Title	スピーチにおける感情的な知覚の多層ファジィ論理的なモデルの構築
Author(s)	黄，純芳
Citation	
Issue Date	2004-09
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1884
Rights	
Description	Supervisor:赤木 正人，情報科学研究科，修士

A MULTI-LAYER FUZZY LOGICAL MODEL FOR EMOTIONAL SPEECH PERCEPTION

By Chun-Fang Huang

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, 2004

A MULTI-LAYER FUZZY LOGICAL MODEL FOR EMOTIONAL SPEECH PERCEPTION

By Chun-Fang Huang (210204)

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
Professor Jianwu Dang
Professor Makoto Miyahara

August, 2004 (Submitted)

ABSTRACT

This thesis proposes a multi-layered perceptual model, which attempts to simulate human ability to perceive emotions from speech. Differing from most existing studies that deal with the relationship between acoustic features measured in speech signals and emotions the speech intended, the proposed perceptual model included 3 layers, emotion, primitive feature, and acoustic feature, where primitive feature is defined as an adjective that is used to describe emotional speech. To accomplish the perceptual model, two approaches should be considered. The top-down approach is adopted to construct a framework for the model, and the bottom-up approach is adopted to verify the existing framework. This thesis focuses on the top-down approach, mainly

The purpose of this thesis is to construct the proposed multi-layered perceptual model by the top-down approach. Two relationships, the emotion and the primitive feature, the primitive feature and the acoustic feature, are investigated sequentially. With regard to the relationship between the emotion and the primitive feature, three perceptual experiments are conducted. The first experiment examines the utterances in terms of the emotions. The other two experiments, although with different purposes, find suitable primitive features collectively. One is to construct the psychological distance model, and the other is to evaluate the adjectives. Finally, the relationship between the emotion and the primitive feature is built by applying fuzzy logic with these experiment results. With regard to the relationship between the primitive feature and the acoustic feature layer, two attributes of sound, pitch and loudness, are investigated. According to the analytic results, part of the relationship is build.

Regarding future work, more analysis of acoustic features is needed and also the relationship between the primitive feature and the acoustic feature should be refined by fuzzy logic. Moreover, the model should be verified by bottom-up approach. The significance of the perceptual model is that it clarifies the human ability to perceive emotional speech from an engineering perspective and it also covers the vagueness of human nature. It opens a possibility to many fields of practical applications, for example text-to-speech processing, emotional morphing processing, or improving human-machine interface to create a “human interface”.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

This chapter describes the background and previous work of this thesis, and what gap this thesis would like to fill. It states the overall purpose and the tasks to perform. Finally it gives the following contents of the thesis.

1.1 Background

Speech is one of the most convenient and important ways we humans use to communicate with each other. Apparently we do not use only linguistic meaning to convey our feeling but we also consciously or unconsciously inject our emotions into speech. When a listener perceives either consciously or unconsciously what emotions a speaker would like to convey, he[1] then decides either consciously or unconsciously how to react to the speaker in a variety of ways. Such interactive paths establish our daily communication. Interpreting emotions from the linguistic form of speech, i.e., words, may be relatively straight-forward. It is generally thought there are relatively unique mappings between the linguistic word and its meaning(s). However, emotions that are injected into speech via not words per se, but changes in the acoustic properties of the speech, i.e., changes in intonation, pitch, duration, loudness, etc., are totally different. There is not necessarily a straight-forward connection between these types of emotions injected into speech and the perception by listeners of these emotions.

We speak and perceive emotions from speech all the time in daily life. However, the perception of emotions from speech is such a mysterious ability of a human that it is still difficult to give it satisfactory explanation. If there is a model that could explain and simulate the ability from an engineering approach, then the model could be applied to many applications to benefit the human community, for example, improvement of human-machine interface to create a “human interface” [1], text-to-speech processing, or emotional morphing processing.

---

[1] It did not imply the listener was male only; “he” or “his” was used hereafter for convenient.

1.2 Previous Work

Concerning perception of emotional speech, two areas of previous research are reviewed here. One is with regard to emotions versus acoustic features. It has been studied most. The other is with regard to sound attributes versus acoustic features. It relates to perception in particular. The following section discusses some previous research of these two areas.

1.2.1 Emotions versus acoustic features

A number of studies attempted to identify and measure acoustic parameters in speech signals that reflect emotional states of speakers. For example, Williams and Stevens studied recordings of pilots speaking directly prior to a fatal crash (fear) or the classic Hindenburg radio announcement (anguish), and compared the simulated emotions of anger, fear, sorrow and “neutral” and the live recording of the Hindenburg crash broadcast. They found that different emotions had different effects on fundamental frequency (F0), which included its average value, the average pitch range, the characteristic shape of the contour, the rate of F0 change along the contour – and the speech rate [2][3]. Hiratate studied the relationship between two categories of anger, cold anger and hot anger, and acoustic features, which included the changing rate of F0 and power at the accent portions, and the duration of vowels in accent portions [4].For emotional speech synthesis, Cahn [5] considered speak model and perceptual model for representation of emotional state. He applied the relationship between emotions and acoustic features in terms of pitch, timing, and voice quality to build the perceptual model.

Their studies well reported either those certain acoustic features would possibly affect perception of emotion or that certain emotional speech has certain acoustic features. Their approaches mainly concerned the direct relationship between emotions and acoustic features. However, even more acoustic features were found to be related to emotions, they all overlooked one important point. Human ability to percept emotions from speech was not directly based on acoustic features. From observations of perception emotional speech, “his voice sounds dark” might lead to a judgment about “he is sad”. Direct studying the relationship between these two factors still could not provide a complete and appropriate solution for solving emotional speech. Some other factor, the linguistic meaning to describe voice, in the middle between them was overlooked.

1.2.2 Sound attributes versus acoustic features

Regarding the perception of sound attributes, Ueda and Akagi [6] reported how amplitude envelope shapes, sound-pressure level, and duration affect perception

attributes, sharpness and brightness, and found a better definition of sharpness and brightness. Regarding to others perception of sound attributes, for example pitch and loudness, Zwicker have been reported a number of studies about the relationship between them and acoustic features [7]. Researches in this area well explained the relationship between the perception of sound attributes and the acoustic features. It was not specific to emotional speech, but it could help to bridge the gap of the current research, the middle factor between acoustic features and emotions.

1.2.3 Perception and Fuzzy logic

Most of studies that analyzed experimental results were by using statistics. Different from using statistical approach, Massaro proposed a fuzzy logical model of perception (FLMP) to solve vagueness of human nature because fuzzy logic provide a natural representation of the degree of match [8]. He showed how fuzzy logic model could be implemented in modeling perceptual system and how audible and visible sources are so naturally coordinated by the perceptual system. The study of this thesis was not focused on collaboration between audible and visible perception, and the idea still applied. To explain human nature in a “precise” and “rigorous” way, fuzzy logic corresponded to the needs.

1.3 Gap in the Research

According to previous research, apparently, using only the relationship between emotions and acoustics feature is not enough. Linguistic form to describe voice in the middle exists in human emotional perception. The studies of the relationship between emotions and linguistic form and the relationship between linguistic form and acoustic features are needed. The relations in these two relationships may be multiple and may be related to other aspects of knowledge. The vagueness of human nature should also be explained. Traditional mathematical model is insufficient. Fuzzy logic should be pulled in to make explanations sounds.

A model which explains and simulates human ability to perceive emotional speech from an engineering perspective but also covers fuzziness within human nature in the perceptional process is necessary.

1.4 Purpose and Tasks of the Current Research

The purpose of the research is to construct a model by which a human’s ability to perceive emotions from speech can be simulated. When such a perceptual model is

accomplished, it could be applied to other practical applications, for example text-to-speech processing or emotional morphing speech processing. For the purpose of this study, there are four tasks:

- To examine the voice data
- To investigate primitive features
- To build fuzzy inference systems
- To analyze acoustic features

In the following sections, the perceptual model of emotional speech is introduced first. Then the details of the tasks are described.

1.4.1 Multi-layered perceptual model of emotional speech

This thesis proposes a multi-layered perceptual model by which a human's ability to perceive emotions from speech could be simulated. Differing from previous research, the multi-layered perceptual model involves with not only emotions and acoustic features, but also the original sound attributes, namely primitive features. By studying sound attributes that we human originally perceive from sound, it provides a useful way to understand the perception of emotions from speech and to be able to simulate it. One other character of the proposed model is it will deal with the relationships by using fuzzy logic, since many physical meanings of fuzzy logic are just corresponded to the property of perception, for example, it deals with the vagueness and linguistic form.

