Rule-based emotional voice conversion utilizing three-layered model for dimensional approach

By Yawen Xue

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

September, 2015

Rule-based emotional voice conversion utilizing three-layered model for dimensional approach

By Yawen Xue (1310201)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

and approved by
Professor Masato Akagi
Associate Professor Masashi Unoki
Professor Jianwu Dang

August, 2015 (Submitted)

Abstract

In the field of human-computer-interface (HCI), improving user experiences by providing genuine human communication with computer is one of the motivations. A speech-to-speech translation (S2ST) system plays a consequential role for converting a spoken utterance from one language into another using computer technology to enable people who speak different languages to communicate with each other. Conventional S2STs focus on processing linguistic information only, which is deficient in synthesizing affective speech, such as emotional rather than neutral speech. Therefore, a system that can recognize and synthesize emotional speech would be momentous. This research consider the emotional speech synthesis part which use the conversion method to get the emotional speech.

To construct a system for synthesizing emotional speech, several studies have already obtained some achievements. Most methods are based on a concatenative approach, like unit selection, or a statistical parametric approach, like the Hidden Markov Model (HMM) with the Gaussian Mixture Model (GMM). Both methods can synthesize emotional speech with good quality when the emotion is present in a category such as happy, sad, or angry. However, they can only synthesize the emotional speech with the average emotion intensity (not strong or weak emotions) in the emotion category, and both need a huge database for training, although it is difficult to collect many human responses when listening to emotional speech. In human speech communication, people sometimes strengthen or weaken emotional expressions depending on the situation. Thus, a small number of discrete categories is not sufficient to mimic the emotional speech in daily life.

Therefore, some researchers proposed a multi-dimension approach to express emotion on a continuous-valued scale instead of categorical methods. The advantage of the dimensional approach is not only that they can represent the emotion in all degrees but also providing a standard criterion for all research because the value of position in multi-dimensional is done by human experiment in the same dimension. And by using the rule-based synthesis method, tendencies of the variations in multi-dimensional space can be acquired using a small database. With the tendencies of variation, the synthesized speech can convey all degrees of an emotion.

The purpose of this study is to propose a emotional conversion system utilizing three-layered model for dimensional approach. The ultimate goal of our work is to improve the conventional emotional speech conversion system based on dimensional method in order to precisely predict the acoustic features as well as to synthesize affective speech much similar to what intended in dimensional space.

To improve the accuracy of estimated acoustic features, a three-layered model is adopted. According to the Branswikian lens model, people's emotion perception is multi-layered. Human beings do not perceive emotion directly from the acoustic features, so semantic primitives such as bright, high, strong, and so on are also of great importance. In these circumstances, this paper basically utilized the three-layered model proposed by Huang and Akagi (acoustic features layer, semantic primitives layer, and emotion layer). For the emotion layer, in this study, a dimensional emotion space is used to model the human

emotions. The Valence-Activation (V-A) axes in the two-dimensional emotion space can describe the strength, such as very or slightly happy, which gives a more flexible interpretation of emotional states. An Adaptive-Network-based Fuzzy Inference System (ANFIS) connects the three layers and estimates corresponding values of dimensional axes.

The modification of estimated acoustic features for emotional speech synthesis is revised following the work of Hamada, especially for the fundamental frequency (F0) related acoustic features. The Fujisaki model is adopted to extract the trajectory of the F0 contour from which we can obtain the F0 related acoustic features all at once. This work was done by Hamada and we adopt this method in three-layered model.

In this study, we improved the conventional emotional speech conversion system based on dimensional method in order to precisely predict the acoustic features as well as to synthesize affective speech much similar to what intended in dimensional space. The higher correlation coefficient comparing to the two-layered model [9] shows that three-layered model estimates acoustic features more accurately than the previous two-layered model. Results of subjective evaluations revealed that emotional speeches converted by three-layered model using new modification method, Fujisaki method can give the intended impression to a much similar degree as than the previous two-layered model in the emotion dimension. And the naturalness of converted speeches achieve an average score, 3.2 whose highest is 5 by subject evaluations. Above all, a conclusion can be made that an emotional conversion system utilizing three-layered model in dimensional approach can achieve better quality converted emotional speech than previous method.

Keywords: Emotional voice conversion, three-layered model, dimensional approach, fuzzy inference system.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

In the field of human-computer-interface (HCI), improving user experiences by providing genuine human communication with computer is one of the motivations. A speech-to-speech translation (S2ST) system plays a consequential role for converting a spoken utterance from one language into another using computer technology to enable people who speak different languages to communicate with each other [1].

Several information can be conveyed by speech communication. It can be roughly divided into three categories: linguistic information, paralinguistic information and nonlinguistic information [2]. Linguistic information represents the discrete categorical information explicitly represented by the written language or uniquely inferred from context. Linguistic information contains the lexical, syntactic, semantic, and pragmatic information .

Paralinguistic information shows the discrete and continuous information added by the speaker to modify or supplement the linguistic information such as attitude, intonation or intention. It can not be extracted from the linguistic information. One sentence can be uttered in different ways to express the intention, intonation or attitude of the speaker. For example, the same sentence "Tomorrow I want to go to library" uttered two different style: "*Tomorrow*, I want to go to library", "Tomorrow, I want to go to *library*". The emphasis part is different so the meaning has a bit change although the linguistic information is the same.

Nonlinguistic information gives the information not generally controlled by the speaker, such as the speaker's emotion, gender, age, etc although it is fact that a 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.

The procedure of S2ST is shown in as Figure 1.1. Firstly, there is a speech in Language A. Using Automatic Speech Recognition (ASR) system, the text in the speech can be acquired. Then, Translation System (TS) can translate the linguistic information in Language A into Language B. Lastly, Text To Speech (TTS) plays an important roles for

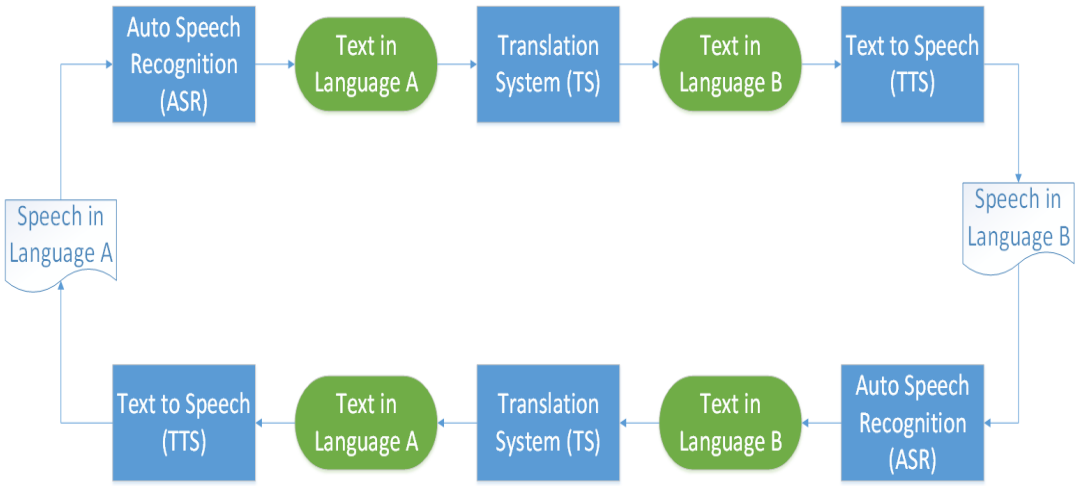

Figure 1.1: Schematic of speech-to-speech translation (S2ST) system

turn the text into speech which is spoken in Language B. This procedure is reversible.

Conventional S2STs focus on processing linguistic information only, which is deficient in synthesizing affective speech, such as emotional rather than neutral speech. While the three categories information conveyed from the speaker all play an important role for human interaction. Therefore, a system that can recognize and synthesize emotional speech would be momentous. This research consider the emotional speech synthesis part which use the conversion method to get the emotional speech.

1.2 Problems

To construct a system for synthesizing emotional speech, several studies have already obtained some achievements. Most methods are based on a concatenative approach, like unit selection, or a statistical parametric approach, like the Hidden Markov Model (HMM) with the Gaussian Mixture Model (GMM) [3] [4]. Both methods can synthesize emotional speech with good quality when the emotion is present in a category such as happy, sad, or angry. However, they can only synthesize the emotional speech with the average emotion intensity (not strong or weak emotions) in the emotion category, and both need a huge database for training, although it is difficult to collect many human responses when listening to emotional speech [5]. In human speech communication, people sometimes strengthen or weaken emotional expressions depending on the situation. Thus, a small number of discrete categories is not sufficient to mimic the emotional speech in daily life [6].

Therefore, some researchers proposed a multi-dimension approach to express emotion on a continuous-valued scale instead of categorical methods [7] [8]. The advantage of the dimensional approach is not only that they can represent the emotion in all degrees but also providing a standard criterion for all research because the value of position in

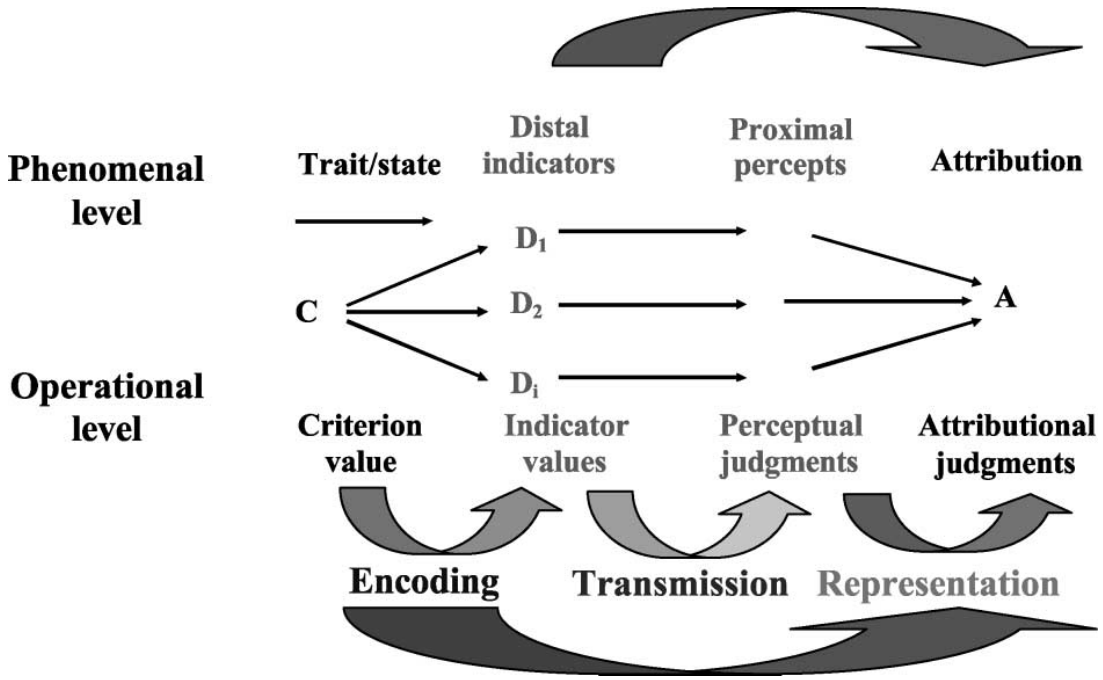

Figure 1.2: A Brunswikian lens (1956) model of the vocal communication of emotion. (Scherer, 2003) [10]

multi-dimensional is done by human experiment in the same dimension. And by using the rule-based synthesis method, tendencies of the variations in multi-dimensional space can be acquired using a small database. With the tendencies of variation, the synthesized speech can convey all degrees of an emotion.

1.3 Motivation and aims

The purpose of this study is to propose a emotional conversion system utilizing three-layered model for dimensional approach. The ultimate goal of our work is to improve the conventional emotional speech conversion system based on dimensional method in order to precisely predict the acoustic features as well as to synthesize affective speech much similar to what intended in dimensional space.

An emotional speech synthesis system based on dimensional approach has already been proposed [9]. However, they can only get low accuracy of estimating acoustic features and not similar affective speech as intended. One of the reasons is that the relationship they build is just from the acoustic features to emotion which not follow the perception of human as human can not exactly figure out the value of acoustic features but from some kinds of adjective to describe emotion. Another reason is that the modification method of the estimated acoustic features is not well-suitable.

To improve the accuracy of estimated acoustic features, a three-layered model is adopted.

