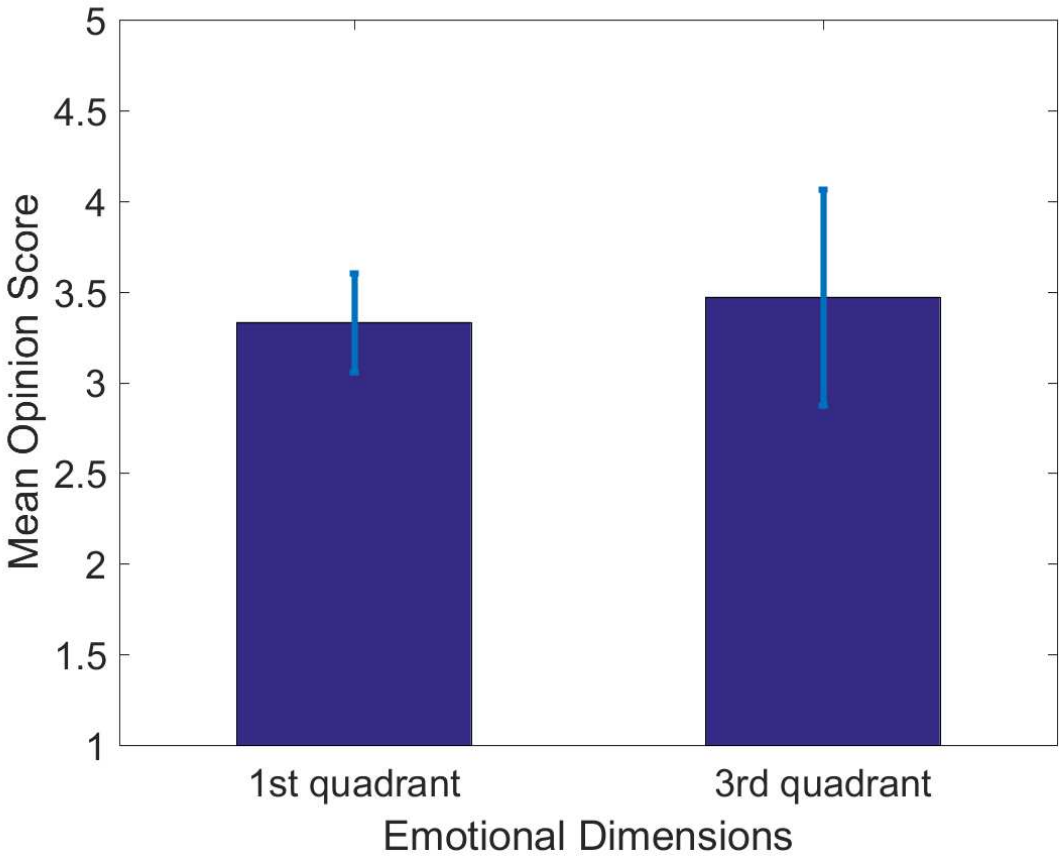

Figure 6.10: Evaluation results (The mean opinion score) of English synthesized voices.

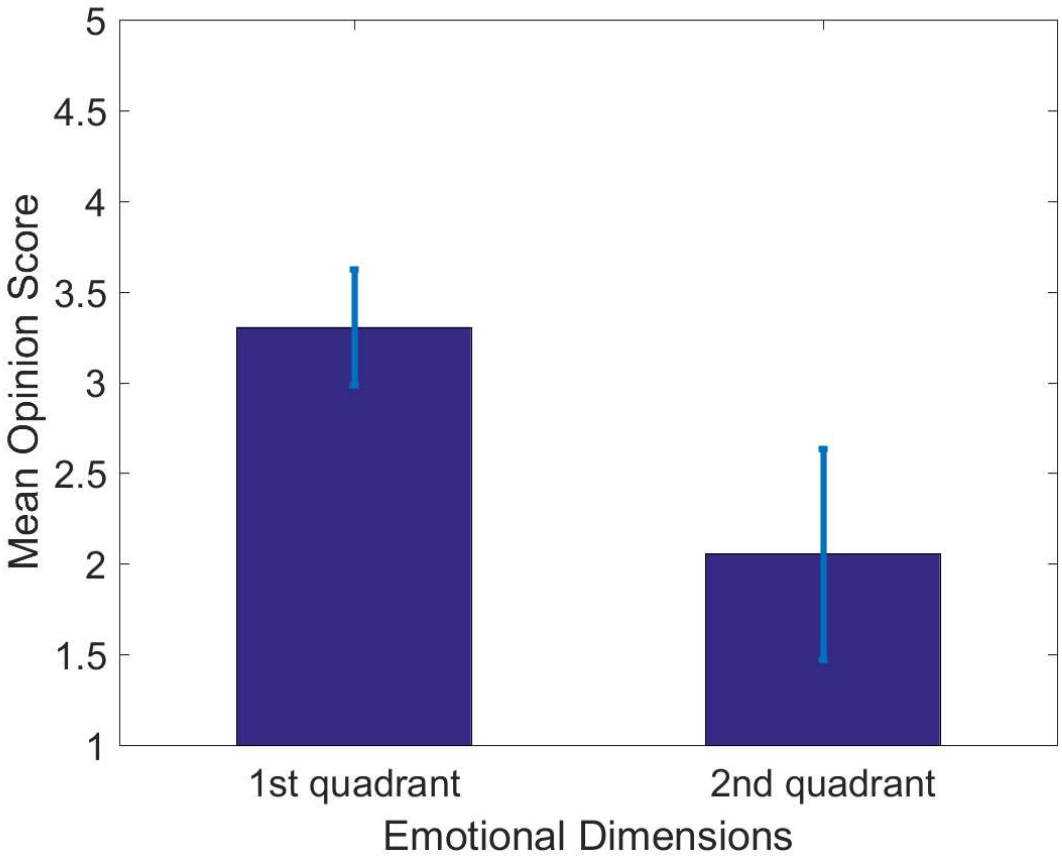

Figure 6.11: Evaluation results (The mean opinion score) of Japanese synthesized voices.