The conceptual model consists of three layers, emotion, primitive feature, and acoustic feature. It is illustrated in Figure 1-1. The emotion represents emotions in speech a listener perceives. There are 5 categories of emotion, Neutral, Joy, Cold Anger, Sad, and Hot Anger, which are studied in this thesis. The primitive feature contains a set of linguistic form of adjectives. The adjectives are used by listeners to describe emotional speech which he listened to, for example, high, low, quiet etc. The acoustic feature concerns the measurable physical features in speech such as average power, highest pitch, vowel duration etc.

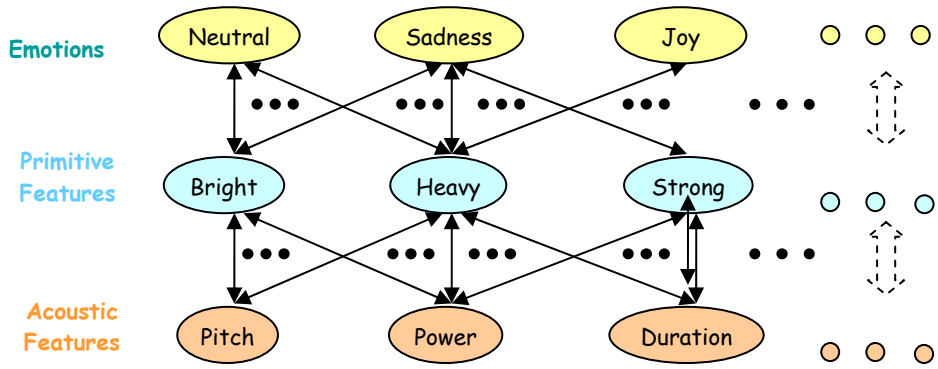


Figure 1-1 Conceptual Diagram of Multi-Layered Perceptual Model for Emotional Speech

The model is inspired by the observations of linguistic form of interpreting emotions from speech. When listening to a voice of a speaker, a receiver first senses "the voice sounds dark" or "the voice sounds very blue", and then we interpret "I think he is sad". We never say "the voice sounds fundamental frequency is 400 Hz". The observations indicate three points.

1. Primitive features are the intermediate relationship between emotions and acoustic features
2. Humans describe their perception of phenomenon with vague linguistic form but not precise value. The vagueness nature of humans should be considered.
3. Although human nature is vague, a precise analytical/mathematic approach to deal with the vagueness of humans is needed.

The first point results in the three-layered architecture of the model. The last two points result, results in the application of fuzzy logic. Linguistic form is vague, uncountable, and also it is related to human knowledge of communication. The characteristics of linguistic form are corresponded to the concepts of fuzzy logic. More detailed explanation of applying fuzzy logic is provided in Chapter 4.

1.4.2 How to construct?

To build the perceptual model, the top-down approach and the bottom-up approach should be adopted. The top-down approach is used to construct the framework of the proposed model. The framework contains the relationship between the emotion and the primitive feature, and the relationship between the primitive feature and the acoustic feature. After the framework is constructed, the bottom-up approach will be adopted to verify the framework.

With regard to the top-down approach, firstly, the voice database in terms of emotions should be examined by perceptual experiments. Secondly, what adjectives should be taken as the members of the primitive feature should be investigated by conducting two perceptual experiments. Thirdly, fuzzy inference systems should be built to describe the relationship between the emotion and the primitive feature. Finally, how acoustic features affect each primitive features should be analyzed.

1.5 Organization of the Thesis

The remainder of this thesis is organized as the following.

Chapter 2, Experiment 1: Voice Data Examination, describes the details of a perceptual experiment. The experiment is used to examine each sentence in the voice data in terms of 5 emotions.

Chapter 3, Finding Primitive Features, presents the details of two perceptual experiments which are used to decide what adjectives are suitable to be the member of the primitive feature. One experiment is conducted for constructing a psychological distance model and the other is to rate adjectives.

Chapter 4, Fuzzy Inference System, describes construction of fuzzy inference systems. The reason why fuzzy logic is used to model the relationship between the emotion and the primitive feature is stated. The process of the construction of the fuzzy inference systems is presented.

Chapter 5, Acoustic Features Analysis, discusses the analysis of acoustic features.

Chapter 6, Conclusion, is the conclusion of this thesis. It summarizes what has been done, the significance of current research, and future work.

Appendix A, Supplement Data of Experiment Results, supplies supplement data results of experiments conducted in this thesis.

Chapter 2

Experiment 1: Voice Data Examination

In order to build the framework for the purposed perceptual model, first of all, the utterances in the voice database should be examined in terms of how the listeners rate their perception of the emotions. The rating results would be utilized in choosing the appropriate utterances concerning emotions in the subsequent experiments. This chapter describes details of the perceptual experiment that was carried out to examine the utterances in terms of 5 emotions. The purpose, the experiment setting, the experiment method, the results and their discussion are presented.

2.1 Purpose

The purpose of the experiment was to choose appropriate utterances from the voice database for further perceptual experiments.

2.2 Corpus, Subjects, and Equipments

2.2.1 Corpus

The corpuses used in the experiment were selected from a voice database produced by Fujitsu laboratories. There were 19 sentences. Each sentence was uttered in 5 emotions, which were Neutral, Joy, Cold Anger, Sadness, and Hot Anger. Neutral had one utterance and others had two utterances for each. Thus each sentence was represented by nine utterances. A total of 171 utterances were used in the experiment and all of them were recorded by professional voice actresses in Japanese. The detailed information of the corpus is shown in Table 2-1, Table 2-2, and Table 2-3.

Table 2-1 Specifications of Voice Data

Item	Value
Sampling frequency	22050 Hz
Quantization	16bit
Sentences	19

Table 2-2 19 Sentences Used in Experiment 1

This table lists all sentences used in Experiment 1. First column shows the id numbers of the sentences. Second column shows the linguistic format of the sentences in Japanese. Id 14 was not in the database.

Id	Sentence
1	新しいメールが届いています
2	頭にくることなんてありません
3	待ち合わせは青山らしいんです
4	新しい車を買いました
5	いらないメールがあったら捨てて下さい
6	そんなの古い迷信です。
7	みんなからエールが送られたんです。
8	手紙が届いたはずです。
9	ずっとみています。
10	私のところには届いています。
11	ありがとうございました
12	申し訳ございません
13	ありがとうは言いません
15	気が遠くなりそうでした。
16	こちらの手違いもございました。
17	花火を見るのにゴザがいりますか。
18	もうしないと言ったじゃないですか。
19	時間通りに来ない訳を教えてください。
20	サービスエリアで合流しましょう。

Table 2-3 171 Utterances Used in Experiment 1

First column shows the utterances id numbers. Second column shows the intended emotions. The UID is composed of a letter and a numerical code. The letter represented what emotion it was (a: Neutral, b, c: Joy, d, e: Cold Anger, f, g: Sadness, h, i: Hot Anger) and the numerical code presented what sentence it was. a014, b014, c014, d014, e014, F014, g014, h014, and i014 were not used.

UID	Emotion
a001 ~ a020	Neutral
b001 ~ b020	Joy
c001 ~ c020	
d001 ~ d020	Cold Anger
e001 ~ e020	
F001 ~ F020	Sadness
g001 ~ g020	
h001 ~ h020	Hot Anger
i001 ~ i020	

2.2.2 Subjects

The subjects were 12 graduate students. Average age was 25 and the difference between maximum and minimum was 4. No one had had a hearing disease before. All were native Japanese speakers.

2.2.3 Equipments

All experiments that were conducted in this thesis used identical equipment. One sound proof room was used. The detailed information of the other equipment was listed in Table 2-4, and, the configuration of the equipment was depicted in Figure 2-1.

Table 2-4 Equipment of Experiment 1

Equipment	Specification
Server for sound presenting	DAT + LINK & Work Station (Linux)
Headphone	STAX SR-404
Headphone amp	STAX SRM-1/MK-2
D/A converter	STAX DAC-TALENT BD.