According to the Branswikian lens model [10] shown in Figure 1.1, people's emotion perception is multi-layered. Human beings do not perceive emotion directly from the acoustic features, so semantic primitives such as bright, high, strong, and so on are also of great importance. In these circumstances, this paper basically utilized the three-layered model proposed by Huang and Akagi [12] [11](acoustic features layer, semantic primitives layer, and emotion layer). For the emotion layer, in this study, a dimensional emotion space is used to model the human emotions. The Valence-Activation (V-A) axes in the two-dimensional emotion space can describe the strength, such as very or slightly happy, which gives a more flexible interpretation of emotional states. An Adaptive-Network-based Fuzzy Inference System (ANFIS) [13] [14] connects the three layers and estimates corresponding values of dimensional axes.

The modification of estimated acoustic features for emotional speech synthesis is revised following the work of Hamada [15], especially for the fundamental frequency (F0) related acoustic features. Before [9] separately extracted and modified the F0 related acoustic features such as average F0 value, highest F0 value, F0 mean value of rising slope, and rising slope of the first accentual phrase in accordance with the extracted rules in the model. However, as there are strong connections among them, these F0 related acoustic features are unsuitable to handle independently. In this study, the Fujisaki model [16] is adopted to extract the trajectory of the F0 contour from which we can obtain the F0 related acoustic features all at once. This work is done in [15] and we adopt this method.

The difficulty in this research is to find out whether the extracted rules are suitable for the intended emotion as the relationship between the three layers is nonlinear. Correlation coefficient is used to calculate the difference between the estimated acoustic features and the extracted acoustic features. And listening test is done to test the affectiveness and naturalness of the synthesis speech.

The input and output of our system are the dimensional parameter values in V-A space and the corresponding acoustic feature displacements, respectively. ANFIS is used to connect the three layers from the top, acoustic feature layer , to the middle, semantic primitive layer, and from semantic primitive layer to the bottom, emotion space layer. The related acoustic features of every semantic primitive are selected when synthesizing the emotional speech. Listening tests were carried out to verify whether the synthesis speech can give a position similar to what anticipated. On the basis of the listening test, effectiveness is discussed.

1.4 Outline of the thesis

In order to give a clear belief introduction of this thesis, this section is for giving the structure of the thesis. The main methods is firstly introduced and then the database used in this study is explained. The procedure of estimating and modifying is considered which is followed by the evaluation of these part.At last, the conclusion will be given to show what have been achieved in this system. The thesis is organized as following :

- **Chapter 2** introduces the methods used in this system. Dimension approach gives the representation of emotion which is different from the category method. The

three-layered model is the basic of this system which shows the structure. And Fuzzy Inference System is the connection of the three-layered model which gives the estimation values of the according input.

- **Chapter 3** presents the outline of the system. The function of each part and the connection between each part.

- **Chapter 4** shows the elements of the system. The database and the procedure of obtaining the values of semantic primitives and acoustic features will be explained in detail. The process that all elements are normalized into the range from 0 to 1 is shown in this chapter.

- **Chapter 5** gives the two main parts of the system, the estimation and modification of acoustic features. The estimation of acoustic features using Fuzzy Inference System will be shown. Then the modification using the estimated acoustic features will be illustrated. After this part, the conversed emotional speech can be obtained when input the value in dimensional space.

- **Chapter 6** represents the evaluation of the proposed system. Correlation coefficient is used to evaluate the estimation part and listening test is done to evaluate the modification part.

- **Chapter 7** summarizes the contribution and achievements of the study. In addition, there are some remained problems in this study and in this part, the future work related to solve the remained problems will be shown.

Chapter 2

Research method

In this section, the methods used in this system will be explained in details. As our system is based on the three-layered model, it provides a bridge from the emotion perception by human to the speech signal following the process of human perception. It is the most important part which will be illustrated in Chapter 2.1. Then there is a problem comes out. Which method to use for describing emotion as emotion is the non-linguistic information convey by the speaker? This issue will be discussed in Chapter 2.2. Fuzzy Inference System (FIS) provides a estimation method from the input, the top layer, to the output, the bottom layer using three-layered model which will be shown in Chapter 2.3.

2.1 Three-layered model

The three-layered model is the structure of this system and its concept will be explained here. In 2008, Huang and Akagi proposed a three-layered model for expressive speech perception [12] [11] based on the Brunswik's lens model [10], which states that humans perceive the emotion of expressive speech not directly from the academic terms such as F0 contour, power envelop or power spectrum but from a series adjectives such as fast, bright, or strong. In addition, the process of human perception can be divided into two part: the first part is to perceive the degree of all adjectives which can be used to describe emotion and the second part is to perceive the degree of emotion from all the adjectives. On that case, the three-layered model is used to mimic human perception.

In this study, the emotional speech conversion system, the three-layered model is adopt according to the Figure 2.1. It consists of three layers: acoustic feature layer is at the top of the model, at the middle of the model is the semantic primitive layer and the emotion dimension layer is at the bottom of the model.

Acoustic feature layer is used to extract acoustic values from the speech signal, such as power envelope, spectrum and fundamental frequency. The representation of emotion in this system is the dimensional approach which will be explained in the next section. Semantic primitives are viewed as a set of adjectives which is often used by listeners to describe the expressive such as fast, slow, high, low, etc. Emotion layer represents the

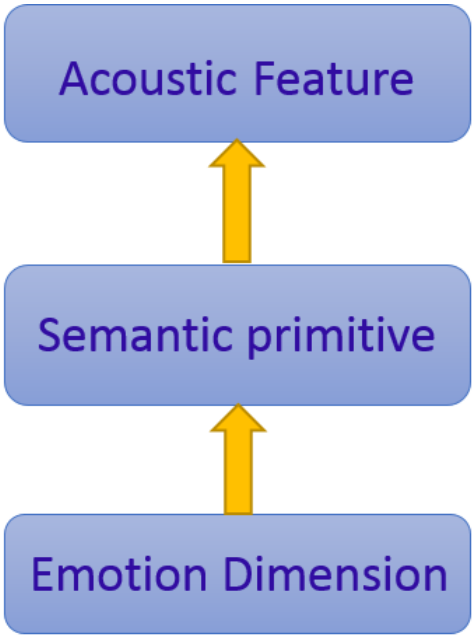

Figure 2.1: Structure of three-layered model in the emotional speech conversion system

emotion states which mainly have two methods and in this study the multi-dimensional approach is adopt.

On the basis of the three-layered model, the recognition system proposed by Elbarougy and Akagi [17] is adopted with opposite input and output to synthesize emotional speech. However, the recognition system is irreversible as the relationship between the three layers is nonlinear. The difficulty in this research is to find out whether the extracted rules are suitable for the intended emotion as the relationship between the three layers is nonlinear.

Opposite to the work of the speech emotion recognition system, the input of the emotional speech synthesis system is the position in valence-activation space, and the output is the according acoustic features as shown in Figure 2.1.

2.2 Dimensional approach

For emotional voice conversion, the method for representing emotion is of great importance. So far there are two main approaches that are related to emotion representation: categorical and dimensional approaches.

Categorical approach divided emotion into some groups which assumes that human can distinguish emotion clearly from one to another. And it just use some simplest emotion category label such as happy, sad or anger to represent emotion. Most of the previous study related to emotional speech synthesis or conversion consider about categorical approach. However, sometimes it is difficult for researchers to decide the number of category to describe the real-life emotion. And different researchers have different definitions and

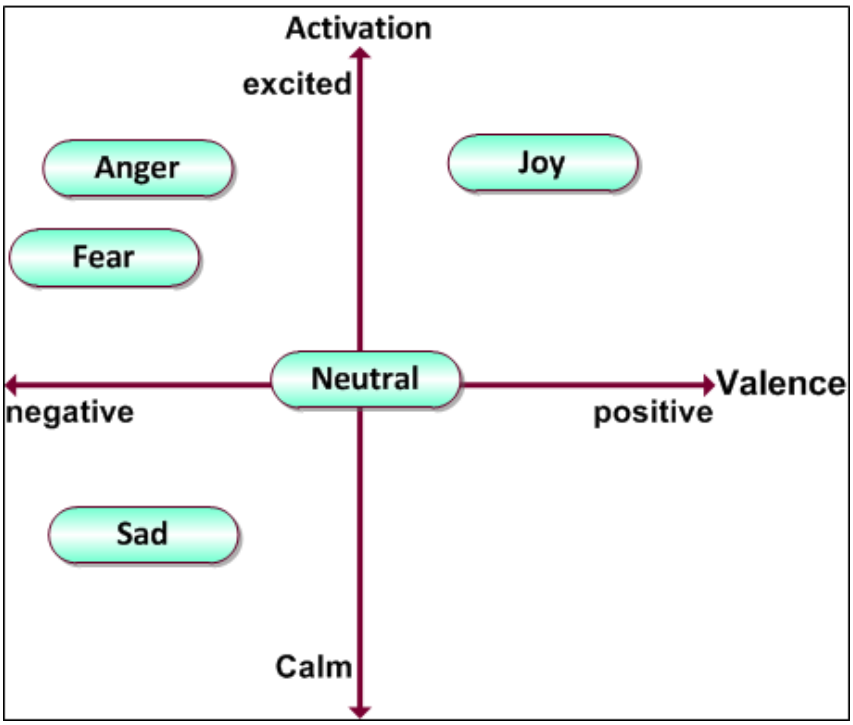

Figure 2.2: A two-dimension emotional space using valence and activation axis

degrees of one emotion. On that case, it is difficult to make a standard criterion for all the researchers. Therefore, some discrete categories and simple labels may not exactly give a good description of emotion in human daily life.

On the other hand, some researcher proposed dimension approach from the literature of psychology which represents emotion state as a position in a multi-dimensional space [18]. The used dimensions in this representation are developed in nature which shows the basic concept of emotion opposited to the specification of the emotion categories. The names of the dimension are not from the data itself but from the interpreting of data by individual researchers. Several methods are used to give a description of emotion in the multi-dimensional approach such as three-dimensional representation and two-dimensional representation.

Three-dimensional representation shows three dimensions, the first one represents valence which is from positive to negative, the second one shows the activation which is from excited to clam and the third one just gives the value of dominance which interpret power from strong to weak.

For two-dimensional representation, it just use the two dimensions, valence and activation which neglects the dominance dimension. Dominance represents power which can be used to distinguish anger and fear [19] as the database used in this study do not have the emotion fear. So that the two-dimensional representation is adopt here which is shown in Figure 2.2.

The advantages of the dimensional approach is that it can show different degrees of

emotion intensity. Emotion in human life can change due to the variance of environment from very to happy. And human sometimes produce the emotional speech in different degree of intensity to express different feeling. What's more, in human-machine interaction area, sometimes it is necessary to produce non-extreme or mild speech rather than some typical ones. And it is based on the human experimental which can give a standard of representing emotion. Therefore in this work, the two-dimensional continuous model is utilized to represent emotional state, valence and activation.

2.3 Fuzzy inference system

The three-layered model is the structure of this system and it need a tool to connect the three layers for estimating acoustic features. Fuzzy logic is considered as it turns human knowledge into mathematical system utilizing If-Then rules and it is based on the non-linear relationship of arbitrary complexity [20]. The relation between emotion dimension and the acoustic features is also non-linear. Moreover, fuzzy logic is build on the foundation of natural language, and semantic primitives in our system is just natural language. Figure 2.3 shows the structure of fuzzy inference system (FIS) which is consist of the following five different parts.

- A fuzzification interface converting crisp inputs into linguistic variables
- A defuzzification interface turning fuzzy values to crisp outputs
- A decision-making part acted as a fuzzy inference manipulating engine
- A database consists of membership functions of fuzzy sets
- A rule base defining some If-Then rules

Although there is a powerful inference system of fuzzy inference system, sometimes it is difficult to transform human knowledge into a rule base as it has no study ability. On the contrary, Neural Network (NN) can give a strong learning ability. Adaptive Neuro Fuzzy Inference System combines two advantages together. It overcomes this problem by using artificial neutral networks that can identify fuzzy rules and turn the parameters of membership functions automatically.

A non-linear mapping from the input (emotion dimension) to the output (acoustic feature) is the basic of a fuzzy-logic system. Fuzzy sets and fuzzifiers are utilized to turn the inputs from arithmetical field to fuzzy field. Then, fuzzy inference manipulating engine and fuzzy rules are used to the fuzzy filed. Using defuzzifiers, the results will be convert back to arithmetical field. Gaussian functions are conducted for fuzzy sets and linear functions are utilized for outputs on ANFIS method. The standard deviation, the coefficients of the output linear functions and mean of the membership functions are utilized as network parameters for ANFIS. At the final node of the system, the summation of outputs is computed. The final node is the rightmost node of a network.