6.2.2 Results

Emotion Perception

In Fig. 6.3, Fig. 6.4 and Fig. 6.5, evaluated positions in the valence and activation space are shown with solid line in each quadrant among multiple languages. The oval is calculated using average and standard deviation in each quadrant. Here, the dashed lines are the inputs of the system and the solid lines are the evaluated results by subjects in the listening test. The dashed lines are what we want, and the solid lines are what we actually obtained. The blue lines represent the joy voices and the green lines show the sad voices. As among three languages, blue lines are all in the first quadrant and green lines are all in the third quadrant. It means that subjects can perceive emotions category well among three languages.

In Fig. 6.6, Fig. 6.7 and Fig. 6.8, 9 stimuli with either largest or smallest value of valence or activation are used to represent the intended and obtained intensity of emotion in each quadrant. The red dashed line shows directions from the intended positions to obtained positions and the rectangles are the intended intensity and the quadrilaterals are the obtained intensity from the listening test in the V-A space. We can see that the quadrilaterals in the first quadrant show a small area but the tendency of emotional intensity is the same except one valence values of Japanese voices. And the quadrilaterals in the third quadrant shows a large area and the tendency of emotional intensity are similar as intended except for the synthesized voice whose intended position is VA(-1.6,-1.2) which is caused by the estimation part. From the results of emotion category and degree, we can confirm that this system can convert neutral speech to emotional ones with the same intended category and similar intensity for multiple languages.

Naturalness

The evaluation result for the naturalness of synthesized speech is shown in Fig. 6.9, Fig.6.10 and Fig.6.11. Mean opinion score (MOS) of each quadrant is calculated separately. From these results, we can see that all naturalness scores are above or near 2, that means not bad. The excellent synthesized speech in terms of naturalness was Chinese joy voices. MOS of English synthesized voices are all above 3, that means ordinary natural and naturalness of sad voices are low for Chinese and Japanese. The reason why sadness

was not good is because the duration control of sad speech sometimes fails in some points so that the synthesized voices were long but the interval in each phrase was not obvious. Therefore, the synthesized speech seemed like machine-like. More precise control of duration ratios between voiced and unvoiced periods is needed to be researched.

6.3 Discussion

These results show given the same direction from neutral speech to emotional states in V-A space, the conversion system built with one language can convert neutral speech to emotional ones among multiple languages. The conversion system for emotion trained in Japanese is used to convert a tone language, Chinese and a stress language, English. The results from the listening tests by subjects confirmed that the synthesized voices can convey the same category and similar intensity of emotion among different languages which means that the conversion system for emotion built for Japanese is compatible to other languages.

Chapter 7

Discussion

In the previous section, the modeling and representation of the non-linguistic information, emotion is considered. In order to convert neutral speeches to emotional ones with emotion degree-controllable following human emotion perception mechanism, a rule-based voice conversion system with emotions represented in dimensional space using the inverse three-layered model has already been proposed.

In Section 4, with the purpose of investigating the related acoustic features to each emotion dimensions (arousal and valence), the perception experiment was conducted by evaluating the synthesized voices with one acoustic feature replaced from neutral to emotional utterances. Four acoustic features: F0, spectral sequence, power envelope and duration were considered. The statistical method, ANOVA is used for analyzing the perception results. It is shown that F0 and spectral sequence make a great contribution to both dimensions while power envelope and duration are important to arousal dimension. These findings are in lines with the previous researches which focus on using the statistical method between acoustic features and emotion dimensions [25] [81].

Perception results in dimensional space show that only replacing F0 from emotional utterance to neutral utterance, joy and sad emotion can be perceived as original emotion category although the degree of emotion is decreased. For hot anger emotion, it is found that only replacing F0, it will be perceived in the first quadrant. On the other hand, if the only spectral sequence is replaced from neutral to emotional, joy, sad and anger emotion can be perceived as original and the degree of emotion is also decreased.

In this research prosody-related acoustic features, duration, F0 and power envelope was

controlled by interpolation method, Fujisaki model and target prediction model. Results of listening test shown in Chapter 5 reveal that the evaluated position of joy and sad emotion are all in the first and third quadrant separately. And the hot anger emotion is perceived in the first quadrant. This result is in line with the findings in Chapter 4. It indicates that the proposed method achieved the success following the purpose of this research.

Based on this idea, the system is applied to multiple languages without training. And the input positions are in the first and third quadrant in the dimensional space. Perception results show that this system trained in one language is suitable for multiple languages by modifying the prosody-related acoustic features.

On the other hand, the influence of emotions on human speech is manifested mainly in prosody, however, emotional state of a speaker is accompanied also by physiological changes causing a shift of individual formants, different amount of low-frequency and high frequency energy. Results in Chapter 4 reveals that not only the prosody related acoustic features, spectral sequence also plays an important role in emotional speech, especially for anger emotion. So the controlling of spectral parameters by parameterizing spectrum utilizing temporal decomposition and Gaussian mixture model [111] [147] [148] for emotional voice conversion is tried in this research.

Temporal decomposition is used to represent the continuous variation of the speech events as a linear-weighted sum of a number of discrete elementary components. Gaussian components in Gaussian mixture model are utilized to represent the distribution based on the spectral envelope. By modifying the parameters of a Gaussian component, the amplitude of spectral and formant frequency can be controlled well.

The satisfactory naturalness of resynthesized speech shows that spectrum can be parameterized well by using temporal decomposition and Gaussian mixture model. While when a large degree is applied for modification, the naturalness of the synthesized speech is drastically decreased. That means more works are needed for controlling the spectral sequence.

In this research, the prosody-related acoustic features were explored. The successful perception results confirm that the proposed method is suitable for controlling the degree of emotion in voice conversion system. In the future, the optimization method for con-

trolling the spectral sequence to fit the desired acoustic features related to spectrum will be researched. The hypothesis can be made that the complete voice conversion system for all kinds of emotion and non-linguistic information can be built if the spectral sequence can be controlled well. Furthermore, the combination of prosody and spectral related acoustic features will make this system capable for multiple languages from any speaker.