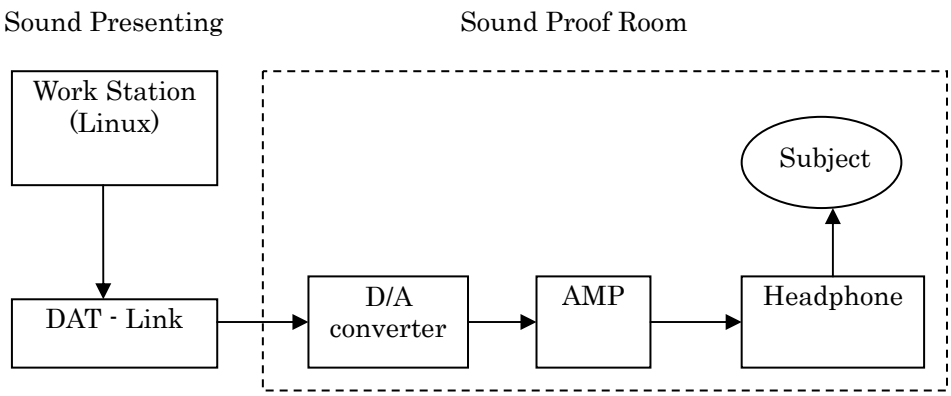


Figure 2-1 Configuration of the Equipment Used in All Experiments of this thesis

2.2.4 Method

The subjects were asked to rate 171 utterances in evaluation forms according to the perceived degree of 5 emotions. There was a total of 5 points for one utterance. That is, if a subject perceived that an utterance belonged to one emotion without any doubt then he gave it 5 points. Conversely, if a subject was confused within two or even more emotions then he divided 5 points into these emotions depending on the degree to which he perceived the utterance belonged to the emotions. Every utterance was presented 2 times with a pause of 2 sec. 9 utterances of one sentences was presented continuously but in a randomized order. One sample evaluation form for the sentences is shown in Figure 2-2.

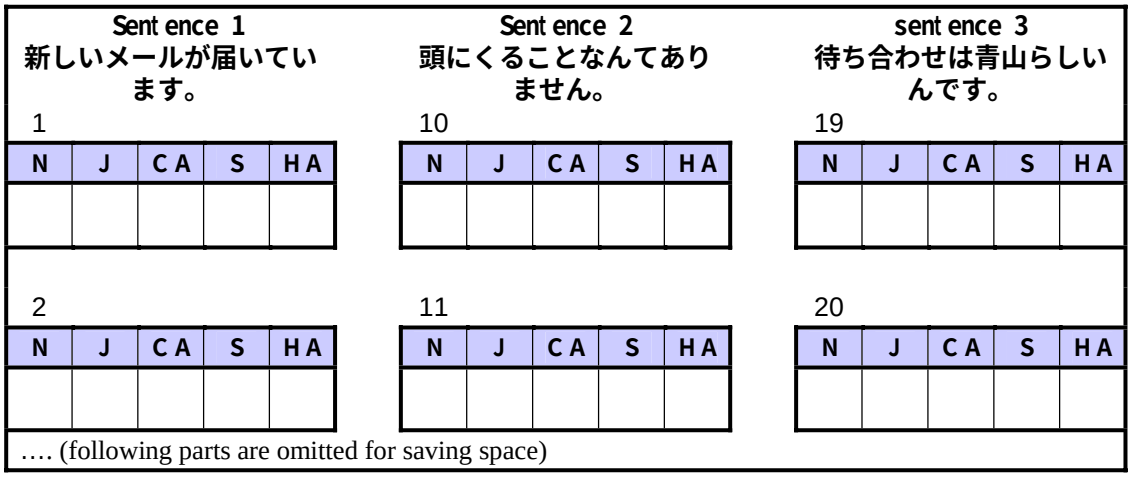
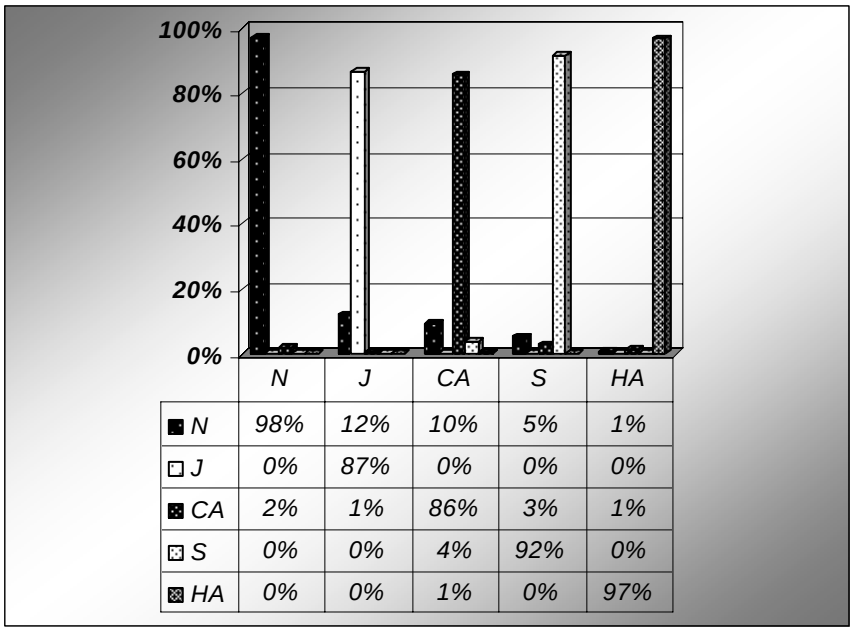


Figure 2-2 One Sample Evaluation Form Used in Experiment 1

2.3 Results

With regard to the rating results, the number of times each of the 171 utterances was judged as containing one of the 5 emotions, Neutral (N), Joy (J), Cold Anger (CA), Sadness (S), and Hot Anger (HA) was count. The results are presented in Section A-1 (on page 62). In addition, rating percentages in terms of the intended emotions of N, J, CA, S, and HA are shown in Figure 2-3. As the results show, most of the utterances could be perceived according to the intended emotions with very high percentage. Cold Anger had the lowest percentage and it was confused with Neutral easily. However, the one most confused with Neutral is Joy. The rating result will be used when considering which utterance should be selected as corpus in the following experiments.



	N	J	CA	S	HA
N	98%	12%	10%	5%	1%
J	0%	87%	0%	0%	0%
CA	2%	1%	86%	3%	1%
S	0%	0%	4%	92%	0%
HA	0%	0%	1%	0%	97%

Figure 2-3 Percentage of ratings of the 5 intended categories

The 5 rows were the 5 intended emotions and the 5 columns were the same 5 emotion rated by subjects.

Chapter 3

Primitive Features

Regarding the purposed perceptual model, each of the 171 utterances had been rated in terms of 5 emotions by the previous perceptual experiment. A key problem is the following. What should be the primitive features for the purposed perceptual model? In order to deal with this question, two experiments are conducted to choose suitable adjectives as the members of primitive features.

Chapter 3 gives the detailed descriptions of two experiments. The first experiment is to construct a psychological distance model of 15 utterances by applying a multidimensional scaling technique. This experiment is carried out to help with selecting suitable adjectives as primitive features in the second experiment. The second experiment is to rate adjectives. According to rating results and directions of all adjectives that are superimposed onto the psychological distance model, 15 adjectives are selected as suitable primitive features for the purposed perceptual model.

3.1 Experiment 2: Construction of Psychological Distance Modal

A psychological distance model was considered as a model that illustrated the similarity among stimuli. The 15 utterances chosen according to the results of the first experiment were used. By asking subjects to rate how similar pairs of the 15 utterances were, this experiment constructed a psychological distance model for these 15 utterances. The resultant psychological distance model represented the similarity among the utterances. Moreover it also showed the positions of 5 emotions within it. The resultant model was primarily used to help with choosing suitable adjectives to be primitive features in the next experiment.

3.1.1 Purpose

The purpose of this experiment was to construct a psychological distance model of 15 utterances, with 3 utterances for each of 5 emotions. It was the first step for selection of suitable primitive features.

3.1.2 Corpus, subjects, and equipments

Corpus

In the results of the first experiment, 171 utterances were rated according to the perceived degree of emotion for each of the 5 emotions. The utterance with a higher perceived degree of emotion was considered to be more typical for a certain emotion. In this experiment, 15 utterances were chosen from the 171 utterances as corpus. Detail descriptions of the corpuses are shown in Table 3-1.There were 3 utterances for each of the 5 emotions, the most typical utterance (Perceived Degree = 0.95 ~ 1), the middle typical utterance (Perceived Degree = 0.8 ~ 0.95), and the least typical utterance (N3, J2, C1, S3, and H3).