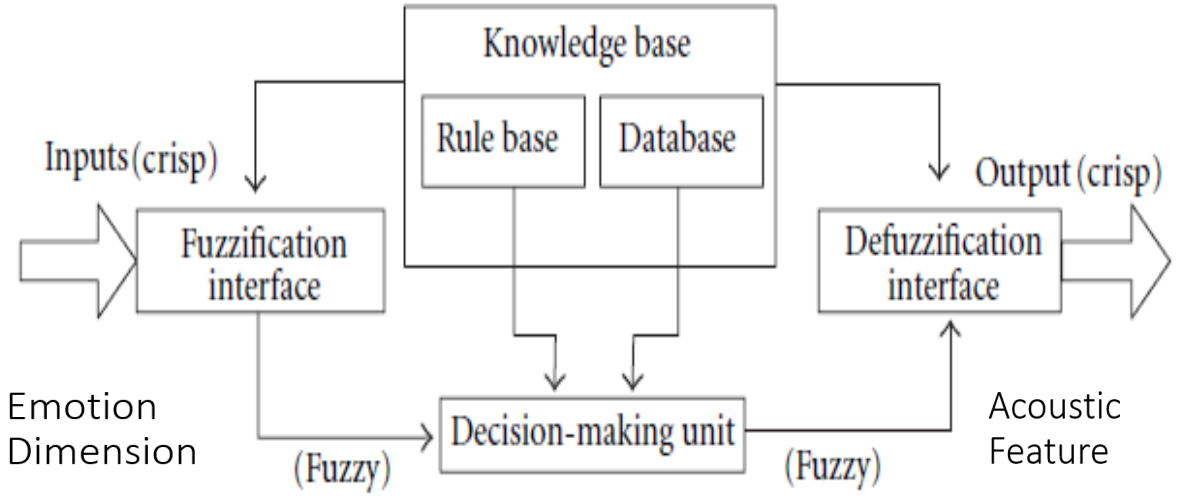

Figure 2.3: Structure of Fuzzy Inference System (FIS)

ANFIS is a system with multi-input and single-output. Therefore in our system the number of ANFIS is decided by the number of outputs in each two layers. The details of using the FIS will be introduced in Chapter 5.

Chapter 3

Outline of the system

This section outlines the emotional speech conversion system proposed by the author. For emotional speech conversion system, the ultimate goal is to transfer the neutral speech to emotional one. In order to obtain the emotional speech, the relationships between the value of the position in the multi-dimension and value of acoustic features of emotional speech is necessary to find. As the relationship between them is non-linear, the Fuzzy Inference System (FIS) is considered. And then the method for transferring the acoustic features of neutral speech of the acoustic features to emotional speech is also of great importance. Then the modification method will be used to synthesize the emotional speech using the modified acoustic features.

Therefore, the ultimate goal can be divided into two sub-goals. The first one is to get a accurately estimated acoustic features which are more related to emotion. The second one is to find a better method to use the modified acoustic features to synthesize emotional speech.

Based on the two sub-goals, the system can be divided into two parts. The first part is applied to estimate acoustic features of emotional speech, and the second is the modification of acoustic features of neutral speech to emotional ones according to the displacements of estimated acoustic features of emotional speech and neutral ones which is shown 3.1. The input of this system is the position value in dimensional space and the output is the converted emotional speech.

Figure 3.2 shows the flow chart for estimating the acoustic features. The procedure of estimation can be divided into two parts. The first part is to train the Fuzzy Inference System (FIS) to get the model which is called model creation. And the second part is to use the trained model to estimate the acoustic features of emotional speech when giving the value in the multi-dimensional space.

In the model creation part, the evaluated emotion dimensions from listening tests, evaluated semantic primitives from listening tests, and the extracted acoustic features from database are firstly needed. And the database which is utilized in this study are detailed in Chapter 4. To connect the three layers, we use the fuzzy inference system (FIS). FIS Modeling_1 is applied for connecting the emotion dimensions layer and the semantic primitives layer of which the inputs are the values of Valence and Activation (V-A) and outputs are semantic primitives. The inputs of FIS Modeling_2 are the values

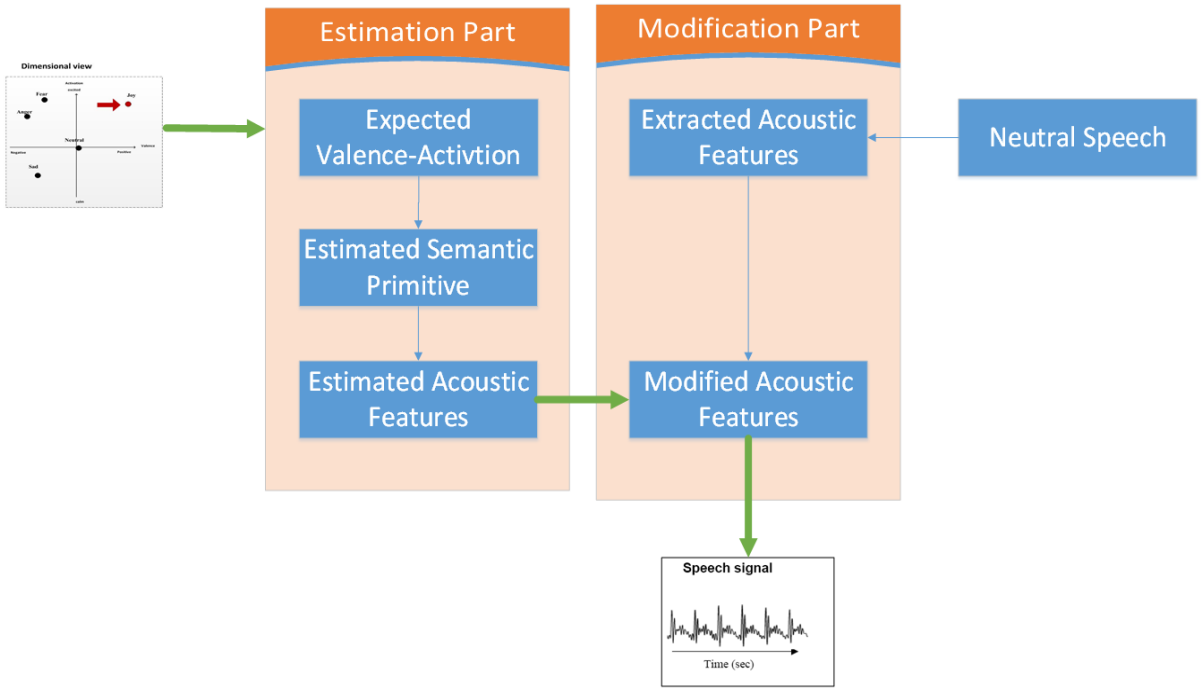

Figure 3.1: Schematic graph of the emotional voice conversion system

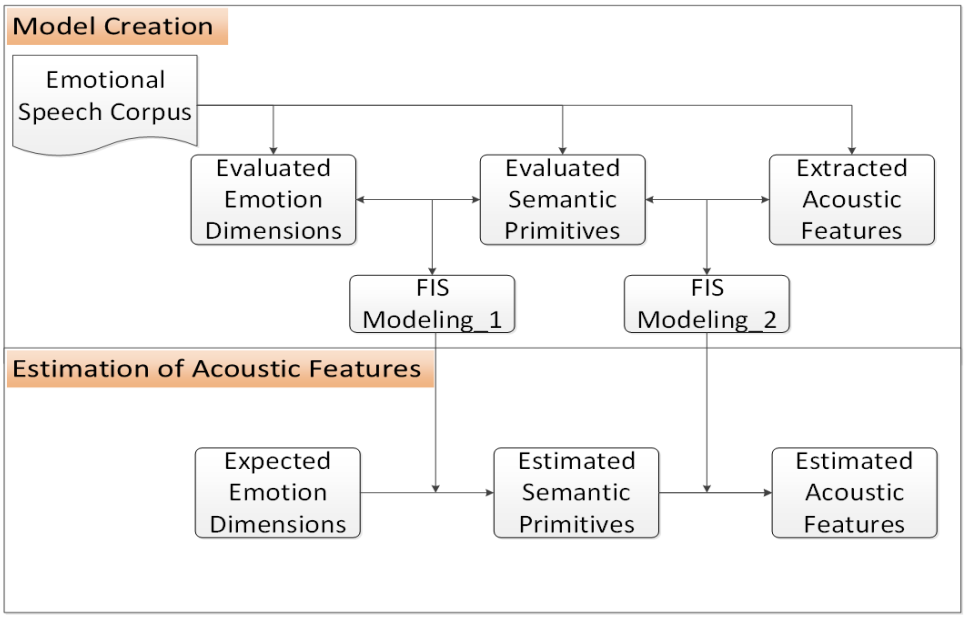

Figure 3.2: Flow chart for estimating acoustic features

of the semantic primitives, and outputs are the values of acoustic features. This is fully explained in Chapter 5.

After FISs have been built, we firstly give the input, position values in dimensional speech to FIS Modeling_1 to get the estimated semantic primitives. Then the estimated semantic primitives are the inputs of FIS Modeling_2 to obtain the estimated acoustic features. To obtain the emotional converted speech, the modified values from the estimated acoustic features of emotional speech and the extracted acoustic features from neutral speech are firstly need to be found. Then the modified acoustic features of emotional speech are used to synthesize emotional speech using some tools and models. This will be discussed in Chapter 5.

Chapter 4

Elements of the system

In this section, the description of database used in this study is explained first. Then, in accordance with our method, the acoustic features extracted from the database, semantic primitives, and emotion dimension acquired by listening tests are presented in the next subsections.

4.1 Speech Materials and Subjects

In this research, the purpose of emotional speech synthesis is to present a speech wave that can give a specific impression listeners perceive. Thus, we can use acted voices that can produce specific impressions in our mind. On that case, acted emotion are suitable for training and testing the system. A Japanese database recorded by Fujitsu Laboratory is adopt in this study to train, test and explore the affectiveness of the proposed system.

Fujitsu database which contains 179 utterances spoken by a professional female voice actress is used in this study. The 179 utterances consist of five different emotion states: neutral, happy, sad, cold anger, and hot anger. In addition, there are 20 different Japanese sentences in this database which is shown in Table 4.1. The left column shows Japanese Sentence and the right column gives the translation version in English.

Table 4.1: The content of sentences in Japanese and the according translations in English

id	Sentences in Japanese	Translation in English.
1	新しいメルが届いています。	You've got a new mail.
2	にることなんてありません。	There is nothing frustrating.
3	待ち合わせは青山らしいんです。	I heard that we would meet in Aoyama.
4	新しいをいました。	I brough a new car.
5	いらないメルがあったらててください。	Please delete any unwanted e-mails.
6	そんなの古い迷信ですよ。	That's an old superstition.
7	みんなからエルが送られたんです。	Many people sent cheers.
8	手が届いたはずです。	You should have reveived a letter.
9	ずっとています。	I will think about you.
10	私のところには届いています。	I have received it.
11	ありがとうございました。	Thank you.
12	申しございません。	I am sorry.
13	ありがとうは言いません。	I won't say thank you.
14	旅行するには二人がいいのです。	I'd like to travel just the two of us.
15	がくなりそうでした。	I felt like fainting.
16	こちらの手いもございました。	There were our mistakes.
17	花火をるのにゴザがいりますか。	Do we need a straw mat to watch fireworks.
18	もうしないと言ったじゃないですか。	You said you would not do it again.
19	通りに来ないを教えてください。	Please tell me why you don't come on time?
20	サビスエリアで合流しましょう。	Meet me at the service area.

Table 4.2: 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)

Each sentence has one for neutral and two for other emotion states. The total number of utterances is 179 because one cold anger utterance is missing from the database. Table 4.2 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 4.3.

4.2 Acoustic Feature Extraction

For emotional speech conversion, the acoustic features play an important role because they can hugely affect the effectiveness of the conversion speech. In the field of F0, power envelope, power spectrum, and duration, 16 acoustic features are put to use in accordance with the work of Huang and Akagi [12]. In addition [17] add five more acoustic features related to the voice quality. The acoustic features extracted form database are shown in Table 4.4.