Chapter 8

Discussion and Conclusion

8.1 Discussion

In the previous section, the modeling and representation of the non-linguistic information, emotion is considered. In order to convert neutral speeches to emotional ones with emotion degree-controllable following human emotion perception mechanism, a rule-based voice conversion system with emotions represented in dimensional space using the inverse three-layered model has already been proposed.

In Section 4, with the purpose of investigating the related acoustic features to each emotion dimensions (arousal and valence), the perception experiment was conducted by evaluating the synthesized voices with one acoustic feature replaced from neutral to emotional utterances. Four acoustic features: F0, spectral sequence, power envelope and duration were considered. The statistical method, ANOVA is used for analyzing the perception results. It is shown that F0 and spectral sequence make a great contribution to both dimensions while power envelope and duration are important to arousal dimension. These findings are in lines with the previous researches which focus on using the statistical method between acoustic features and emotion dimensions [25] [81].

Perception results in dimensional space show that only replacing F0 from emotional utterance to neutral utterance, joy and sad emotion can be perceived as original emotion category although the degree of emotion is decreased. For hot anger emotion, it is found that only replacing F0, it will be perceived in the first quadrant. On the other hand, if the only spectral sequence is replaced from neutral to emotional, joy, sad and anger emotion

can be perceived as original and the degree of emotion is also decreased.

In this research prosody-related acoustic features, duration, F0 and power envelope was controlled by interpolation method, Fujisaki model and target prediction model. Results of listening test shown in Chapter 5 reveal that the evaluated position of joy and sad emotion are all in the first and third quadrant separately. And the hot anger emotion is perceived in the first quadrant. This result is in line with the findings in Chapter 4. It indicates that the proposed method achieved the success following the purpose of this research.

Based on this idea, the system is applied to multiple languages without training. And the input positions are in the first and third quadrant in the dimensional space. Perception results show that this system trained in one language is suitable for multiple languages by modifying the prosody-related acoustic features.

On the other hand, the influence of emotions on human speech is manifested mainly in prosody, however, emotional state of a speaker is accompanied also by physiological changes causing a shift of individual formants, different amount of low-frequency and high frequency energy. Results in Chapter 4 reveals that not only the prosody related acoustic features, spectral sequence also plays an important role in emotional speech, especially for anger emotion. So the controlling of spectral parameters by parameterizing spectrum utilizing temporal decomposition and Gaussian mixture model [111] [147] [148] for emotional voice conversion is tried in this research.

Temporal decomposition is used to represent the continuous variation of the speech events as a linear-weighted sum of a number of discrete elementary components. Gaussian components in Gaussian mixture model are utilized to represent the distribution based on the spectral envelope. By modifying the parameters of a Gaussian component, the amplitude of spectral and formant frequency can be controlled well.

The satisfactory naturalness of resynthesized speech shows that spectrum can be parameterized well by using temporal decomposition and Gaussian mixture model. While when a large degree is applied for modification, the naturalness of the synthesized speech is drastically decreased. That means more works are needed for controlling the spectral sequence.

In this research, the prosody-related acoustic features were explored. The successful

perception results confirm that the proposed method is suitable for controlling the degree of emotion in voice conversion system. In the future, the optimization method for controlling the spectral sequence to fit the desired acoustic features related to spectrum will be researched. The hypothesis can be made that the complete voice conversion system for all kinds of emotion and non-linguistic information can be built if the spectral sequence can be controlled well. Furthermore, the combination of prosody and spectral related acoustic features will make this system capable for multiple languages from any speaker.

8.2 Conclusion

A voice conversion system for emotional speech which utilized dimensional space to represent emotion in order to control the degree of emotion following the human emotional speech perception mechanism is proposed in this research.

For emotion representation, two dimensions, valence(from positive to negative) and arousal (from excited to calm), are considered. So the degree and category of emotion can be freely controlled by changing the position values in dimensional space.

For modeling process of emotion, the inverse three-layered model is proposed as the structure between emotion dimensions and acoustics. The inverse three-layered model follows the Brunswik's functional lens model which assumes that the perception of emotion by humans is multi-layered.

For the method of synthesizing emotional speech, rule-based emotional voice conversion system is proposed which can obtain the rules among different degree of emotions using limited training database.

The significant acoustic features related to each dimension are explored by synthesizing speech, certain acoustic features of which are from emotional speech and others, from neutral speech. Perceptual evaluations in V-A space show that F0 and spectral information are the most important factors related to arousal and valence.

By replacing the F0 and spectral information of neutral speech from joyful, cold anger and sad emotional speech, the synthesized speech can be perceived as having the same original emotional category, although the degree is decreased by replacing either of them. But by replacing only the F0 of the neutral utterance to the F0 from the hot anger utterance, the synthesized utterance is perceived as a joyful emotion. If only spectral

information is replaced by that from the hot anger utterance, the synthesized voice can be perceived as hot anger while the degrees in both valence and arousal dimensions are decreased. These results support the previous studies that voice quality and F0 contribute much to emotions [86] [87].

The voice conversion system has two parts: rule extraction and rule application. ANFIS, which embraces the concept of human perception of emotion as fuzzy logic, connects the three layers as a non-linear mapping. The low mean absolute error between the estimated value from ANFIS and the reference shows that ANFIS and the inverse three-layered model has the ability to build the non-linear relationship between acoustics and the emotion dimensions. The rules of acoustic features for modifying the neutral speech are extracted using the estimated acoustic features from ANFIS and the extracted acoustic features from neutral speech. In order to convert the neutral speech to the desired emotional speech in dimensional space, the Fujisaki model and target prediction model for parameterizing F0 and power envelope separately are conducted. STRAIGHT is used as the analysis-synthesis tool in this system.