Table 3-1 Detail Descriptions of Corpuses

First column represents id numbers. Second column represents emotions. Third column represents utterances id numbers. Forth column represents perceived degrees of 15 utterances used in Experiment 2.

ID	Emotion	UID	Perceived Degree
N1	Neutral	a001	1
N2	Neutral	a002	0.945455
N3	Neutral	a009	0.972727
J1	Joy	b001	1
J2	Joy	b019	0.640909
J3	Joy	c016	0.845455
C1	Cold Anger	d003	0.686364
C2	Cold Anger	e010	0.859091
C3	Cold Anger	e018	0.959091
S1	Sadness	F009	0.872727
S2	Sadness	g004	0.981818
S3	Sadness	g017	0.772727
H1	Hot Anger	h001	0.927273
H2	Hot Anger	h002	1
H3	Hot Anger	h004	0.763636

Subjects and equipments

The subjects and the equipment were identical to Experiment 1.

3.1.3 Method

The method of paired comparison was implemented in this experiment. Since there were 15 utterances in the experiment, taking each pair within the15 utterances as stimuli, there were a total of 15 x 14 = 210 pairs as stimuli. The subjects were asked to rate each utterance pair according to how similar they perceived the pair. There were 3 levels between the two extreme ends, totally similar and totally different, making a total of 5 levels. Before each utterance pair was presented to the subjects, a notification “bee” tone was presented. Each utterance pair was presented with a gap of 2 sec. The whole experiment was taken by an evaluation program written in Visual Basic 6.0 and a screenshot of the program is shown in Figure 3-1. As the screenshot shows, the subjects were allowed to listen to one utterance pair again by clicking the “Retry” button. The experiment was conducted 5 times.

Figure 3-1 Screenshot of the Evaluation Program in Experiment 2

The radio buttons were used by the subjects to evaluate one utterance pair. There were 5 levels including the two extreme ends. The “Retry” button was used by the subjects to listen to one utterance pair again. The “Next” button was used to move to next utterance pair.

3.1.4 Data analysis and results

This experiment evaluated the similarity among 15 utterances. A number of studies have

reported that the perceptual space is multidimensional [9][10][11][12][13]. Since there is a possibility that the perception of similarity had multidimensional characteristics, the experimental results were mainly analyzed by multidimensional scaling techniques (MDS) [14][15]. A statistical application, SPSS 11.0J for Windows was used and Kruskal's method was applied [16]. It was convenient that MDS could provide a visual presentation of the model of similarities among these 15 utterances. Finally, the result was a psychological distance model which described the similarities among the utterances.

To build the psychological distance model according to the rating results, there were two steps.

Firstly, a psychological distance matrix for 15 sentences was formed according to the rating results. The psychological distance matrix was a dissimilarity matrix as shown in Table 3-2. In this matrix, the number represented the distance between two utterances. In other words, a larger number indicated a longer distance between two utterances, i.e. less similarity between them. Conversely, a smaller number indicated a shorter distance between two utterances, i.e. more similarity between them.

Secondly, the psychological distance matrix calculated in the first step was analyzed by MDS in SASS which resulted in the psychological distance model.

Table 3-2 Psychological Distance Matrix of 15 Utterances

In the matrix, larger numbers indicate less similarity between two utterances. Conversely, smaller numbers indicate more similarity.

	N1	N2	N3	J1	J2	J3	C1	C2	C3	S1	S2	S3	H1	H2	H3
N1	0														
N2	1.05	0													
N3	1.12	1.02	0												
J1	4.25	4.2	4.27	0											
J2	2.79	2.73	3.05	1.98	0										
J3	4.4	4.41	4.43	1.19	2.23	0									
C1	4.18	3.94	4.15	4.83	4.68	4.94	0								
C2	4.36	4.33	4.37	4.96	4.68	4.9	1.83	0							
C3	4.33	4.27	4.1	4.95	4.77	4.92	1.39	1.11	0						
S1	4.35	4.18	4.4	4.96	4.76	4.93	4.11	3.53	3.78	0					
S2	4.41	4.43	4.44	4.87	4.55	4.78	4.41	4.04	3.99	2.01	0				
S3	4.22	4.27	4.4	4.65	4.47	4.63	4.62	4.33	4.16	2.53	1.42	0			
H1	4.37	4.43	4.5	4.73	4.8	4.84	3.01	3.18	3.38	4.86	4.82	4.93	0		
H2	4.74	4.7	4.64	4.83	4.85	4.76	3.23	3.52	3.38	4.92	4.89	4.94	1.08	0	
H3	4.58	4.47	4.63	4.65	4.77	4.76	3.26	3.51	3.41	4.92	4.93	4.97	1.17	1.11	0

By MDS, as implied by the name of the analysis technique, data results could be represented by multi-dimensions. For MDS, orientation and how many dimensions used to specify the resultant distant model should be considered. Following are discussions about it.

First, orientation of presentation was arbitrary within MDS. The most important set of information was the distance among instances (i.e. utterances) that indicates similarity between them.

Second, concerning how many dimensions should be used to specify the resultant distant model, the stress measure was the most common measure, that was used to evaluate how well (or poorly) the model reproduced the distance matrix. In other words, the stress value represented appropriate dimensions. The smaller the stress value, the more appropriate it was to reproduce the distance model to the distance matrix [15].

According to the above discussions, in order to decide how many dimensions

would be appropriate to use to present the resultant distance model, three criteria were taken into account:

1. Stress should be less then 10%
2. The dimension should be stable
3. Its presentation should be easy to comprehend

The results, the stress value against presenting dimension, are shown in Figure 3-2. Accordingly, 3-dimension with stress equal to 7 % was chosen to present the psychological distance model. The resultant psychological distance model is shown in Figure 3-3. Moreover, in order to look into information within 3-dimension, the 3-dimension model was illustrated by showing dimension 1 against dimension 2 in Figure 3-4, dimension 1 against dimension 3 in Figure 3-5, and dimension 2 against dimension 3 in Figure 3-6.

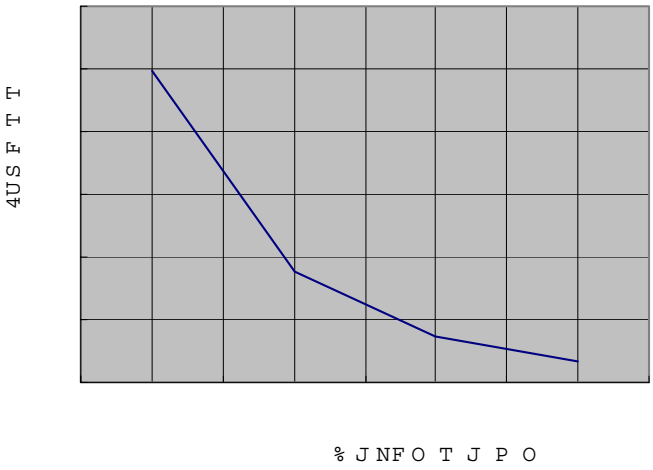


Figure 3-2 Stress by Dimension

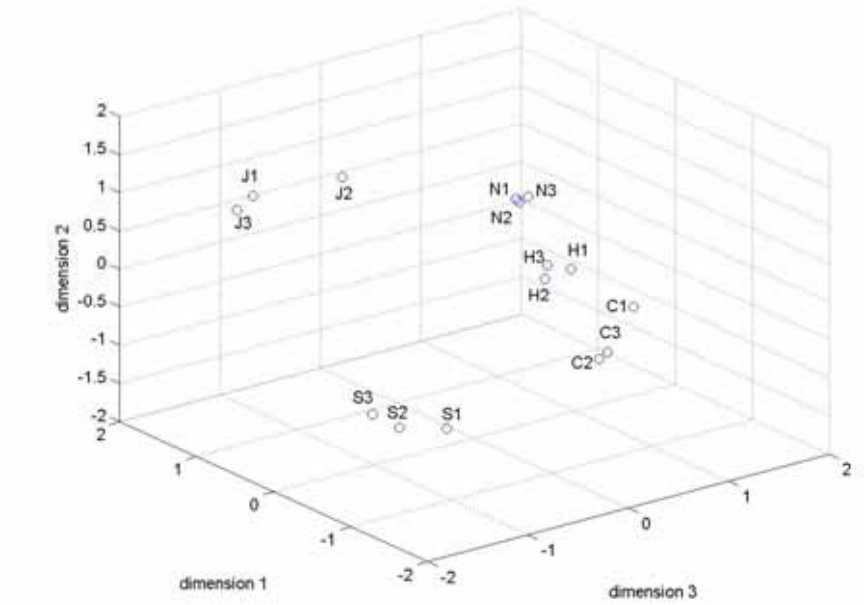


Figure 3-3 3-Dimensional Presentation of the Psychological Distance Model of 15 Utterances with 7% Stress