Except for the acoustic features related to duration that is extracted by segmentation manually, the rest are obtained by the high quality speech analysis-synthesis system STRAIGHT [21]. Also, the same as the work of Hamada et al. [9], five acoustic features related to the voice quality were focused on as they are important for perceiving the expressive voice. All together, 21 acoustic features are classified into the following subgroups:

F0 related features: The acoustic features related to F0 were extracted from each phrase in each sentence: F0 mean value of a rising slope of the F0 contour (F0_RS),

Table 4.3: 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

highest F0 (F0_HP), average F0 (F0_AP), and rising slope of the F0 contour for the first accentual phrase (F0_RS1).

Power envelope related features: The acoustic features related to power were extracted from each phrase in each sentence: Mean value of power range in accentual phrase (PW_RAP), 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), and the average power (PW_RHT) were measured.

Power spectrum related features: First formant frequency (SP_F1), second formant frequency (SP_F2), and third formant frequency (SP_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) is used to measure voice quality and was calculated using the following equation:

$$SP\_TL = A_1 - A_3 \tag{4.1}$$

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 [22]. To describe acoustic consonant reduction [23], spectral balance (SP_SB) is adopted. It was calculated in accordance with the following equation:

$$SP\_SB = \frac{\sum f_i \cdot E_i}{\sum E_i} \tag{4.2}$$

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

Duration related features: Total length (DU_TL), consonant length (DU_CL), and ratio between consonant length and vowel length (DU_ RCV).

Voice quality: According to Menezes et al. [24], H1-H2 is concerns glottal opening, which means the mean value of difference between the first and second harmonics for vowels /a/, /e/, /i/, /o/, and /u/ per utterance. MH_A, MH_E, MH_I, MH_O, and MH_U were used as indexes of voice quality.

Table 4.4: The extracted acoustic features

	Acoustic Features
F0 (4)	**F0** mean value of **R**ising **S**lope (F0_RS), F0 Highest Pitch (F0_HP), F0 Average Pitch (F0_AP) F0 Rising Slope of the 1st accentual phrase (F0_RS1)
Power envelope (4)	mean value of Power Range in Accentual Phrase (PW_RAP), Power Range (PW_R), Rising Slope of the 1st accentual phrase (PW_RS1), the Ratio between the average power in High frequency portion (over 3 kHz) and the Total average power (PW_RHT)
Spectrum (5)	1st Formant frequency (SP_F1), 2nd Formant frequency (SP_F2), 3rd Formant frequency (SP_F3), Spectral Tilt (SP_Ti), Spectral Balance (SP_SB)
Duration (3)	Total Length (DU_TL), consonant length (DU_CL), Ratio between Consonant length and Vowel length (DU_RCV).
Voice Quality (5)	the mean value of the difference between the first harmonic and the second harmonic H1-H2 for vowel /a/,/e/,/i/,/o/, and /u/ per utterance, MH_A, MH_E, MH_I, MH_O, and MH_U, respectively.

4.3 Semantic Primitives Evaluation

In the listening test, 11 graduate students, all native Japanese speakers without any hearing impairment, were asked to evaluate the utterances as subjects. This part is following the work of [17].

As mentioned above, in the three-layered model, the bottom layer is the emotion dimension layer, the middle layer is the semantic primitives layer, and the top layer is the acoustic features layer. Therefore, the value of semantic primitives is essential for building this model as the middle layer construct a bridge to connect the bottom and top layer.

Subjects were asked to give subjective values for 17 adjectives: Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow. These words were selected by Huang and Akagi [12] as they can describe emotional speech in a balanced way. The 17 semantic primitives were evaluated 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"). Separately for each semantic primitive, the inter-rater agreement was measured by pairwise Pearosn's correlation between two subjects' ratings which shows that all subjects agreed from a moderate to a high level.

4.4 Emotion Dimensions Evaluation

The evaluation of emotion dimension is divided into two parts: valence and activation. The 11 subjects were required to rate the 179 utterances on a five-point scale {-2, -1, 0, 1, 2 }. Valence was from -2 (very negative) to +2 (very positive), and activation was from -2 (very calm) to +2 (very excited). The value of the evaluation has a high inter-rater agreement, which shows that all subjects had similar impressions of the emotional speech. This work expands on the work of Elbarougy and Akagi [17].

Chapter 5

Features estimation and modification

This section gives the two main procedure of the system the estimation of acoustic features and the modification of acoustic features. Then as not all acoustic features have a strong relationship with all semantic primitives, the related acoustic features are choosen. After the procedure of estimation and modification, the converted emotional speech can be obtain when giving the input position in Valence-Activation (V-A) space.

5.1 Estimation of acoustic features

For the three-layered model, a connection method between the three layers is needed to obtain the non-linear relationship as shown in Figure 5.1. In Figure 5.1, the input is the value position in V-A space, and then the estimated semantic primitives can be captured. Then the estimated acoustic features are the input of the second connectors and the estimated acoustic features can be obtained lastly. In this system, ANFIS is adopt as the connector of which reason has been shown in Chapter 2.

ANFIS is a system with multi-input and single-output. As is shown in Figure 5.2, two different kinds of ANFIS are for FIS Modeling_1, 17 ANFISs were trained because there are 17 semantic primitives which is the output of FIS Modeling_ 1 the middle layer. FIS Modeling_2 has 21 ANFISs for the same reason that 21 acoustic features are modified to synthesize the emotional speech. The input of the FIS Modeling_ 2 is the semantic primitives evaluated in the listening test when training the model. The values of acoustic features were found to change greatly for emotional speech and neutral speech [25]. Different people have different vocal tracts, which will influence some acoustic features such as formant frequency. For avoiding speaker-dependency and emotion-dependency, all acoustic features were normalized by the neutral speech using (3)

$$\hat{f}_{(i,m)} = \frac{f_{(i,m)}}{\sum_{i=1}^{l} f_{(i,m)}/l} \tag{5.1}$$

where m is the number of acoustic features ($m = 1, \ldots, 21$) and i is the number of utterances in the database. $f_{(i,m)}(i = 1, 2, \ldots, l, \ldots, 179)$ is a sequence value of the mth acoustic feature which comes from the extracted values in the database explained in

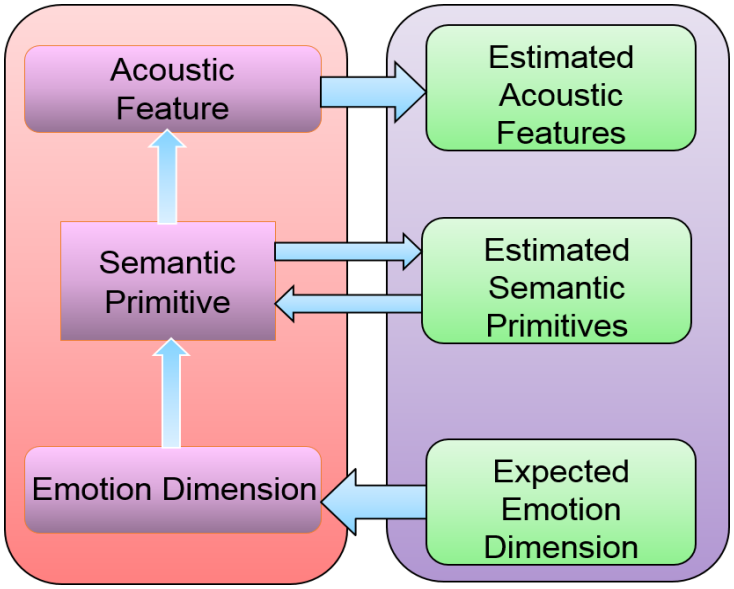

Figure 5.1: The direction of three-layered model in voice conversion system

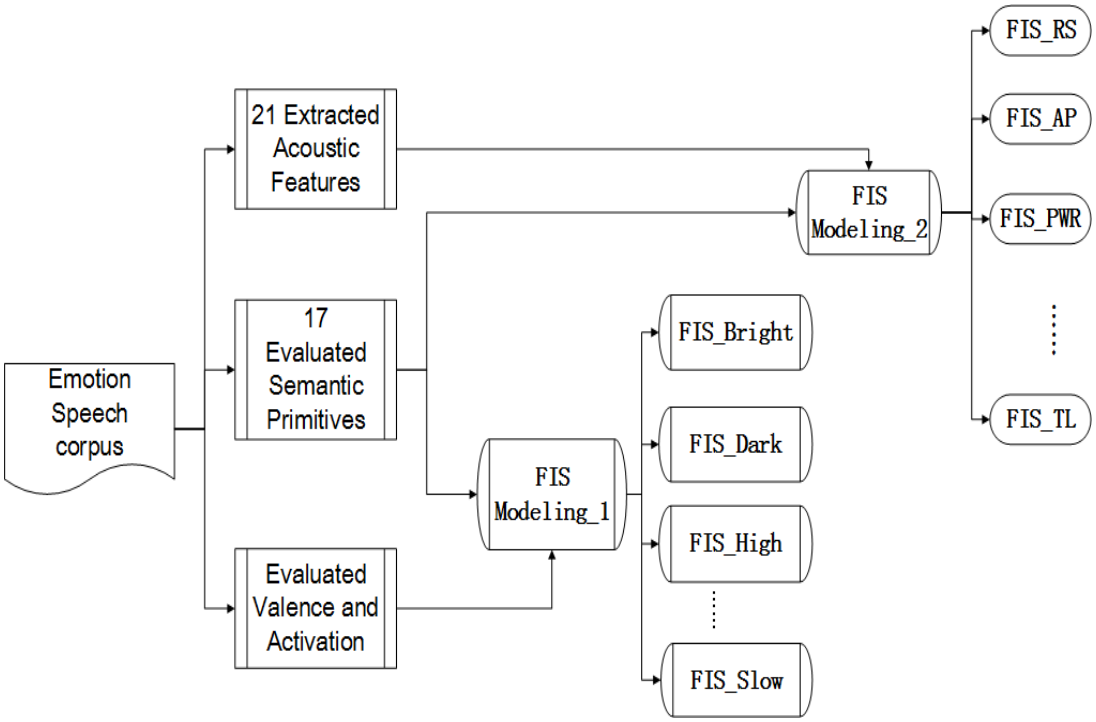

Figure 5.2: The flow chart of building Fuzzy Inference System (FIS)

Section IV(B). The first l represents the value of the neutral acoustic features, and the rest are set to other emotional states. By using (3), $\hat{f}_{(i,m)}$ can be calculated which represents the normalized value from the ith utterances of the mth acoustic feature. The requirement for using ANFIS is that all input and output should be from 0 to 1. Therefore, using the range and minimum value of every variable, all acoustic features, all semantic primitives, and all emotion dimensions were normalized in the range [0,1] when training the model. Using (4), the acoustic features between 0 and 1 can be got

$$\tilde{f}_{(i,m)} = \frac{\hat{f}_{(i,m)} - fmin_m}{fran_m} \tag{5.2}$$

where m is the number of acoustic features ($m = 1, \ldots, 21$) and i is the number of utterances in the database ($i = 1, \ldots, 179$). $\hat{f}_{(i,m)}$ is the normalized value from (3). $fmin_m$ and $fran_m$ is the minimum value and range of the mth acoustic features which have appeared in (4). Using (4), $\tilde{f}_{(i,m)}$ which means the normalized value in the range [0,1] can be used as the input for training ANFIS. For semantic primitives and emotion dimension, the normalized part into [0,1] is the same as acoustic features which is firstly needed to subtract the minimum value and then divide the range value of the semantic primitives or emotion dimension. All the data sets were divided into the training data (90%) and testing data (10%) in order to avoid the over-fitting of the model being developed. ANFIS is first trained using the training data and then validated using the testing data.

After the process of training the FIS model, equation (5) is used to obtain the estimated semantic primitives when given the input value of valence and activation. In (5), v, a represents the value of valence and activation separately, $F1_n$ means the ANFIS of the nth semantic primitives which is the same as FIS Modeling_1 in Figure 4. And sp_n represents the estimated value of the nth semantic primitives where n is the number of semantic primitive ($n = 1, \ldots, 17$).

$$sp_n(v, a) = F1_n(v, a) \tag{5.3}$$

The 17 estimated semantic primitives are the input of the FIS Modeling_2. When training the model, the extracted acoustic features were normalized in the range [0,1] so that the denormalized procedure is needed using (6) to get the actual estimated acoustic features.

$$\check{f}_m(sp_n) = F2_m(sp_n) \times fran_m + fmin_m \tag{5.4}$$

Equation (6) is adopted to obtain the acoustic features for synthesis where m is the number of acoustic features ($m = 1, \ldots, 21$), $fran_m$ is the range of the mth acoustic feature, and $fmin_m$ is the minimum value of the mth acoustic feature which give the same meaning as (4). $F2_m$ represents the ANFIS of the mth acoustic feature which is the same as FISModeling_2 in Figure 4. And $\check{f}_m$ is the estimated acoustic features.