Perceptual evaluation results in V-A space show that the synthesized speech of joyful, sad and cold anger emotion can be perceived well, including the category and the degree, although the perceived degree is decreased compared to the desired values. For hot anger emotion, since spectral modification was not conducted, the synthesized speech of hot anger is perceived as a joyful emotion.

The possibility for extended this emotional conversion system build in one language is explored lastly. Given the same direction from neutral speech to emotional states in V-A space, the conversion system for emotion can convert neutral speech to emotional ones among multiple languages. The conversion system for emotion trained in Japanese is used to convert a tone language, Chinese and a stress language, English. The results from the listening tests by subjects confirmed that the synthesized voices can convey the same category and similar intensity of emotion among different languages which means that the conversion system for emotion built for Japanese is compatible with other languages.

8.3 Contribution

In this research, the significant contribution is to apply the dimensional approach for representing emotion in emotion conversion system instead of the categorical method. Hence the degree of emotion can be controlled freely by changing the position in dimensional space.

Then the propose of the inverse three-layered model following the emotion perception of human beings makes the modified Brunswikin's lens model applicable. And it provides a new structure between acoustic and emotion.

Lastly, the rule-based speech synthesis system realized by the fuzzy inference system independence of the language and speaker, it can make the conventional TTS synthesize emotional speech of the target speaker and target language using only the neutral speech of the speaker in any language. It will have a great impact on practical TTS applications in the future.

8.4 Future works

Exploring the individuality among speakers

All these findings are based on one female voice actress database. Yet, speakers encode their affective states using various acoustic features. In future work, the database will be extended to multiple speakers in order to explore speaker individuality for affectiveness.

Exploring the differences and commonality among languages

This system is trained only in one language. For the purpose to build the affective voice conversion system for multiple languages, the training database can be extended to more kinds of languages. And the commonality and difference among languages can be explored in the future.

Controlling the spectral sequence

The current system can control the prosody related acoustic features. While the spectral sequence is also important for emotional speech. More efforts will be made on temporal

decomposition and Gaussian mixture model for controlling spectral sequences in order to convert neutral speech to any kind of affective speech.

Building the affective voice conversion system for multiple languages

As three kinds of information are embedded in speech, other affective factors such as emphasis, shouted will be researched in the future for building the complete affective voice conversion system.

Lastly, previous research has already revealed the commonality and difference in cross-cultural emotion perception [88] [89] [90]. Since this system does not have a restriction on linguistic information, this will be a good approach for exploring the applications to multiple languages and multiple speakers in the future.

Bibliography

[1] H. Fujisaki, "Information, prosody, and modeling-with emphasis on tonal features of speech," *Proc. Speech Prosody,* pp. 1-10, 2004.

[2] J. Tao and T. Tan, "Affective computing: A review," *International Conference on Affective computing and intelligent interaction. Springer Berlin Heidelberg,* pp. 981-995, 2005.

[3] R. W. Picard, "Affective computing," MIT Press, Cambridge, 1997.

[4] D. Erickson, "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Science and Technology,* vol. 26, no.4, pp. 317-325, 2005.

[5] A. Rilliard, T. Shochi, J. C. Martin, D. Erickson and V. Aubergé, "Multimodal indices to Japanese and French prosodically expressed social affects," *Language and speech,* vol. 52, no. 2-3, pp. 223-243, 2009.

[6] T. Shochi, A. Rilliard, V. Aubergé and D. Erickson, "Intercultural Perception of English, French and Japanese Social Affective Prosody," *The role of prosody in Affective Speech*, pp. 31, 2009.

[7] M. Schröder, "Expressive speech synthesis: Past, present, and possible futures," *Affective information processing, Springer*, London, pp. 11-126, 2009.

[8] M. Schrder. "Emotional speech synthesis: a review". *Proc: INTERSPEECH* pp. 561-564, 2001.

[9] M. Akagi, X. Han, R. Elbarougy, Y. Hamada, and J. Li. "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in

multiple languages," *APSIPA 2014 - Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference, December 912 Siem Reap, Cambodia Proceedings*, pp. 1-10, 2014.

[10] W. Wahlster, "Verbmobil: foundations of speech-to-speech translation," *Springer Science & Business Media*, 2000.

[11] S. Nakamura, K. Markov, H. Nakaiwa, G. I. Kikui, H. Kawai, T. Jitsuhiro and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no.2, pp. 365-376, 2006.

[12] B. Zhou, X. Cui, S. Huang, M. Cmejrek, W. Zhang, J. Xue and S. Maskey, "The IBM speech-to-speech translation system for smartphone: Improvements for resource-constrained tasks," *Computer Speech & Language*, vol. 27, no. 2, pp. 592-618, 2013.

[13] V. Arranz, E. Comelles, D. Farwell, C. Nadeu, J. Padrell, A. Febrer and K. Peterson, "A speech-to-speech translation system for Catalan, Spanish, and English," *In Conference of the Association for Machine Translation in the Americas,*" pp. 7-16. Springer, Berlin, Heidelberg, 2004.

[14] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis," *IEEE Trans Speech Audio Proc*, vol. 7, pp. 2401-2404, 2003.

[15] J. Tao, Y. Kang and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no.4, pp. 1145-1154, 2006.

[16] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken English," *Speech Communication,* vol. 51, no. 3, pp. 268-283, 2009.

[17] R. Aihara, R. Takashima, T. Takiguchi and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing,* vol.2, no.5, pp. 134-138, 2012.

[18] J. Yadav and K. Rao, "Prosodic mapping using neural networks for emotion conversion in Hindi language," *Circuits, Systems, and Signal Processing,* vol. 35, no.1, pp. 139-162, 2016.

[19] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features," *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on. IEEE,* pp.1-5, 2016.

[20] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication,* vol. 40, no.1, pp. 5-32, 2003.

[21] B. Fehr, J. Russell, "Concept of emotion viewed from a prototype perspective," *Journal of experimental psychology: General,* vol. 113, no.3, pp. 464-486, 1984.