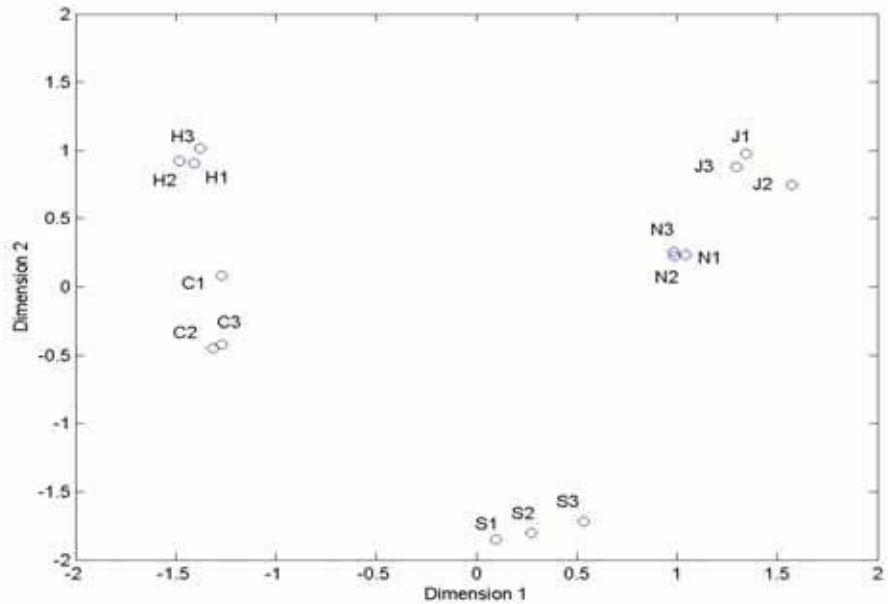


Figure 3-4 2-Dimensional (Dimension 1 against Dimension 2) Presentation of the Psychological Distance Model of 15 Utterances with 7% Stress

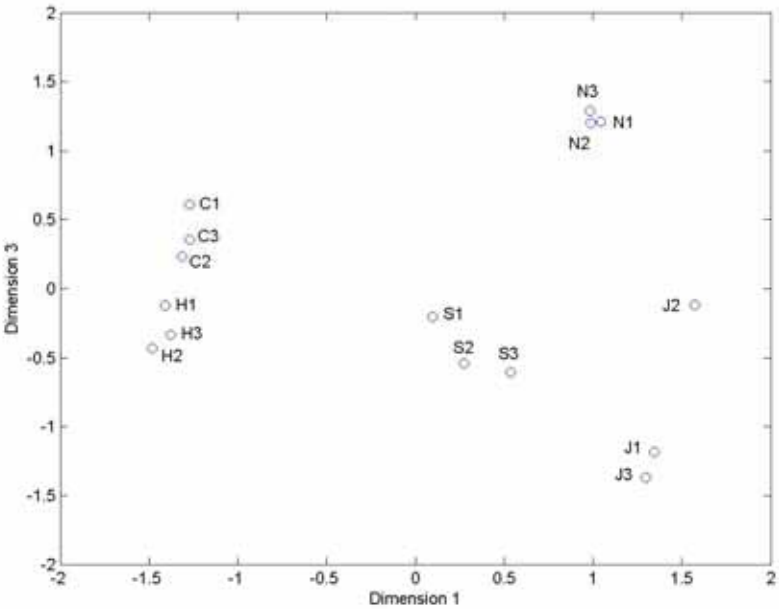


Figure 3-5 2-Dimensional (Dimension 1 against Dimension 3) Presentation of the Psychological Distance Model of 15 Utterances with 7% Stress

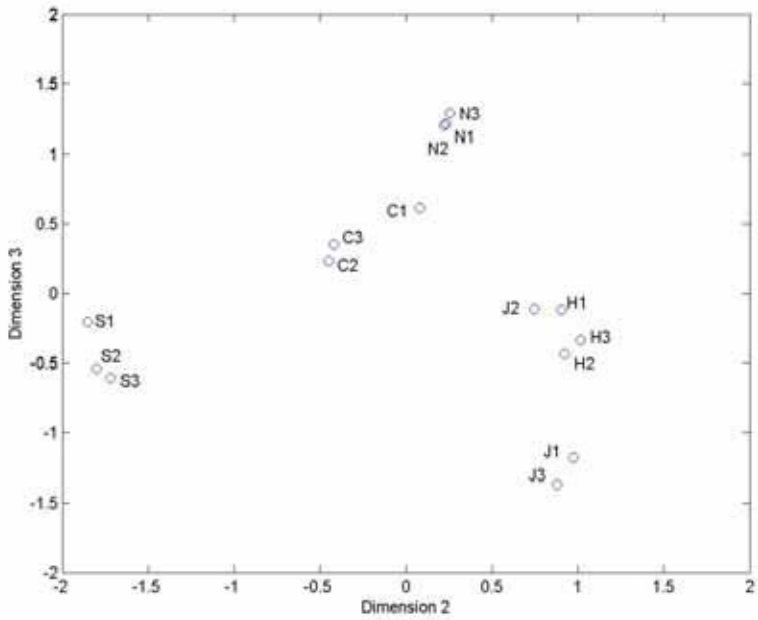


Figure 3-6 2-Dimensional (Dimension 2 against Dimension 3) Presentation of the Psychological Distance Model of 15 Utterances with 7% Stress

3.1.5 Discussion

A qualified psychological distance model could tell the similarity among instances correctly. That is, if two instances were very similar then they should be very close to each other. Conversely, if two objectives were very different then they should be a far distance from each other. The following discussions describe whether the resultant psychological distance model was qualified to describe the similarity among these 15 utterances.

By observing profiles shown in Figure 3-3, Figure 3-4, Figure 3-5, and Figure 3-6, it could be noticed that those sentences which had the same intended emotion clustered together. The details of each emotion are discussed below.

About Neutral (N1, N2, and N3), they clustered together in every profile. A similar phenomenon was also shown within Hot Anger (H1, H2, and H3). Such phenomenon could be explained in that Neutral and Hot Anger had a higher degree of perceptibility by listeners than the other emotions. That meant they were easier to be perceived by listeners.

About Joy (J1, J2, and J3), from Figure 3-5and Figure 3-6, it was noticed that J1 and J3 clustered together but J2 kept a slight distance from J1 and J3. Similarly, regarding Cold Anger, (C1, C2, and C3), C1 was somewhat distant from C2 and C3. In fact, it could be explained by comparing the degree of perceptibility by listeners of 3 utterances of Joy and Cold Anger in Table 3-1. It showed that J2 was the least typical sentence for Joy and C1 was also the least typical sentence for Cold Anger. The degree of perceptibility of both J2 and C1 was rather low compared to the other 2 utterances, in other words, J2 was not “so-Joy” and C1 was not “so-Cold Anger”. That was why J2 and C1 were not clustered with the other 2 more typical sentences.

About Sad (S1, S2, and S3), the distances among each sentence were bigger than that for the other emotions. It could be explained that it was because those sentences of Sad had lower degree of perceptibility by listeners than other emotions. In other words, Sad was more difficult to be perceived.

According to the discussions above, the psychological distance model was considered to be reasonable and qualified. In the next experiment, it will be applied to help find suitable primitive features.

3.2 Experiment 3: Rating Adjectives

Before performing Experiment 3, a pre-experiment was conducted and 34 adjectives were chosen from 60 adjectives. Then, in this experiment, the subjects were asked to rate the 34 adjectives. According to the combination of the rating results and how each

of the 34 adjectives were related to a certain emotion, depending on what direction an adjective was superimposed onto the psychological distance model, 15 adjectives were selected from the 5 emotions to be the members of the primitive features for the proposed perceptual model.