5.2 Related Acoustic Features of Semantic Primitives

Since not all 21 acoustic features have a strong relation with the 17 semantic primitives, the acoustic features with less relation with all semantic primitives are not put into use to simplify modifications of acoustic features without spurious and wrong estimations of them. The related acoustic features of every semantic primitive were selected for synthesizing the emotional speech. The selection procedure is based on the following hypothesis: acoustic features highly related to semantic primitives hugely affect how emotional speech is synthesized. The selection procedure was done in the following four steps:

Step (1): Choosing one utterance with the maximum extent of one semantic primitive, such as Bright, from our database.

Step (2): Extracting the values of 17 semantic primitives ($\check{sp}_n(n = 1, \ldots, 17)$) of the utterance with the highest value of Bright from database and putting them in the FIS Modeling_2 using (6) so that the values of 21 acoustic features ($\check{f}_m(m = 1, \ldots, 21)$) can be obtained. On the other hand, the values of 17 semantic primitives ($\overline{sp}_n(n = 1, \ldots, 17)$) of the utterance with the neutral voice as the input of FIS Modeling_2 and the according acoustic features ($\overline{f}_m(m = 1, \ldots, 21)$) can be extracted.

Step (3): The following function is used to calculate percentage variation between the brightest and the neutral speech of one acoustic feature

$$per_m = \frac{\check{f}_m(\check{sp}_n)}{\overline{f}_m(\overline{sp}_n)} \tag{5.5}$$

Step (4): Selecting the highly correlated acoustic features for synthesizing. Considering the number of acoustic features related to semantic primitives, the percentage (per_m) above 1.4 and below 0.7 is chosen.

After the related acoustic features of every semantic primitive have been obtained, the number of semantic primitives to which every acoustic feature is related is calculated as shown in Figure 5.3. This figure also shows that the number of acoustic features related to the semantic primitives have some tendencies which is limited to 16 at most.

5.3 Modification of acoustic features

After the ANFIS was used to obtain the estimated acoustic features, voice morphing was done using the estimated acoustic features as shown in Figure 5.4. All 21 acoustic features obtained from ANFIS were modified in accordance with the following equation:

$$fmod_m(v, a) = f_{(1,m)} \times \frac{\check{f}_m(sp_n(v, a))}{\check{f}_m(sp_n(0, 0))} \tag{5.6}$$

where v means the value of the valence, and a means the value of the activation. m is the number of acoustic features ($m = 1, \ldots, 21$). $f_{(1,m)}$ is the extracted value without any normalization of the mth acoustic feature from the 1st utterance in the database which is

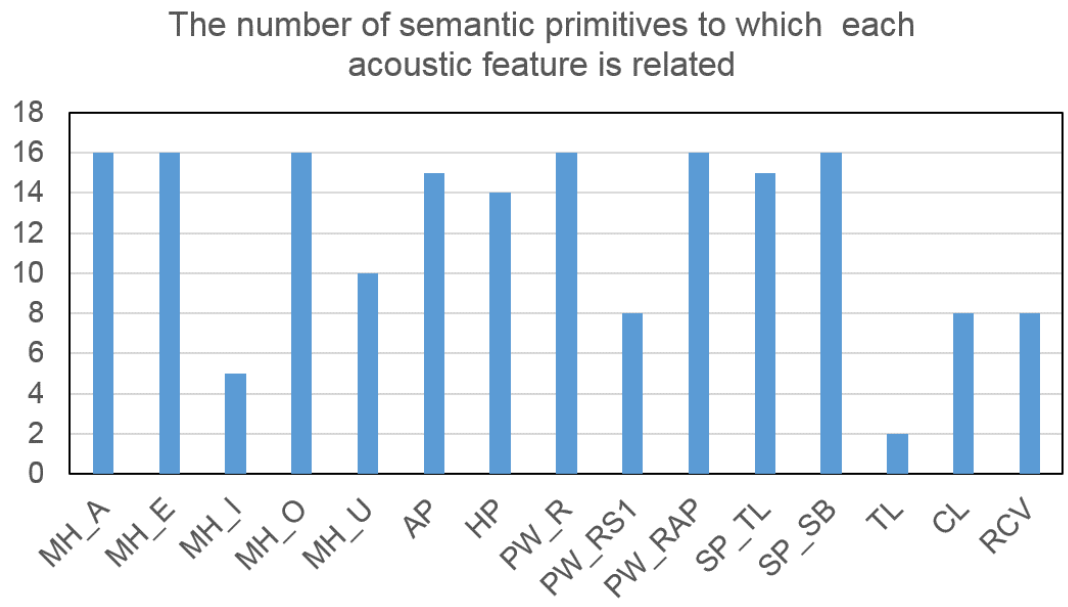

Figure 5.3: Number of semantic primitives to which every acoustic feature is related

the neutral voice. $\tilde{f}_m(sp_n(v,a))$ is the estimated acoustic feature value using (5) and (6) when input the values of valence and activation. $\tilde{f}(sp_n(0,0))$ is the estimated acoustic feature value using (5) and (6) when the values of valence and activation are both 0. $fmod_m$ is the modified value of the mth acoustic feature. The original voice was morphed using the modified values of acoustic features.

Duration and spectrum parts were the same as those in the work of Huang and Akagi [9], and the spectrum of glottal waveform part is the same as that in the concept of Hamada et al. [9]. In segmentation part, durations of phoneme, phrase, and accent parts were measured manually. By using STRAIGHT [21], power envelope and spectral sequence were extracted. The spectrum of glottal waveform is extracted using the ARX-LF model [26]. This procedure was exactly the same for the two-layered model [9].

The modification method of F0 related acoustic features (such as average pitch, highest pitch, f0 mean value of rising slope, and rising slope of the first accentual phrase) is changed in this paper. Because it is not suitable to extracted the F0 related acoustic features separately as previous work done. The Fujisaki model [16] is adopted to extract the trajectory of F0 contour from which we can obtain the F0 related acoustic features all at once. The Fujisaki model is a mathematical model represented by the sum of phrase components, accentual components, and the base line (Fb). The F0 contour can be expressed by the following equation:

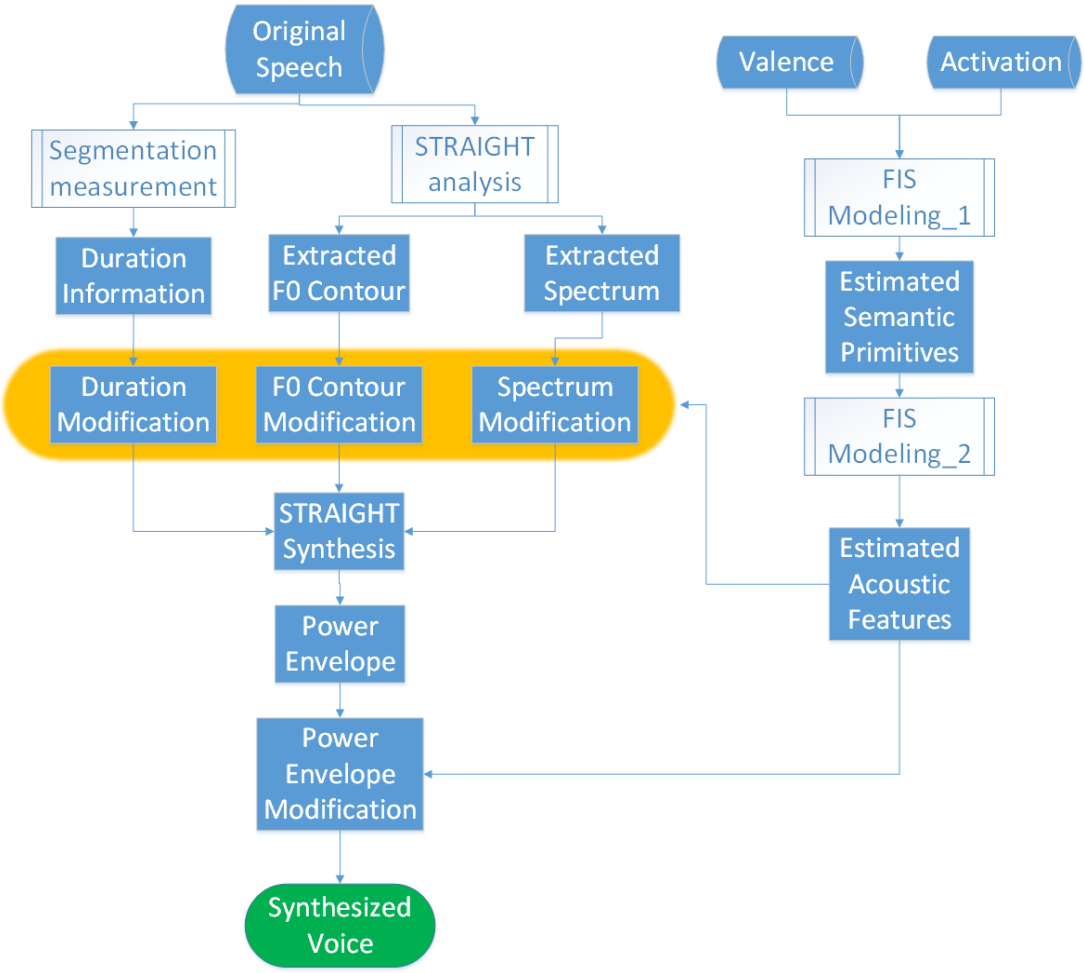

Figure 5.4: Process of modifying voice

$$\ln F_0(t) = \ln Fb + \sum_{i=1}^{I} Ap_i Gp_i(t - T_{0i}) + \sum_{j=1}^{J} Aa_j\{Ga_j(t - T_{1j}) - Ga_j(t - T_{2j})\} \quad (5.7)$$

$$Gp_i(t) = \begin{cases} \alpha_i^2 t \exp(-\alpha_i t), & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (5.8)$$

$$Ga_j(t) \begin{cases} \min[1 - (1 + \beta_j t)\exp(-\beta_j t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases} \quad (5.9)$$

where $G_{p(t)}$ represents the impulse response function of the phrase control mechanism, 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, and the work of Mixdorff [27] is adopted in this paper where $\alpha = 1.0/s$ and $\beta = 20/s$. The parameters are T0 and T1 for the duration and Ap, Aa, and Fb for the magnitude of F0. First, duration was converted in accordance with (T0 and T1), and then the estimated values related to F0 were converted and modified with the parameter values of the Fujisaki model (Ap, Aa, and Fb). By

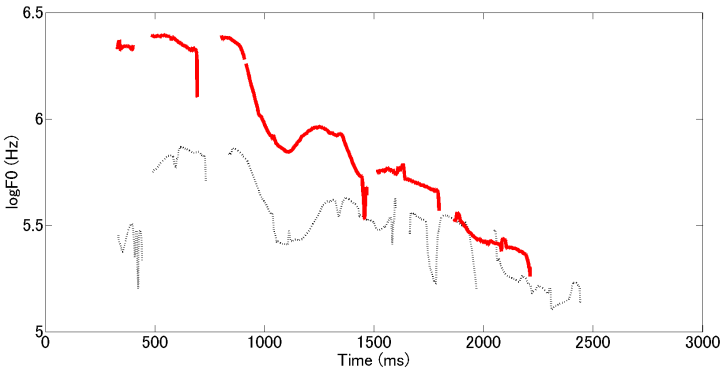

Figure 5.5: F0 trajectory of a neutral speech (dashed) and synthesized speech (solid) [15].

controlling fundamental frequency, neutral speech was converted into emotional speech related to a position on the V-A space which is shown in Figure 5.5.

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.

Chapter 6

Evaluation

After presenting the estimation and modification procedure in this system, the converted speech can be obtained. In this part, the evaluations of the estimation and modification will be shown separately.

6.1 Evaluation of estimation

6.1.1 Correlation Coefficient

The three layers are connected using ANFIS so that by giving the value of activation and valence, the estimated semantic primitives and acoustic features can be obtained. However, the accuracy of the estimated acoustic features and semantic primitives has not been explored yet.

In Figure 6.2, the correlation coefficient of 3-layer subtract layer can be got. For the correlation coefficient of semantic primitives, the results are shown in Figure 6.3 which suggest that most of FISs from the semantic primitives can work well as 15 of the 17 semantic primtives can get the correlation coefficient above 0.9 and all are above 0.8.