[22] K. Scherer, P. Ekman, "On the nature and function of emotion: A component process approach," *Approaches to emotion,* vol. 2293, pp. 317, 1984.

[23] R. E. Plutchik, H. R. Conte, "Circumplex models of personality and emotions," American Psychological Association, 1997.

[24] R. Banse, K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology,* vol. 70, no. 3, pp. 614, 1996.

[25] M. Schröder, "Expressing degree of activation in synthetic speech," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 4, pp. 1128-1136, 2006.

[26] M. Schröder and M. Grice, "Expressing vocal effort in concatenative synthesis," *Proc. 15th International Conference of Phonetic Science,* pp. 797-800, 2003.

[27] H. Schlossberg, "Three dimensions of emotion," *Psychological review,* vol. 61, no. 2, pp. 81-88, 1954.

[28] M. Grimm and K. Kroschel, "Emotion estimation in speech using a 3d emotion space concept," *Robust Speech Recognition and Understanding. InTech,* 2007.

[29] K. Scherer, "Vocal communication of emotion: A review of research paradigms, " *Speech communication,* vol. 40, no.1, pp. 227-256, 2003.

[30] K. Scherer and R. Klaus, "Methods of research on vocal communication: Paradigms and parameters," *Handbook of methods in nonverbal behavior research,* pp. 136-198, 1982.

[31] A. Kappas, U. Hess and K. Scherer, "Voice and emotion," *Fundamentals of nonverbal behavior,* vol. 200, 1991.

[32] T. Bänziger, G. Hosoya and K. Scherer,"Path Models of Vocal Emotion Communication," *PLoS ONE,* vol. 10, no. 9: e0136675. pone.0136675.

[33] E. Brunswik, "Historical and thematic relations of psychology to other sciences," *Scientific Monthly,* vol. 83, pp: 151161, 1956.

[34] C. F. Huang and M. Akagi, "A three-layered model for expressive speech perceptionh," *Speech Communication* , vol. 50, no. 10, pp. 810-828, 2008.

[35] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical science and technology,* vol. 35, no.2, pp. 86-98, 2014.

[36] X. Li and M. Akagi, "Multilingual Speech Emotion Recognition System based on a Three-layer Model," *Prof. Interspeech2016*, pp. 3608-3612, 2016.

[37] R. Barra-Chicote, J. Yamagishi, S. King and J. M. Montero, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication,* vol. 52, no. 5, pp. 394-404, 2010.

[38] R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication,* vol. 52, no. 5, pp. 394-404, 2010.

[39] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *In Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings, 1996 IEEE International Conference,* vol. 1, pp. 373-376. 1996.

[40] M. Bulut, S. Narayanan and A. K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," *In Seventh International Conference on Spoken Language Processing,* 2002.

[41] E. Rank and H. Pirker, "Generating emotional speech with a concatenative synthesizer," *In Fifth International Conference on Spoken Language Processing,* 1998.

[42] J. Jia, S. Zhang, F. Meng, Y. Wang and L. Cai, "Emotional audio-visual speech synthesis based on PAD," *IEEE transactions on audio, speech, and language processing,* vol. 19, no. 3, pp. 570-582, 2011.

[43] H. R. Pfizinger, N. Amir, H. Mixdorff, and J. Bsel, "Cross-language Perception of Hebrew and German Authentic Emotional Speech," *International Congress of Phonetic Sciences,* Hong Kong, 17-21, August, 2011.

[44] J. Yamagishi, T. Nose, H. Zen, Z. Ling, T. Toda, K. Tokuda and S. Renals, "Robust speaker-adaptive HMM-based text-to-speech synthesis," *IEEE Transactions on, Audio, Speech, and Language Processing,* vol. 17, no. 6, pp. 1208-1230, 2009.

[45] T. Toda and K. S. Tukuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems,* vol. 90, no. 5, pp. 816-824, 2007.

[46] P. Birkholz, L. Martin, K. Willmes, B. J. Kröger and C. Neuschaefer-Rube, "The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study," *The Journal of the Acoustical Society of America,* vol. 137, no. 3, pp. 1503-1512, 2015.

[47] Heiga Zen, <http://rtthss2015.talp.cat/download/RTTHSS2015_Zen.pdf>.

[48] R. Elbarougy and M. Akagi, "Cross-lingual Speech Emotion Recognition System Based on a Three-Layer Model for Human Perception," *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific,* IEEE, 2013.

[49] X. Li, and M. Akagi, "Multilingual Speech Emotion Recognition System Based on a Three-Layer Model," *INTERSPEECH,*, pp. 3608-3612 2016.

[50] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *Plos One,* vol. 8, no. 4, e60603, 2013.

[51] P. Birkholz, B. J. Kröger and C. Neuschaefer-Rube, "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a modified two-mass model of the vocal fols," *Interspeech,* pp. 2681-2684, 2011.

[52] P. Birkholz, B. J. Kröger and C. Neuschaefer-Rube, "Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal fold," *First International Workshop on Performative Speech and Singing Synthesis, 2011, Vancouver, BC, Canada.*

[53] P. Birkholz, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulation of the vocal tract system," *Interspeech,* pp. 1125-1128, 2004.

[54] P. Birkholz, "Control of an articulatory speech syntehsizer based on dynamic approximation of spatial articulatory targets," *Interpseech,* pp. 2865-2868, 2007.

[55] P. Brikholz, B. J. Kröger and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant-vowel sequences," *IEEE Trans. Audio Speech Lang. Process.* pp. 1422-1433, 2010.

[56] I. R. Murray, M. D. Edgington, D. Campion and J. Lynn, "Rule-based emotion synthesis using concatenated speech," *In ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion,* 2000.

[57] E. Zovato, A. Pacchiotti, S. Quazza and S. Sandri, "Towards emotional speech synthesis: A rule based approach," *In Fifth ISCA Workshop on Speech Synthesis,*2004.