3.2.1 The purpose

The purpose of the experiment was to give each of 34 adjectives a rating in terms of how much appropriately it described each of the 15 utterances. It was the second step for selection of suitable primitive features.

3.2.2 A pre-experiment for choosing 34 adjectives from 60 adjectives

The purpose of the pre-experiment was to narrow down adjectives for Experiment 3. As mentioned in Chapter 1, the primitive features used in this thesis were actually adjectives that were suitable to describe emotional speech. Ueda proposed 50 adjectives that were considered most often used to describe tone and sound [17].The 50 adjectives were selected from 114 adjectives according to his previous research. There were so many adjective that could be possibly used to describe sound, tone, or voice. However, for this study what was needed should be those related to emotional speech. In order to remove unnecessary adjectives, it is necessary conduct a pre-experiment.

A total of 60 adjectives were chosen as candidates in the pre-experiment. Except the 50 adjectives proposed by Ueda, an extra 10 adjectives that were considered relevant to emotional speech were also added into.

The subjects listened to each of utterances and were asked to circle which adjective is appropriate to describe the utterances. As listed in Table 3-3, 34 adjectives that were relatively more frequently circled are chosen as the candidates for Experiment 3.

Table 3-3 34 Adjectives Chosen from Per-Experiment

ID	Adjective (Japanese)	Adjective (English)
1	明るい	bright
2	暗い	dark
3	声の高い	high
4	声の低い	low
5	強い	strong
6	弱い	weak
7	太い	thick
8	細い	thin
9	堅い	hard
10	柔らかい	soft
11	重い	heavy
12	軽い	light
13	鋭い	sharp
14	鈍い	dull
15	耳障りな	rough
16	流暢な	fluent
17	荒っぽい	violent
18	滑らかな	smooth
19	うるさい	noisy
20	静かな	quiet
21	ざわついた	noisy
22	落ち着いた	calm
23	落ち着きのない	unstable
24	きれいな	clean
25	汚い	dirty
26	濁った	muddy
27	明らかな	clear
28	あいまいな	vague
29	明瞭な	plain
30	かすれた	husky
31	抑揚のある	well-modulated
32	単調な	monotonous
33	早い	quick
34	ゆっくり	slow

3.2.3 Corpus, subjects, and equipments

The corpus, the subjects, and the equipments were the same as the previous experiments.

3.2.4 Method

The subjects were asked to rate each of 34 adjectives in terms of how appropriate the adjectives described the utterances they heard. It was a 4 point scale from 0 to 3, where “0” indicated that the adjective was inappropriate and “3”, the adjective was appropriate. The experiment was a one-side test. The experiment was taken by a program written in Visual Basic 6.0 and a screenshot of the program is shown in Figure 3-7. As the screenshot shows, the subjects were allowed to listen to one utterance again by clicking the “Retry” button. The experiment was conducted twice.

Figure 3-7 Screenshot of the Evaluation Program in Experiment 3

The radio buttons were used by the subjects to evaluate the adjectives. For each adjective, there were 4 levels covering two extreme ends. "Retry" button was used by the subjects to listen to one utterance again. "Next" button was used to move to next utterance.

3.2.5 Data analysis and Results

Section A-2 (on page 68) listed the rating results of 34 adjectives for each of the 15 utterances. In order to see how each adjective was related to emotional speech, 34 adjectives were superimposed onto the psychological distance model by multiple regression analysis. (3-1) was the regress equation.

$$y = a_1 x_1 + a_2 x_2 + a_3 x_3 \quad \textbf{(3-1)}$$

, where x_1, x_2, and x_3 were position (x_1, x_2, x_3) of one utterance in the psychological distance model, and y was the rating of an adjective against the utterance. Calculating regression coefficients a_1, a_2, and a_3, by performing a least squares fit, Figure 3-8, Figure 3-9, and Figure 3-10 present the 34 adjectives plotted within dimension-1 against dimension-2, dimension-1 against dimension-3, and dimension-2 against dimension-3, respectively, of the psychological distance model. A line in the plot indicates an adjective by marking its id number (id numbers are shown in Table 3-3). The pointed direction of the arrow head of each line indicated that the adjective was

increasingly related to the utterances. For example, the adjective “clean” (ID: 24) was more related to the utterance J2 than to the utterance C2. In the other words, the adjective “clean” was more related to Joy than to Cold Anger. In this way, it was possible to find which each adjective was related to which emotion. In addition, the multiple correlation coefficient of each adjective was calculated. According to

1. The direction of each adjective in the psychological distance model, and
2. The multiple correlation coefficient of each adjective,

15 adjectives listed in Table 3-4 are chosen as suitable primitive feature from 34 adjectives.

The detailed data of regression coefficients a_1, a_2, a_3, and multiple correlation coefficient of each adjective were listed in Section A-3 (on page 70).

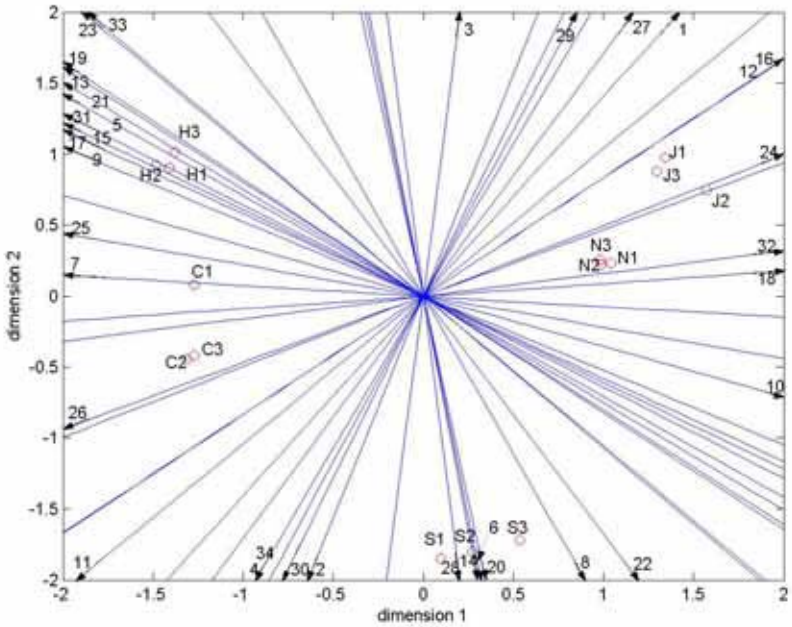


Figure 3-8 Direction of Each of 34 Adjectives of Dimension-1 against Dimension-2

The figure was plotted with arrow head lines in dimension-1 against dimension-2 of the psychological distance model.

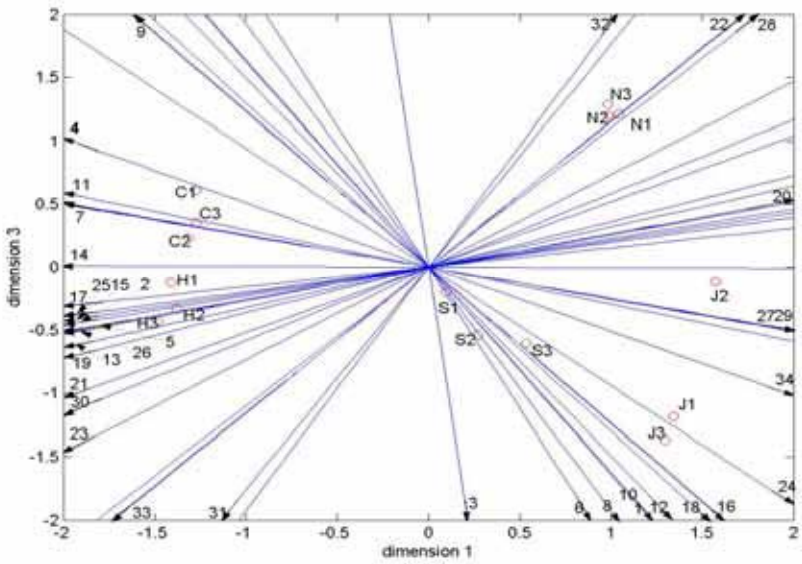


Figure 3-9 Direction of Each of 34 Adjectives of Dimension-1 against Dimension-3

The figure was plotted with arrow head lines in dimension-1 against dimension-3 of the psychological distance model.