Correlation coefficient $R2^{(j)}$ (j = 1, 2, ... n) between the estimated semantic primitive a and the evaluated semantic primitive b can be determined by the following equation:

$$R1^{(j)} = \frac{\sum_{i=1}^{n}(a_i - \overline{a})(b_i - \overline{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \overline{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \overline{b})^2}} \tag{6.1}$$

where $\overline{a}$ and $\overline{b}$ are the average values for $x = \{a_i^{(j)}\}$, $y = \{b_i^{(j)}\}$, respectively.

Correlation coefficient $R2^{(j)}$ between the estimated acoustic feature y and the extracted acoustic feature $x$can be determined by the following equation:

$$R2^{(j)} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{6.2}$$

where $\overline{x}$ and $\overline{y}$ are the average values for $x = \{x_i^{(j)}\}$, $y = \{y_i^{(j)}\}$, respectively.

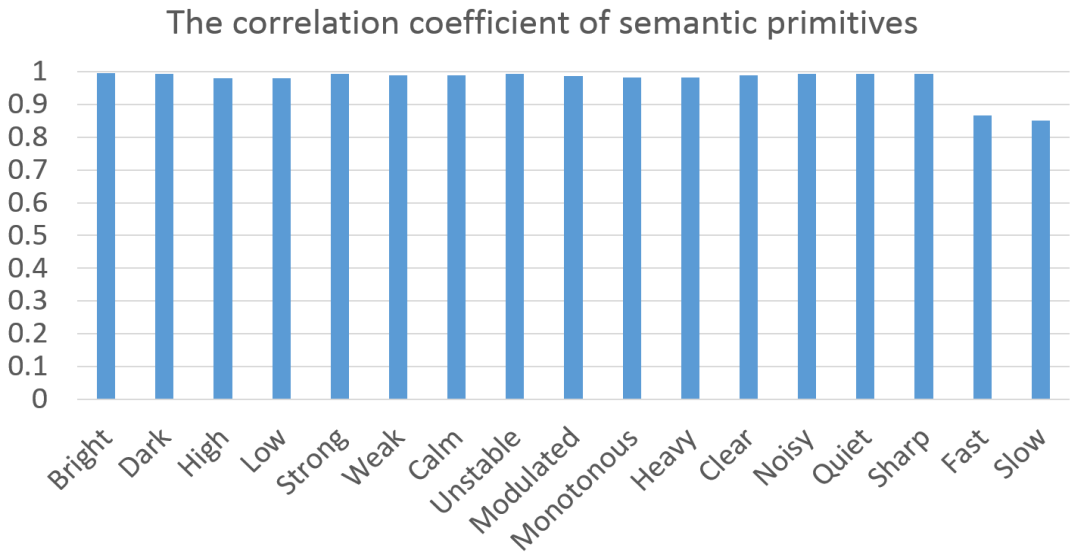

Figure 6.1: Correlation coefficient of semantic primitives from 3-layered model

The two-layered model of Hamada et al. [9] uses the same method for evaluating the system performance. By using the two synthesis systems separately, the values of valence and activation of 179 utterances are given as the inputs, and acoustic features can be obtained. Figure 6.1 displays the correlation coefficient results of the two and three-layered models (blue and red columns, respectively). In Figure 6,2 the subtraction from the correlation coefficients of three-layered to the correlation coefficients of two-layered model can be got.

From these figures, we can see that among the performances of estimating the 21 acoustic features, correlation coefficients were higher for 12 acoustic features when using the three-layered than two-layered model, the same for seven acoustic features when using both models, and two acoustic features lower than two-layered model. The 12 acoustic features include significant ones for speech emotion recognition [29].

6.1.2 Mean Absolute Error

On the other hand, Mean Absolute Error (MEA) which shows the distance between the estimated acoustic features and extracted acoustic features is adopt. The MAE is measured for each acoustic features according to the following equation:

$$E^{(j)} = \frac{\sum_{i=1}^{m} |x_i^j - y_i^j|}{N} \tag{6.3}$$

where $x_i^{(j)}(i = 1, 2, ..21)$ is the estimated values of jth acoustic features using the system and $y_i^j(i = 1, 2, ...21)$ is the value of jth extracted acoustic features.

Figure 6.3 and figure 6.4 show the results of MEA from the two- and three- layered

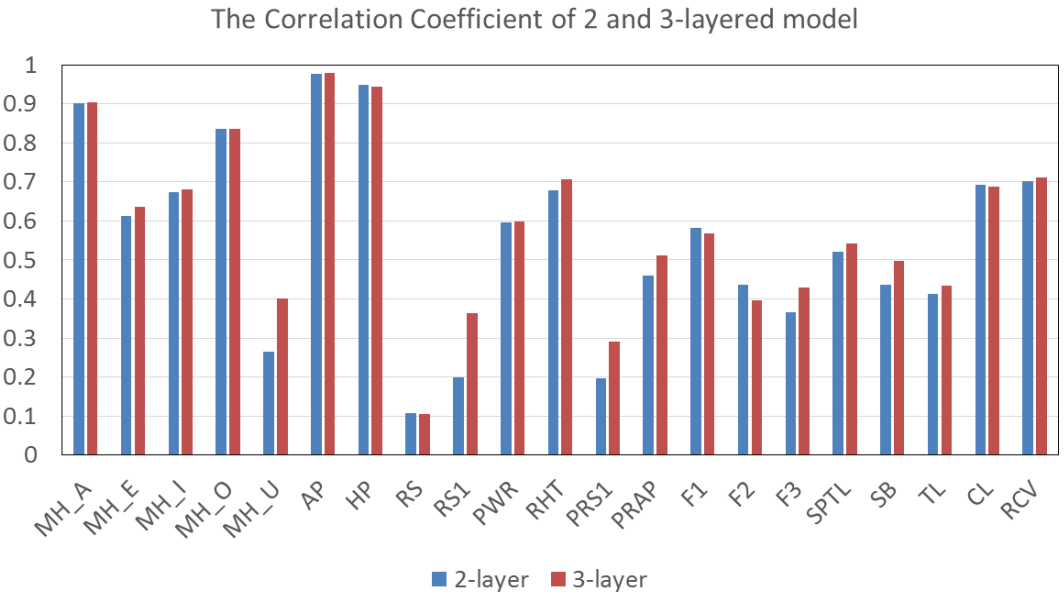

Figure 6.2: Correlation coefficient of acoustic features from two- and three-layered models

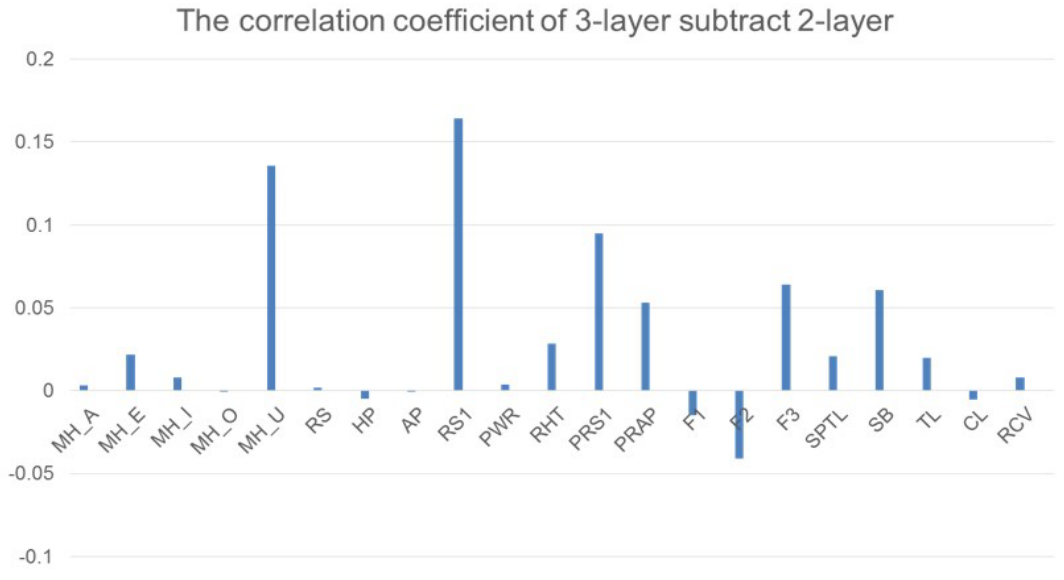

Figure 6.3: Correlation coefficient of acoustic features from two-layer subtracts three-layered models

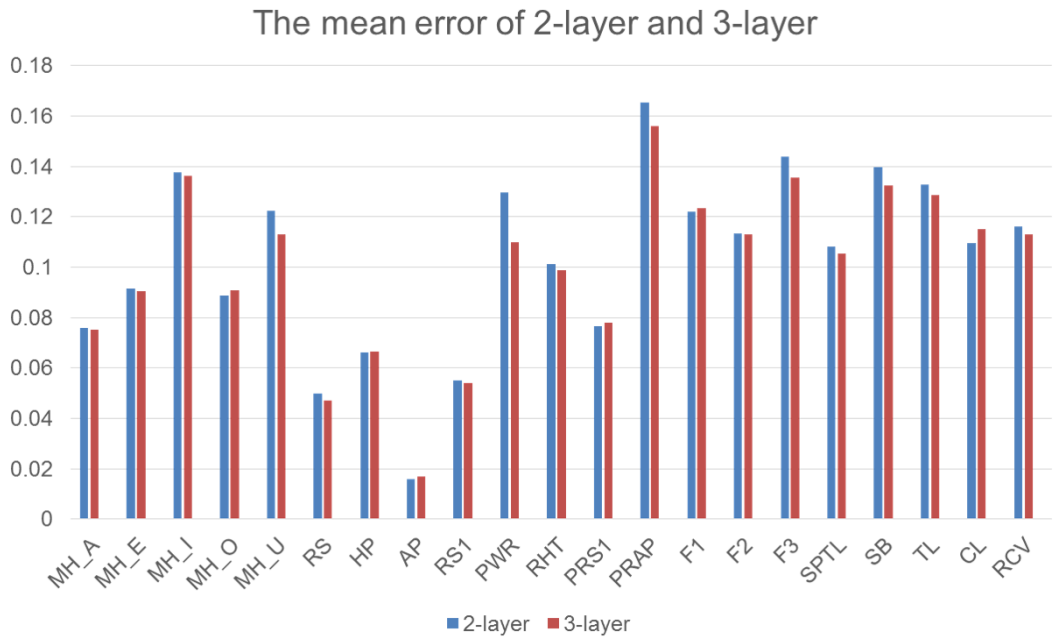

Figure 6.4: Mean Absolute Error of acoustic features from two- and three-layered models

mode and the results of MEA from two-layered model subtract three-layered model. Using the 3-layer, MEAs of 15 acoustic features have been reduced compared with 2-layer.

Using the 3-layer, the MEAs of 15 acoustic features have been reduced compared with 2-layer. The 15 acoustic features include significant ones for speech emotion recognition [29]. 4 acoustic features are almost the same as the 2-layer and 2 acoustic features are higher than 2-layer.

As most correlation coefficients of three-layered model are increased and most MEAs of three-layered model are decreased, a conclusion can be made that the three-layered model more accurately estimated the acoustic features than the two-layered model.

6.2 Evaluation of modification

6.2.1 Listening Test

Listening tests are required to verify whether converted emotional speech can be perceived as the intended position. What's more, the naturalness of the converted emotional speech is evaluated by doing listening test which is shown in the following parts.

- **Subjects:** Seven Japanese students (six males and one female; mean age: 25 years old) who have normal hearing level without any hearing loss were invited to do the listening tests. The number of subjects in the listening tests is the same as the two-layered model.