[58] Y. Li, K. Sakakibara, D. Morikawa and M. Akagi, "Commonalities of glottal sources and vocal tract shapes among speakers in emotional speech," *In The 11th International Seminar on Speech Production (ISSP 2017), Tianjin, China.*

[59] S. Lee, S. Yildirim, A. Kazemzadeh and S. Narayanan, "An articulatory study of emotional speech production," *In Ninth European Conference on Speech Communication and Technology, 2005.*

[60] J. Dang, A. Li, D. Erickson, A. Suemitsu, M. Akagi, K. Sakuraba, N. Minematsu and K. Hirose, "Comparison of Emotion Perception among Different Cultures," *Acoustic Science and Technology*, vol. 31, no. 6, 2010.

[61] P. Birkholz, [software], <http://www.vocaltractlab.de/index.php?page=vocaltractlab-download>

[62] T. Toda, L. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu and J. Yamagishi, "The voice conversion challenge 2016," in *Proc. Interspeech2016*, pp. 1632-1636, 2016.

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

[64] S. Takamichi, T. Toda, A. Black, G. Neubig, S. Salto. and S. nakamura, "Post-filter to modify the modulation spectrum for statistical parametric speech synthesis," *Audio, Speech and Language Processing, IEEE/ACM Transactions,* vol. 18, no. 5, pp. 1006-1010, 2015.

[65] D. Erro, A. Moreno and A. Bonafonte, "Voice conversion based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 18, no. 5, pp. 922-931, 2010.

[66] D. Erro, E. Navas and I. Hernez, "Emotion conversion based on prosodic unit selection," *IEEE Transactions on Audio, Speech, and Language Processing,* vol.18, no.5, pp. 974-983, 2010.

[67] A. Iida, N. Campbell, F. Higuchi and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Communication,* vol. 40, no.1, pp. 161-187, 2003.

[68] J. Dang, et al."Comparison of emotion perception among different cultures," *Acoustical Science and Technology,* vol,31, no. 6, pp. 394-402, 2010.

[69] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese," *Journal of the Acoustical Society of Japan (E),* vol. 5, no. 4, pp. 233-242, 1984.

[70] M. O'Reilly and A. Chasaide, "Analysis of intonation contours in portrayed emotions using the Fujisaki model," *The Second International Conference on Affective Computing and Intelligent Interaction. Proceedings of the Doctoral Consortium,* 2007.

[71] M. Akagi and Y. Tohkura, "Spectrum target prediction model and its application to speech recognition," *Computer Speech & Language,* vol. 4, no. 4, pp. 325-344, 1990.

[72] H. Kawahara, I. Masuda-Katsuse and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sound," *Speech communication,* vol. 27, no. 3, pp. 187-207, 1999.

[73] C.M. Whissell, "A dictionary of affect in language," *Perceptual and Motor Skills,* vol. 62, no.1 pp. 127-132, 1986.

[74] K. Scherer, "Emotion as a multicomponent process: A model and some cross-cultural data," *Review of Personality & Social Psychology (1984).*

[75] N. Takashi, J. Yamagishi and T. Masuko, "A style control technique for HMM-based expressive speech synthesis," *IEICE TRANSACTIONS on Information and Systems,* vol. 90, no. 9, pp. 1406-1413, 2007.

[76] F. Biassoni, S. Balzarotti and M. Giamporcaro, "Hot or cold anger? Verbal and vocal expression of anger while driving in a simulated anger-provoking scenario," *SAGE Open,* vol. 6, no. 3, 2016.

[77] P. Juslin, "Cue utilization in communication of emotion in music performance: Relating performance to perception," *Journal of Experimental Psychology: Human perception and performance,* vol. 26, no.6, pp. 1797, 2000.

[78] S. BIERSACK and V. KEMPE, "Tracing Vocal Expression of Emotion Along the Speech Chain: Do Listeners Perceive What Speakers Feel?" *In: ISCA Workshop on Plasticity in Speech Perception. 2005.*

[79] P. B. Denes and E. Pinson, E, "The speech chain,". *Macmillan*, 2015.

[80] M. Belyk, S. Brown, "The acoustic correlates of valence depend on emotion family," *Journal of Voice,* vol. 28, no. 4, pp. 523. e9-523.e18, 2014

[81] M. Schröder, R. Cowie and E. Douglas-Cowie, "Acoustic correlates of emotion dimensions in view of speech synthesis," *Seventh European Conference on Speech Communication and Technology. 2001.*

[82] Y. Xue, Y. Hamada, and M. Akagi, "Emotional speech synthesis system based on a three-layered model using a dimensional approach," *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific,* IEEE, pp. 505-514, Hongkong, 2015.

[83] J-SR. Jang, , "ANFIS: adaptive-network-based fuzzy inference system," *IEEE transactions on systems, man, and cybernetics,* vol. 23, no.3, pp. 665-685, 1993.

[84] H. Mixdorff, "A novel approach to the fully automatic extraction of Fujisaki model parameters," *Proc. ICASSP*, Istanbul, Turkey, pp. 1281-1284, 2000.

[85] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion to emotional speech based on three-layered model in dimensional approach and parameterization of dynamic features in prosody," *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific,* IEEE, pp.1-6, 2016.

[86] I. Grichkovtsova, M. Morel and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication,* vol. 54, no. 3, pp. 414-429, 2012.

[87] C. Gobl and A. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication,* vol. 40, no. 1, pp. 189-212, 2003.

[88] X. Han, E. Reda and M. Akagi, "A study on perception of emotional states in multiple languages on Valence-Activation approach," *2015 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP 2015)*, pp. 86-89, 2015.

[89] N. Lim, "Cultural differences in emotion: differences in emotional arousal level between the East and the West," *Integrate Medicine Research,* vol. 5, no. 2, pp. 105-109 2016.

[90] A. Chen, C. Gussenhoven, and T. Rietveld, "Language-specificity in the perception of paralinguistic intonational meaning," *Language and Speech,* vol. 47, no. 4, pp. 311-349, 2004.