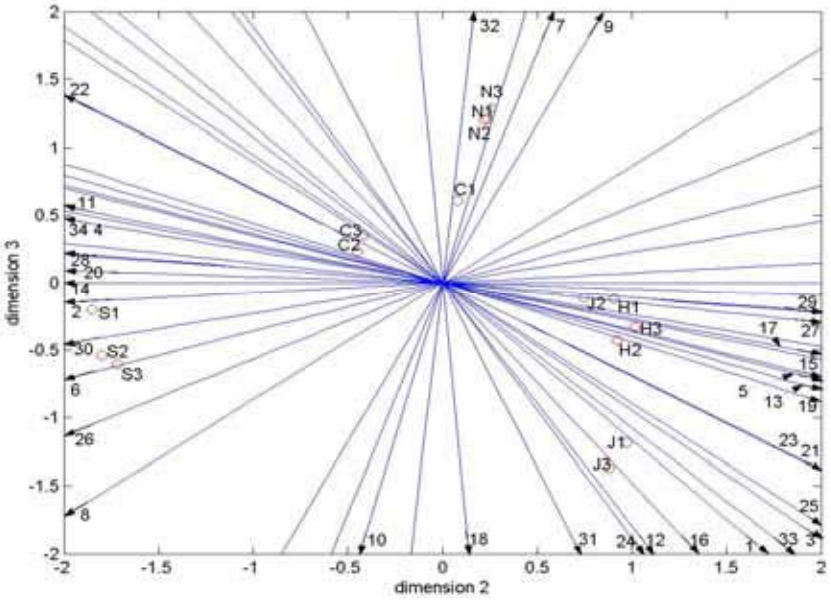


Figure 3-10 Direction of Each of 34 Adjectives of Dimension-2 against Dimension-3

The figure was plotted with arrow head lines in dimension-2 against dimension-3 of the psychological distance model.

Table 3-4 15 Primitive Features for the Proposed Perceptual Model

They were chosen from 34 adjectives in Table 3-3 according to multiple correlation coefficients and the direction in the psychological distance model.

ID	Adjective (Japanese)	Adjective (English)
PF1	明るい	bright
PF2	暗い	dark
PF3	声の高い	high
PF4	声の低い	low
PF5	強い	strong
PF6	弱い	weak
PF7	落ち着いた	calm
PF8	落ち着きのない	unstable
PF9	抑揚のある	well-modulated
PF10	単調な	monotonous
PF11	重い	heavy
PF12	明らかな	clear
PF13	うるさい	noisy
PF14	静かな	quiet
PF15	鋭い	sharp

Chapter 4

Fuzzy Inference System

Thus far, regarding the three-layered proposed perceptual model, the emotion and the primitive feature are investigated. The following task is to build fuzzy inference systems by which the relationship between the emotion and the primitive feature is established. A fuzzy inference system is the process of formulating the mapping from a given input to an output based on the concepts of fuzzy logic, i.e. fuzzy set theory

Chapter 4 gives explanations of fuzzy logic, how it can deal with fuzziness, why a fuzzy inference system is used to build the relationship, and finally describes how a fuzzy inference system of the perceptual model is built.

4.1 What is FIS?

Fuzzy logic was fist proposed by L.A. Zadeh in 1965. He developed the concept of linguistic variables, or fuzzy set in a 1973 paper. The human brain interpreted imprecise and incomplete sensory information provided by perceptive organs. Fuzzy logic provided a systematic calculus to deal with such information linguistically, and it performed numerical computation by using linguistic labels stipulated by membership functions. Moreover, a set of fuzzy IF-THEN rules formed the key component of a fuzzy inference system that could effectively model human expertise in a specific application [18].

Three relevant concepts within fuzzy logic were:

1. Fuzzy set, i.e. linguistic variables were defined as variables whose values were words or sentences.
2. Membership function mapped each element in a fuzzy set to a membership degree between 0 and 1.
3. IF-THEN rules, comprising the input (antecedent) and the output (consequent), were propositions containing linguistic variables.

Here was an example provided for briefly explaining what fuzzy set and membership functions were. Considering the definition of “hot” for temperature, for

instance, from 0 to 36 degrees, how many degrees is “hot”? The left of Figure 4-1illustrated what a classical set looked like. If the temperature degree was over 32, then it was hot, otherwise, it was not hot. That is, even a little difference, for instance 31.9 degrees was judged as “not hot at all”. Apparently it was not the way we humans perceive temperature. The right of Figure 4-1 illustrated what a fuzzy set looked like. In the fuzzy set, the judgment was not only 0 or 1. It was flexible between 0 and 1. The curve defined a function that mapped the input (temperature) to the output (degree of hotness). Specifically it was known as a membership function.

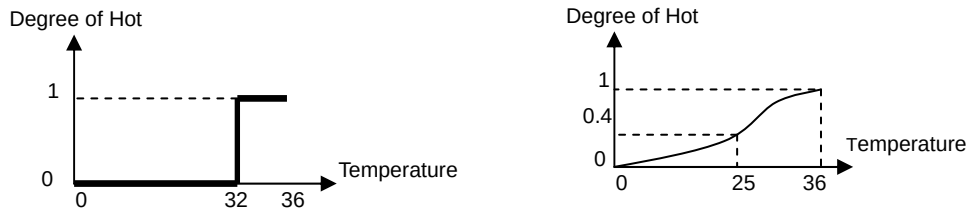


Figure 4-1 Exemplified Comparison between Classical Set and Fuzzy Set

A classical Set (right) and a fuzzy set (left) that describe the degree of hot.

In this study, FIS was built by using MATLAB Fuzzy Logic Toolbox [19]. The schematic diagram of one FIS was shown in Figure 4-2. Information flowed from left to right, 15 primitive features were treated as input variables and the single output represented the how much degree of the emotion. The parallel nature of the 5 if-then rules was the key component of the fuzzy inference system, because it controlled the behavior of the system.

The following five processes explained how the Fuzzy inference system worked. Figure 4-2 illustrated how a fuzzy inference system worked. It was assumed that the membership functions of each variable and the rules had been designed already.

1. **Fuzzyilfy inputs**. Initially, each input was a crisp value. However, within fuzzy logic, the membership degree of each fuzzy set of an input was used instead of one crisp value. This process gave membership degrees to each fuzzy set for each input by the calculation of membership functions.
2. **Calculating the antecedent of rules**. Based on the membership degrees, the antecedent was calculated by a fuzzy operator. In Figure 4-2, operator AND was implied.
3. **Calculating the consequent of rules**. According to the antecedent, this process calculated the formula designed in consequence of each rule.
4. **Aggregation**. It combined the consequence across all rules.
5. **Defuzzidication**. It calculated a weighted average for the combination which came from the fourth process.