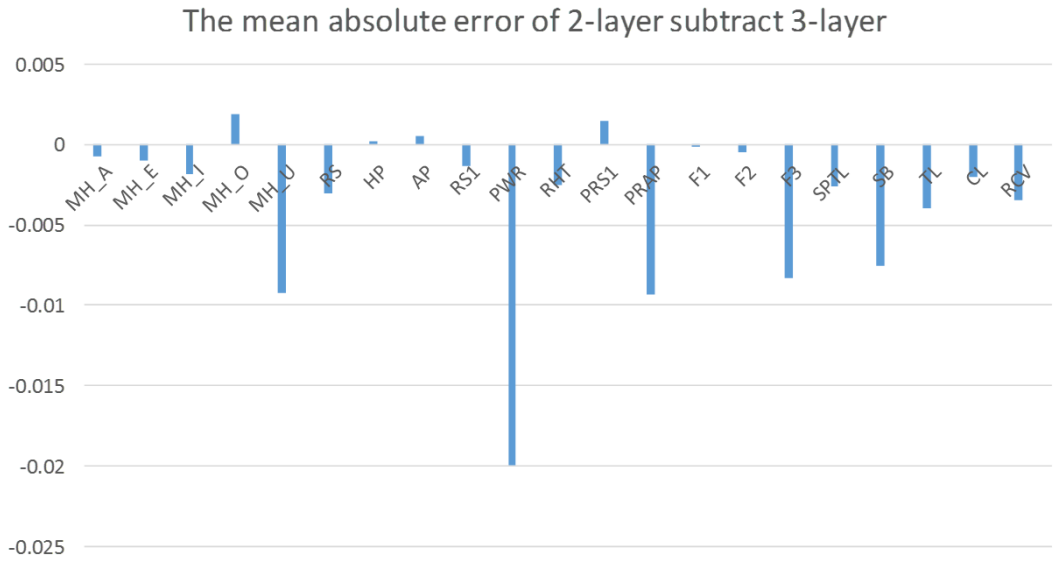

Figure 6.5: Mean Absolute Error of acoustic features from two-layered model subtracts three-layered model

- **Stimuli:** In the listening test for the two-layered model [9], 76 synthesized voices were used with the same position in V-A space as shown is in Figure 6.6. Happy voices are situated in the 1st quadrant, angry voice are in the 2st quadrant, and sad voices are located in the 3rd quadrant. Every quadrant had 25 pieces of synthesis speech, and there was one neutral voice in the center position, which is the original spoken by a professional voice actress in the Fujitsu database. The content of all utterances was
  - /Atarashi meru ga todoite imasu./ (Japanese original).
  - /You've got a new e-mail./ (English translation).

- **Procedure:** In a soundproof room, subjects were invited to listen to the stimuli, which were presented through an audio interface (FIREFACE UCX, Syntax Japan) and headphones (HDA200, SENNHEISER). The mean sound pressure level of the original voice was 65 dB, and the sound pressure level of all stimuli ranged from 63 dB to 67 dB.

For valence and activation, subjects listened to all stimuli twice. The reason for this is that they were supposed to acquire an impression of the whole stimulus the first time and then evaluate one dimension from -2 to 2 in 40 scales. What is more, valence and activation needed to be done separately. The interval was at least one day so that they would not mistake the conception of valence and activation. Valence and activation were evaluated

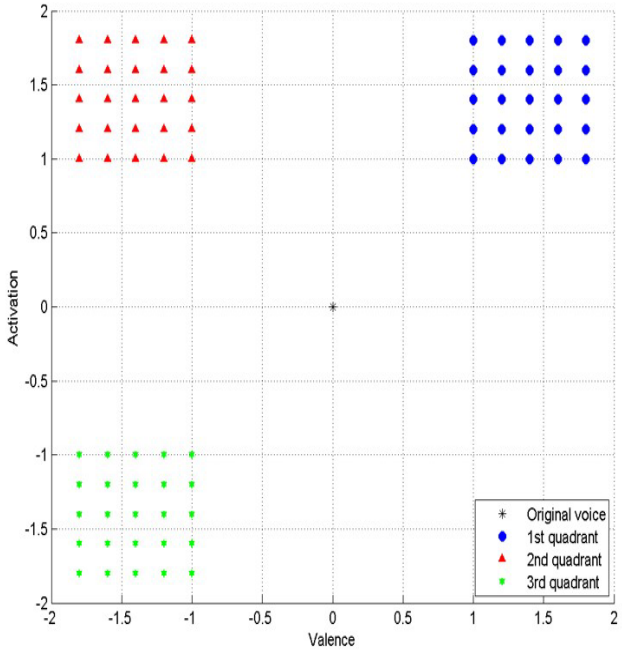

Figure 6.6: Stimuli position in valence-activation space

using forty scales (Valence: Left [Very Negative], Right [Very Positive]; Activation: Left [Very Calm], Right [Very Excited]: range $-2 \sim 2$ by 0.1 step).

For naturalness, subjects evaluate the naturalness of the converted speeches compared with the genuine emotional speech produced by human using the experience in daily life in five scales (Left [Bad], Right [Excellent]: range 1 - 5 by 1 step).Subjects evaluated these scales using the graphic user interface (Figure 6.7 and Figure 6.8). In each evaluation task, subjects could listen to the stimulus repeatedly.

6.2.2 Results of listening test

The evaluated positions in V-A space are shown in Figure 6.9. As the position of stimuli, the number of subjects, and the evaluation in the listening test were the same as for the two- and three-layered models [9], we can compare the results of positions in V-A space between both models [9]. The results of the two-layered model for evaluated position are shown in Figure 6.10.

To investigate the distance between the intended position of stimuli and listeners' evaluation, we calculated the mean absolute error (MAE) (How much error is there between the stimuli's positions and the evaluated positions?). The values of distance between the intended values and evaluated values of valence and activation were calculated separately

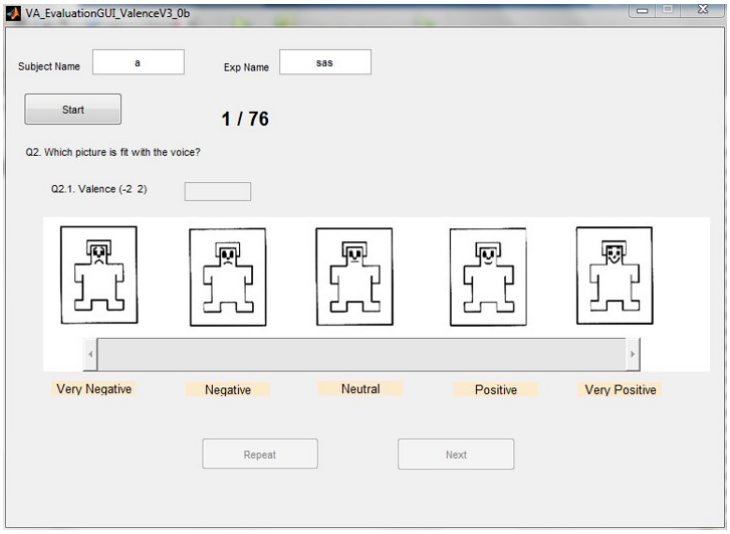

Figure 6.7: Graphic user interface for evaluate Valence and Activation

Figure 6.8: Graphic user interface for evaluate naturalness

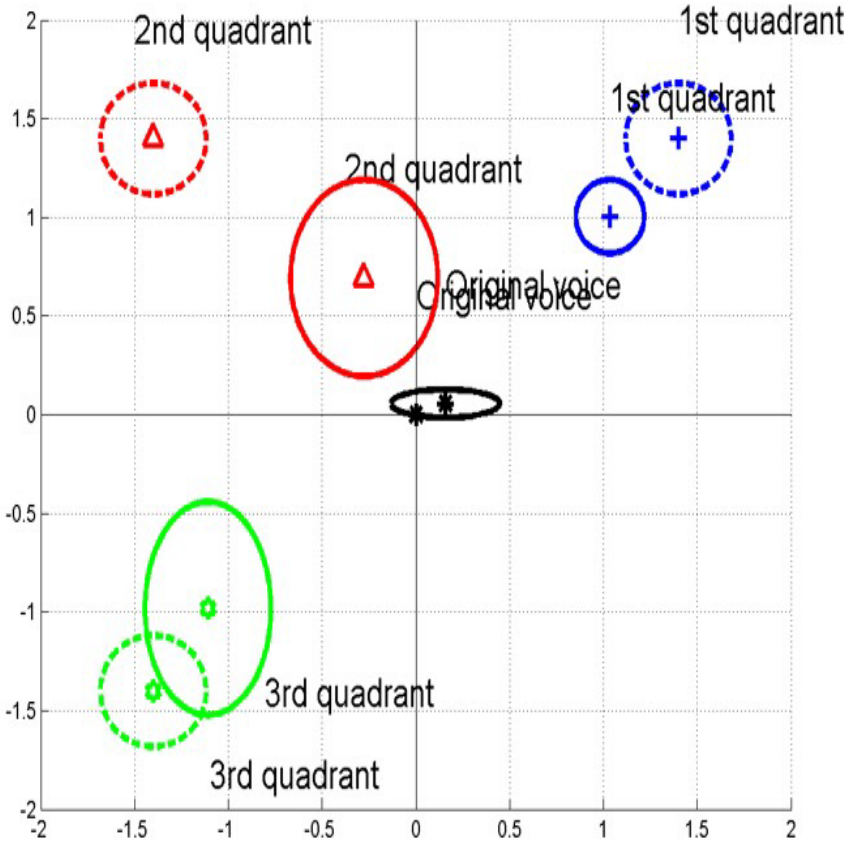

Figure 6.9: Evaluated positions in valence-activation space using three-layered model. Blue, red, and green points are the average values of the 1st, 2nd, and 3rd quadrants, respectively. Each circle describes the standard deviation (solid: evaluated value for synthesize voice; dashed: stimulus value for intended emotional voice)

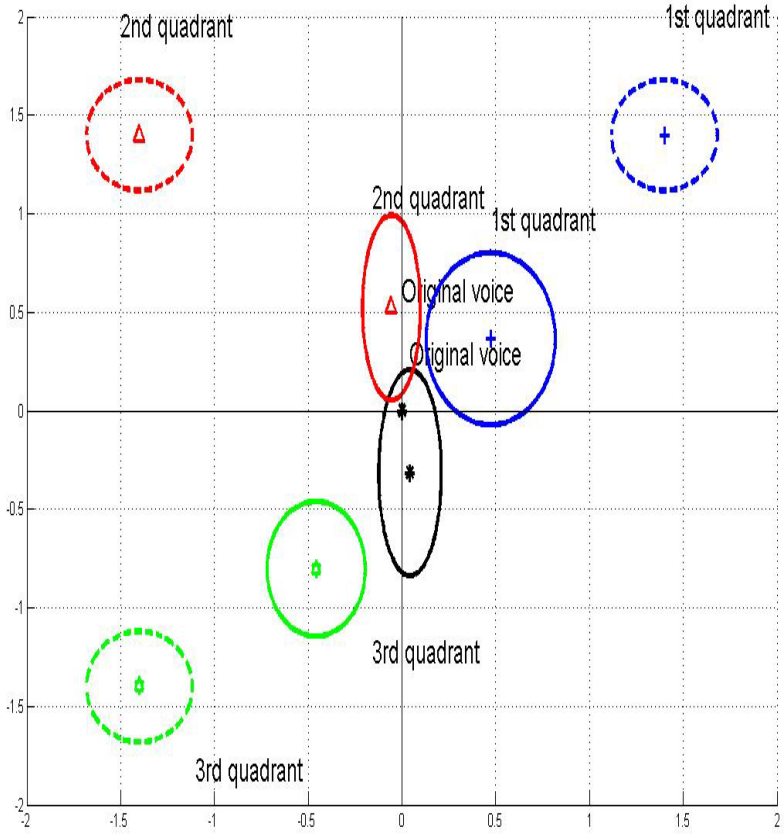

Figure 6.10: Evaluated positions in valence-activation space using two-layered model [9]

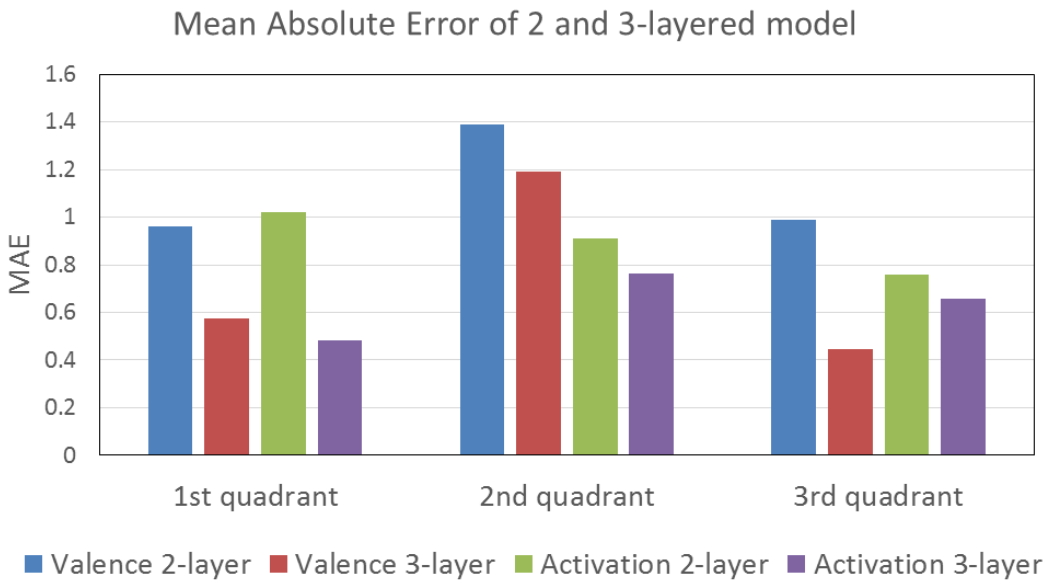

Figure 6.11: MAEs of each quadrant using two- and three-layered models

in each quadrant. The MAE is calculated in accordance with the following equation:

$$MAE^{(j)} = \frac{\sum_{i=1}^{N} |\hat{x}_i^{(j)} - x_i^{(j)}|}{N} \tag{6.4}$$

where $j \in \{V, A\}$, $\hat{x}_i^{(j)}$ is the evaluated value for the synthesis voice and $x_i^{(j)}$ is the value of the intended stimulus. The MAEs for each quadrant by two- and three-layered models [9] are shown in Figure 6.11.