[91] R. W. Rieber and D. K. Robinson, " Wilhelm Wundt in history: The making of a scientific psychology, " *Springer Science & Business Media*, 2001.

[92] H. Scholsberg, "A scale for the judgment of facial expressions," *Journal of experimental psychology*, vol. 29, no.6, pp. 497, 1941.

[93] A. Mehrabian and J. A. Russell, "An approach to environmental psychology," *the MIT Press*, 1974.

[94] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32-80, 2001.

[95] P. Shaver, J. Schwartz, D. Kirson and C. O'connor, "Emotion knowledge: Further exploration of a prototype approach," *Journal of personality and social psychology,* vol. 52, no. 6, pp, 1061, 1987.

[96] C. A. Smith, (1989). Dimensions of appraisal and physiological response in emotion. Journal of personality and social psychology, vol. 56, no. 3, pp. 339.

[97] R. Plutchik, "Emotion: A Psychoevolutionary Synthesis, Harper and Row," *New York, 1980.*

[98] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467-487, 1978.

[99] E. Brunswik, "Perception and the representative design of psychological experiments," *Univ of California Press,* 1956.

[100] K. R. Hammond, "Probabilistic functioning and the clinical method," *Psychological review*, vol. 62, no. 4, pp. 255, 1955.

[101] R. Gifford, "A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior," *Journal of Personality and Social Psychology*, vol. 66, no. 2, pp. 398, 1994.

[102] D. A. J. Reynolds and R. Gifford, "The sounds and sights of intelligence: A lens model channel analysis," *Personality and Social Psychology Bulletin,* vol. 27, no. 2, pp. 187-200, 2001.

[103] Y. Xue, Y. Hamada, R. Elbarougy and M. Akagi,"Voice conversion system to emotional speech in multiple languages based on three-layered model for dimensional space," *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA),* pp. 122-127, 2016.

[104] Y. Hamada, Y. Xue and M. Akagi "Study on method to control fundamental frequency contour related to a position on Valence-Activation space," the *Western Pacific Commission for Acoustics (WESPAC)*, Singapore, P12000176, 2015.

[105] C. F. Huang, D. Erickson and M. Akagi, "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners," Acoustics2008, Paris, pp. 23172322, 2008.

[106] P. Shaver, J. Schwartz, D. Kirson and C. O'connor, "Emotion knowledge: Further exploration of a prototype approach," *Journal of personality and social psychology*, vol. 52, no. 6, pp. 1061, 1987.

[107] H. Gunes, B. Schuller, M. Pantic and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey. In Automatic Face & Gesture Recognition and Workshops", *2011 IEEE International Conference on IEEE.* pp. 827-834. IEEE.

[108] M. Schröder, "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions," *In Tutorial and research workshop on affective dialogue systems,* pp. 209-220. Springer, Berlin, Heidelberg, 2004.

[109] B. Parkinson, "Getting from situations to emotions: Appraisal and other routes."

[110] N. H. Frijda, "The place of appraisal in emotion," *Cognition & Emotion,* vol. 7, no. 3-4, pp. 357-387, 1993.

[111] T. N. Phung, T. S. Phan, T. Vu, M. C. Luong and M. Akagi, "Improving naturalness of HMM-based TTS trained with limited data by temporal decomposition," *IEICE TRANSACTIONS on Information and Systems"*, vol, 96, no. 11, pp. 2417-2426, 2013.

[112] S. Narayanan, K. Nayak, K, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America,* vol. 115, no. 4, pp. 1771-1776, 2004.

[113] P. Martins, I. Carbone, A. Pinto, A. Silva and A. Teixeira, "European Portuguese MRI based speech production studies," *Speech Communication,* vol. 50, no. 11, pp. 925-952, 2008.

[114] T. Ito, K. Takeda and F. Itakura, "Analysis and recognition of whispered speech," *Speech Comunication,* vol. 45, no. 2, pp. 139-152, 2005.

[115] J. C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication,* vol. 20, no. 1-2, pp. 13-22, 1996.

[116] T. Raitio, A. Suni, M. Vainio and P. Alku, "Analysis of HMM-based Lombard speech synthesis," *Interspeech,* pp. 2781-2884, 2011.

[117] D. Rostolland, "Intelligibility of shouted voice," *Acta Acustica united with Acustica,* vol. 57, no. 3, pp. 103-121, 1985.

[118] J. M. Pickett, "Effects of vocal force on the intelligibility of speech sounds," *The Journal of the Acoustical Society of America,* vol. 28, no. 5, pp. 902-905, 1956.

[119] Z. S. Bond, and J. Moore, "A note on loud and Lombard speech," *First International Conference on Spoken Language Processing.* 1990.

[120] D. Rostolland, "Acoustic features of shouted voice," *Acta Acustica united with Acustica,* vol. 57, no. 3, pp. 118-125, 1985.

[121] D. Rostolland, "Phonetic structure of shouted voice," *Acta Acustica united with Acustica,* vol. 51, no. 2, pp. 80-89, 1982.

[122] J. Elliott, "Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics," *Proc. SST-2000: 8th Int. Conf. Speech Sci. & Tech,* pp. 154-159, 2000.

[123] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio and P. Alku, "Analysis and synthesis of shouted speech," *Interspeech,* pp. 1544-1548, 2013.

[124] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication,* vol. 51, no. 12, pp. 1253-1262, 2009.

[125] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *The Journal of the Acoustical Society of America,* vol. 133, no. 5, pp. 3050-3061, 2013.

[126] V. K. Mittal and B. Yegnanarayana, "Production features for detection of shouted speech," *Consumer Communications and Networking Conference (CCNC),IEEE,* 2013.

[127] R. Schulman, "Articulatory dynamics of loud and normal speech," *The Journal of the Acoustical Society of America,* vol. 85, no. 1, pp. 295-312, 1989.

[128] M. Echternach, F. Burk, M. Burdumy, L. Traser and B. Richter, "Morphometric differences of vocal tract articulators in different loudness conditions in singing," *PloS one,* vol. 11, no. 4, pp. e0153792, 2016.

[129] P. Zelinka, M. sigmund and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Communication*, vol. 54, no. 6, pp. 732-742, 2012.

[130] J. Pohjalainen, P. Alku and T. Kinnunen, "Shout detection in noise," *Proc. IEEE Intr. Conf. Acoustics, Speech and Signal Processing (ICASSP),* pp. 4968-4971, 2011.

[131] B. Shikha, K. Banriskhem, M. Prasanna and P. Guha, "Shouted/Normal speech classification using speech -specific features," *Region 10 Conference (TENCON), IEEE*, 2016.

[132] H. Zen, K. Tokuda and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication,* vol. 51, no. 11, pp. 1039-1064, 2009.

[133] P. Boersma and D. Weenik, "Praat: doing phonetics by computer," [software]. <http://www.praat.org/>.

[134] P. Birkholz, Image3D [software]. <http://www.vocaltractlab.de/index.php?page=image3d-about>.

[135] M. Echternach, L. Trase and B. Richter, "Vocal tract configurations in tenors' passaggio in different vowel conditions- a real-time magnetic resonance imaging study," *Jounal of Voice,* vol. 28, no. 2, pp. 262.e1-262.e8, 2014.

[136] M. Echternach, P. Birkholz, J. Sundberg, L. Traser, J. G. Korvink and B. Richter, "Resonatory properties in professional tenors singing above the passaggio," *Acta Acustica united with Acustica,* vol. 102, pp. 298-306, 2016.

[137] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication,* vol. 36, pp. 169-180, 2002.

[138] P. Birkholz, VocalTractLab, [software]. <http://www.vocaltractlab.de/index.php?page=vocaltractlab-download>.

[139] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview." *Phonetica,* vol.49, no. 3-4, pp. 155-180, 1992.

[140] P. Birkholz, L. Martin, K. Willmes, B. J. Kroöger and C. Neuschaefer-Rube, "The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study," *The Journal of the Acoustical Society of America,* vol. 137, no. 3, pp: 1503-1512, 2015.

[141] B. Atal, "Efficient coding of LPC parameters by temporal decomposition". In Acoustics, Speech, and Signal Processing, IEEE International Conference on *ICASSP*, vol. 8 pp. 81-84, 1983.

[142] P. Zolfaghari, H. Kato, Y. Minami, A. Nakamura, S. Katagiri and S. Patterson, "Dynamic assignment of Gaussian components in modelling speech spectra." *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 45, no. 1-2, pp. 7-19.

[143] C. T. Sun, "Rule-base structure identification in an adaptive-network-based fuzzy inference system," *IEEE Transactions on Fuzzy Systems*, vol.2, no. 1, pp. 64-73.

Publications

Journal Paper

[1] Yawen Xue, Yasuhiro Hamada and Masato Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," Speech Communication. (Accept)

[2] Yawen Xue, Michael Marxen, Masato Akagi and Peter Birkholz "Acoustic and articulatory analysis and synthesis of shouted vowels," Speech Communication. (Revision request)

International Conference

[3] Yawen Xue, Yasuhiro Hamada, Reda Elbarougy, and Masato Akagi, Voice conversion to emotional speech based on three-layered model in dimensional approach and parameterization of dynamic features in prosody, Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific. IEEE, pp: 1-6, Jeju, Korean, Dec, 2016.

[4] Yawen Xue, Yasuhiro Hamada, and Masato Akagi, Voice conversion system to emotional speech in multiple languages based on three-layered model for dimensional space, International Conference on Speech Database and Assessments (Oriental COCOSDA). IEEE, Bali, Indonesia, Oct, 2016.

[5] Yawen Xue, Yasuhiro Hamada, Masato Akagi, Emotional voice conversion system for multiple languages based on three-layered model in dimensional space, The Journal of the Acoustical Society of America, vol.140(4), pp:2960-2960, Oct, 2016.

[6] Yawen Xue, and Masato Akagi, A study on applying target prediction model to parameterize power envelope of emotional speech, 2016 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (NCSP'16), Hawaii, USA, Mar, 2016.

[7] Yawen Xue, Yasuhiro Hamada, and Masato Akagi, A method for synthesizing emotional speech using the three-layered model based on a dimensional approach, ISURAC, Zao, Japan, May, 2016.

[8] Yawen Xue, Yasuhiro Hamada, Masato Akagi, Emotional speech synthesis system based on a three-layered model using a dimensional approach, Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific. IEEE, pp: 505-514, Hongkong, Dec, 2015.

[9] Yawen Xue, Yasuhiro Hamada, and Masato Akagi, Rule-based emotional voice conversion utilizing three-layered model for dimensional approach, Proceedings of the Taiwan/Japn Joint Research Meeting on Psychological Acoustics and Electroacoustics, pp. 583-588,Taiwan, Oct, 2015.

[10] Yasuhiro Hamada, Yawen Xue, Masato Akagi. "Study on method to control fundamental frequency contour related to a position on Valence-Activation space," WesPac, Singapore, Singapore, P12000176, 2015.

Domestic Conference

[11] Yawen Xue, Yasuhiro Hamada, Masato Akagi, A method for synthesizing emotional speech using the three-layered model based on a dimensional approach, Acoustic Society of Japan, Autumn, 2015.

[12] Yawen Xue, Yasuhiro Hamada, Masato Akagi, Apply the emotion conversion system based on three-layer model for dimensional space to other languages, Acoustic Society of Japan, Autumn, 2016.

[13] Yawen Xue, Masato Akagi, Acoustic features related to speaking styles in valence-arousal dimensional space, Acoustic Society of Japan, Spring, 2017.

Award

[14] The best student poster presenter in the International Symposium on Human Life Design. (Kanazawa, Japan, 03/2016).

[15] The best poster presenter in the 1st ISURAC (Zao, Japan, 05/2016).