From Figure 6.9 and Figure 6.10, we can find that the evaluated positions from three-layered model in the three quadrants are closer to the intended positions than those from the two-layered model. The position evaluated by two-layered model is close to the center point, which means that the synthesized speech may not express the strong intensity of emotion. In contrast, the position evaluated by the three-layered model is much closer to the intended position. The results of MAEs in Figure 6.11 reveal that the mean absolute value between stimuli position and evaluated position for the three-layered model is about 0.6, which improves on that for the two-layered model, 1.0. However, from Figure 6.9, we can see that anger is not as well-perceived as the other two emotions, resulting in the MAEs of valence and activation in the second quadrant being higher than in the others. In the future, the improvement of anger voice needs to be investigated.

On the other hand, the naturalness results are shown in Figure 6.12. From this figure, we can find that the score of naturalness in the 1st quadrant is higher than the other two quadrant which means that converted speech of happy is better than angry and sad but all are in the acceptable level.

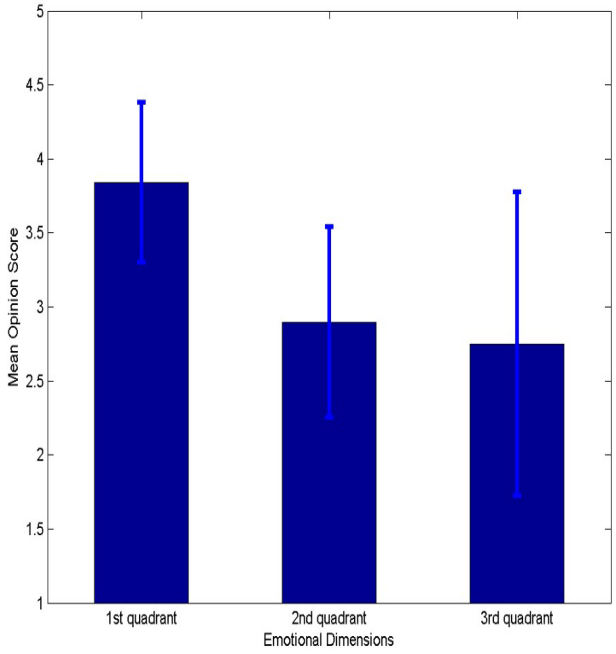

Figure 6.12: Mean Opinion Score in the three quadrants

All in all, the new estimation method outperforms the two-layered model. The synthesized speech using the revised model with the added modification method can give the intended impression. What is more, the distance between the synthesized and intended speech is smaller for the three-layered model than for the two-layered model, which is a great improvement.

Chapter 7

Conclusion

7.1 Summary

In this study, we improved the conventional emotional speech conversion system based on dimensional method in order to precisely predict the acoustic features as well as to synthesize affective speech much similar to what intended in dimensional space. The higher correlation coefficient comparing to the two-layered model [9] shows that three-layered model estimates acoustic features more accurately than the previous two-layered model. Results of subjective evaluations revealed that emotional speeches converted by three-layered model using new modification method, Fujisaki method can give the intended impression to a much similar degree as than the previous two-layered model in the emotion dimension. And the naturalness of converted speeches achieve an average score, 3.2 whose highest is 5 by subject evaluations. Above all, a conclusion can be made that an emotional conversion system utilizing three-layered model in dimensional approach can achieve better quality converted emotional speech than previous method.

7.2 Future work

As we have discussed in Chapter 6, the positions of the converted anger speeches are a little far from intended comparing with the other two kinds of speeches. Power envelope is much related to the angry speech as the semantic primitives value of strong in angry speech are much higher than other ones from the listening test. The new modification method of power envelope such as 2nd order critically damped model [28] will be researched. What's more, for Speech to Speech Translation System, para-linguistic information, such as intonation will be considered in synthesizing speech area using three-layered model.

7.3 Contribution

The rule-based emotional voice conversion utilizing three-layered model for dimensional approach is proposed in this study. Comparing with the previous work, the ultimate goal

of improving the accuracy of acoustic features and enhancing the modification method are obtained which gives a new method of synthesizing emotional speech considering human perception. The converted emotional speech can give a similar expression as intended and good naturalness which is a great improvement. The emotional voice conversion system can be put into many applications such as Speech to Speech Translation System and Story Teller System which can give a great improvement to human daily life.

Bibliography

[1] Akagi, M., Han, X., Elbarougy, R., Hamada, Y., & Li, J. "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," Proc. APSIPA2014, CD-ROM, Siem Reap, Cambodia, 2014.

[2] Fujisaki, Hiroya. "Information, prosody, and modeling-with emphasis on tonal features of speech," Speech Prosody 2004, International Conference. 2004.

[3] Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., & Macias-Guarasa, J. "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," Speech Communication, 52(5): 394-404, 2010.

[4] Yamagishi, J., Nose, T., Zen, H., Ling, Z. H., Toda, T., Tokuda, K.. & Renals, S. "Robust speaker-adaptive HMM-based text-to-speech synthesis," IEEE Transactions on, Audio, Speech, and Language Processing, 17(6), 1208-1230, 2009.

[5] Toda, T., & Tukuda, K.,S ."A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," IEICE TRANSACTIONS on Information and Systems 90.5, 816-824, 2007.

[6] Albrecht, I., Schröder, M., Haber, J., & Seidel, H. P. "Mixed feelings: expression of non-basic emotions in a muscle-based talking head," Virtual Reality, 8(4), 201-212, 2005.

[7] Schröder, M, et al. "Acoustic correlates of emotion dimensions in view of speech synthesis," Proc INTERSPEECH. 2001.

[8] Grimm, Michael, and Kristian K. "Emotion estimation in speech using a 3d emotion space concept," INTECH Open Access Publisher, 2007.

[9] Hamada, Y., Elbarougy, R., & Akagi, M. "A method for emotional speech synthesis based on the position of emotional state in Valence-Activation space," Proc. APSIPA2014, CD-ROM, Siem Reap, Cambodia, 2014.

[10] Scherer, K.R., "Personality Inference from Voice Quality: The Loud Voice of Extroversion," European Journal of Social Psychology, 8, 467-487, 1978

[11] Huang, C-F. and Akagi, M. "A three-layered model for expressive speech perception," Speech Communication 50, 810-828, 2008.

[12] Huang, C. and Akagi, M. "The building and verification of a three-layered model for expressive speech perception," Proc. JCA2007,CD-ROM, 2007.

[13] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system," IEEE Transactions on Systems, Man and Cybernetics, 23.3, 665-685, 1993.

[14] Nauck, Detlef, Frank K, and Rudolf K. "Foundations of neuro-fuzzy systems," John Wiley & Sons, Inc., 1997.

[15] Hamada,Y, Xue,Y and Akagi,M. "Study on method to control fundamental frequency contour related to a position on Valence-Activation space," WesPac, Singapore, Singapore, P12000176, 2015.

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

[17] Elbarougy R, Akagi, M. "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model," Proc. of APSIPA CD-ROM, Los Angers, USA, 2012.

[18] Pereira, Cécile. "Dimensions of emotional meaning in speech," ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion, Belfast, 2000.

[19] Russell, James A., and Albert Mehrabian. "Evidence for a three-factor theory of emotions," Journal of research in Personality 11.3, 273-294, 1997.

[20] Wolkenhauer, Olaf. Data engineering: fuzzy mathematics in systems theory and data analysis. John Wiley & Sons, 2004.

[21] Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," Speech communication, 27(3), 187-207, 1999.

[22] Maekawa,H. "Production and perception of paralinguistic information," Proc of Speech Prosody, Nara. 367-374, 2004.

[23] Van Son, R. J. J. H., and Pols, L. "An acoustic description of consonant reduction," Speech communication 28.2, 125-140, 1990.

[24] Menezes. C, Maekawa. K, and Kawahara. H, "Perception of voice quality in paralinguistic information types," Proc. of the 20th General meeting of the Phonetic Society of Japan. 153-158, 2006.

[25] Vlasenko, Bogdan, et al. "Vowels formants analysis allows straightforward detection of high arousal emotions," Multimedia and Expo (ICME), 1-6, 2011.

[26] Agiomyrgiannakis, Yannis, and Olivier R. "ARX-LF-based source-filter methods for voice modification and transformation," Proc. ICASSP, Taipei, Taiwan, 3589-3592, 2009.

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

[28] Akagi, M. and Tohkura, Y. "Spectrum target presiction model and its application to speech recognition," Computer Speech and Language, 4, Academic Press 325-344, 1990.

[29] Li X-F., & Akagi, M. "Study on estimation of bilingual speech emotion dimensions using a three-layered model," ASJ Autumn meeting, 1-Q-39, 2015.

Acknowledgements

Time flies. Two years have past since the time when I first stepped into Japan is still in my mind. Fot the two years, lots of people gave me so many helps which let me felt I am not in the foreign country.

Firstly, I would like to express the sincere gratitude to Professor Akagi who gave me lots of help when I am in Jaist in the years. Two years before, it is the first time I leave my parents and come to an unfamiliar environment as from the elementary school to undergraduate school, I was all in Dalian, a city in the north part of China. Unknown people around me, unfamiliar language around me and unexpected prices around me make me feel helpless and loneliness. It is Professor Akagi who gives me much help not only from the research but also support me in daily life. I learned many things from Professor Akagi not only about the signal processing but also about how to be a humble person. Although Professor Akagi is already achieved many accomplishments in his life but I can not feel the distance when I ask some simple question from him. Of course, I would like to avoid to ask the simple questions again. All I want to say is thank you very much!

I also want to say thank you to Associate Professor Unoki. He gave me lots of useful suggestions when I gave a representation in weekly meeting. Sometimes I did not realize the problem in my research and presentation method but it is Unoki teacher who told me directly which helps me to perfect my research.

Lastly, I would like to express my thanking for my family and friends both in China and Japan. As we known, when we are far from family, friends are just family member who can share the happiness and sadness around us. What's more, the encouragement from my boyfriend also gives much strength when I faced trouble. Thank you for your love. For my parents, they are the ones who always thought I am the best, I am the person who can get happiness and I missed them so much. Without them, I not not enjoy the two years' life in my master at the best age of my life.

Thank you, all!

Publications

[1] Yawen Xue, Ysuhiro Hamada, Masato Akagi. "Emotional speech synthesis system based on a three-layered model using a dimensional approach," 'Asia-Pacific Signal and Information Processing Association (APSIPA)', Hang Kong, December, 2015 (submitted reviewed).

[2] 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.

[3] Yawen Xue, Ysuhiro Hamada, Masato Akagi. "A method for synthesizing emotional speech using the three-layered model based on a dimensional approach," '2015 ASJ Fall Meeting', 1-Q-40, Aizu, Fukusima, September, 2015 (Appeared, non-reviewed).

[4] Yawen Xue, Yasuhiro Hamada, Masato Akagi. "Rule-based emotional voice conversion utilizing three-layered model for dimensional approach," 'The Taiwan/Japan Joint Research Meeting on Psychological & Physiological Acoustics and Electroacoustics', Taiwan, October, 2015 (Appeared, non-reviewed).