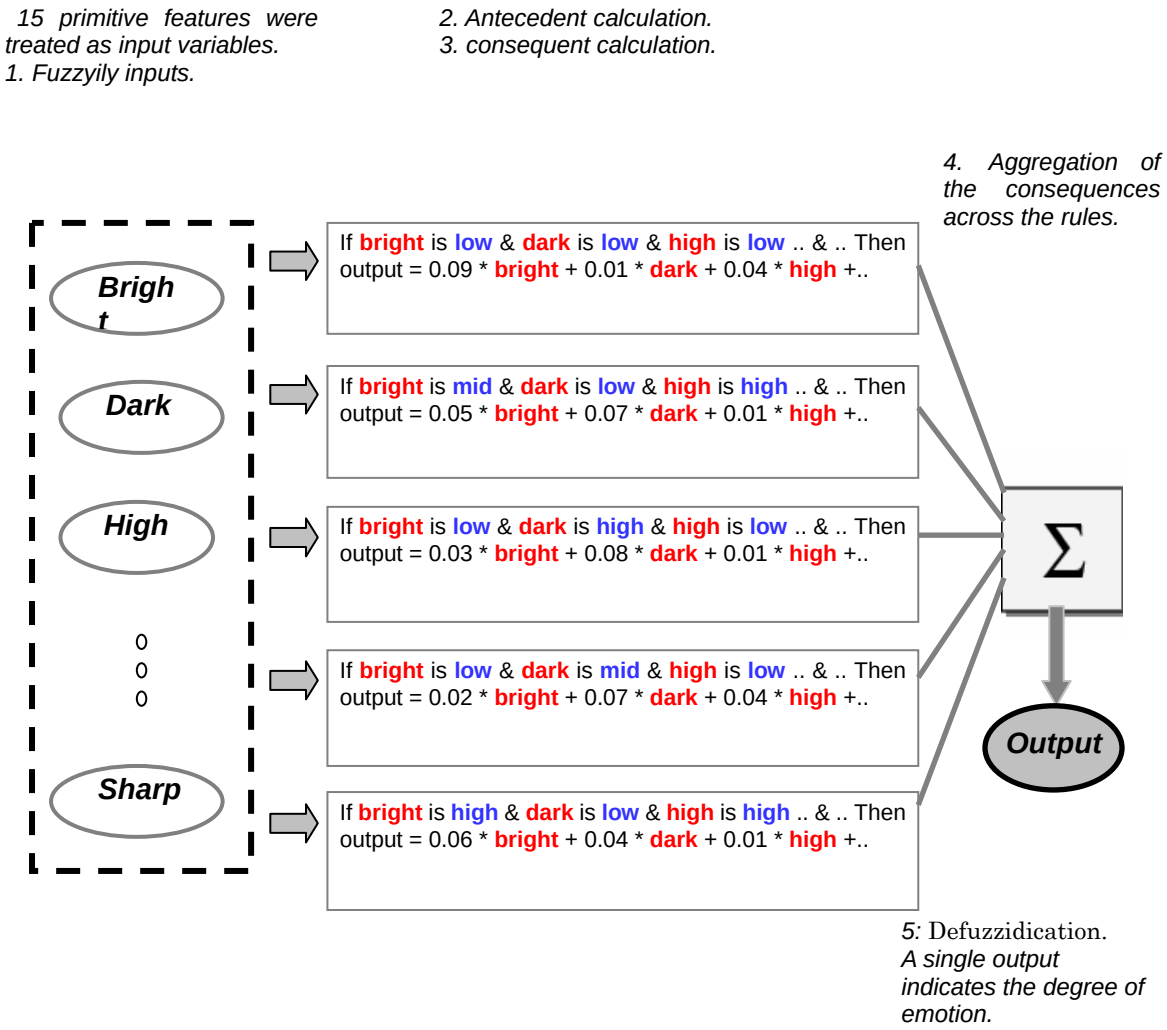


Figure 4-2 Schematic Diagram of FIS

For a FIS, membership functions of each input variable and rules of the whole system should be decided. In the next section, some underlying techniques by which membership functions and rules could be decided were introduced. How FIS was actually built for each of emotion category was described in Section 4.5.

4.2 Preamble

This session explains why FIS was used to build the relationship.

Fuzzy inference is the process of formulating the mapping from a given input to an output based on the concepts of fuzzy logic, that is, fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning [18]. The following list describes general observations about FIS. Each item is followed by given reasons for why FIS was used in this thesis.

1. FIS embedded existing structured human knowledge (experience, expertise, or heuristics) into workable mathematics. It was a tool for embedding structured human knowledge into an analytical model [20].

   This was one of the reasons for using FIS in this thesis. The idea corresponds to what the perceptual model proposed to deal with - human perception of emotions from speech. The problem that needs to be solved here is to establish the relationship between the emotion and the primitive feature in terms of human ability to perceive emotions from speech, where both emotions and primitive feature are, relevant to human knowledge.
2. Fuzzy logic was based on natural language. The basis for fuzzy logic was the basis for human communication. This observation underpinned many of the other statements about fuzzy logic [19]. Natural language is the main way humans use to communicate with each other; however, one individual character of it is vagueness. Fuzzy logic deals with such vague nature by various shapes of membership functions.

   This is one of the reasons for using FIS in this thesis. As mentioned before, primitive features are specified in terms of linguistic form as adjectives. The idea corresponds to the primitive feature of the perceptual model.
3. FIS are non-linear mappings. Fuzzy logic can model nonlinear functions of arbitrary complexity [21]

   This is one of the reasons for using FIS in this thesis. The relationship between the emotion and the primitive feature was nonlinear and complex.

4.3 Purpose

The purpose of this stage is to build an FIS for the emotions that modeled the relationship between one emotion and 15 primitive features.

4.4 Underlying Techniques of FIS

Fuzzy Logic Toolbox within MATLAB [19] was employed here to construct an FIS. Basically, Fuzzy Logic Toolbox provided two approaches to build FIS.

The first approach was intuitive. It was to assign membership functions directly for every input/output variable, to define rule structure, and to determine the method of implication and aggregation. Obviously, to decide these factors intuitively needed familiarity with the targeted system which was expected to be simulated. Whether to apply this approach or not depended on having enough experiences and experiments.

The second approach was to generate FIS automatically by using a ***clustering*** technique or ***adaptive neuro-fuzzy*** technique. A short explanation of both techniques is given in the following section. When there is the situation that a bunch of experimental input/output data had been collected, and it was still not easy to arbitrarily decide how many membership functions were appropriate, what membership functions, what rules structure should be, etc, by only looking at the experimental data, then this approach should be applied to build FIS for data modeling.

For this thesis, experimental data had been collected already. However, about emotion and primitive features, there is not enough experience and knowledge that could support building FIS intuitively by hand. Therefore, the second approach was considered more adequate for solving the problem and was chosen to build FIS.

4.5 Construction of FIS

By applying adaptive neuro-fuzzy technique within the Fuzzy Logic Toolbox of MATLAB, the initial FIS was first built and then improved using the available data. The available data were the perceptible degrees of emotions and primitive features of 50 sentences that had already been collected from the experiments described in Chapter 3. As was expected by FIS, input data would be the perceptible degrees of primitive features and output data would be the perceptible degrees of emotions. During the process of analysis, both training data and checking data were given. Training data contained desired input/output data pairs of the target system to be modeled and checking data helped with avoiding model overfitting during the training. Of the original 50 sentences, 40 sentences were used as training data and 10 sentences were used as checking data.

The following steps were used to analyze experimental data and to construct FIS. A schematic block diagram was depicted in Figure 4-3.

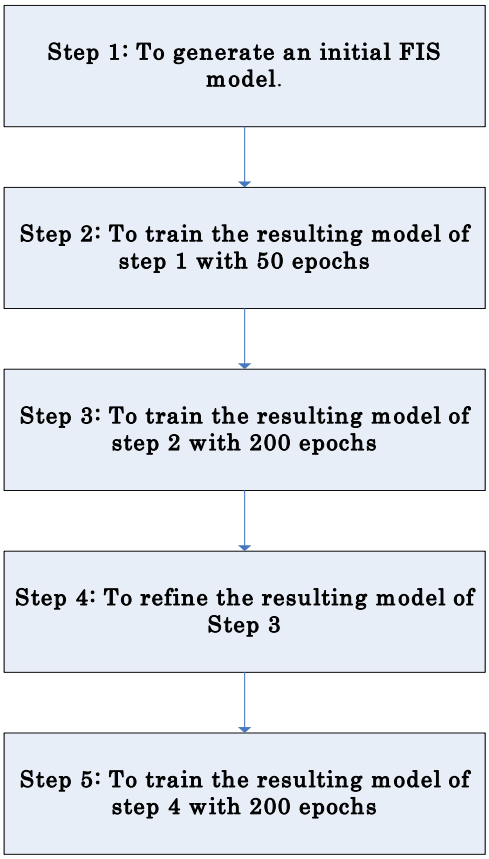


Figure 4-3 Building Steps of the Model

Step 1: To generate an initial FIS model: The first step was to generate an initial FIS model. The purpose of this step was to construct a raw model according to experimental data, which decided how many membership functions for each input variable and rules for the FISes so that it could be trained and adjusted in the following steps. Firstly, an estimation of the number of clusters from training data by using ***Subtractive Clustering*** technique was decided [22]. Then the number of membership functions and rule structure were decided. Subtractive clustering was a fast one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data. There was a trade-off between the number of clusters and the accuracy. Under an acceptable error, this algorithm was finally set up to produce 5 clusters by assigning a cluster radius of 1.5. Finally, an initial FIS model with 5 membership functions for each variable and with a 5-rule structure was generated. It was a first-order Sugeno type model [23]. A typical fuzzy rule in Sugeno type model has the form:

```
IF
      Input 1 = x and Input 2 = y,
 THEN
      Output is z = ax + by + c
```

In order to see how well the resultant model predicted the corresponding data set output values, the resultant model was verified by applying an input data set of training data and an input data set of checking data into the model respectively, and then the root mean squared error (RMSE) of corresponding output data set, named training error, and checking error were calculated.

Step 2: To train the resulting model of step 1 with 50 epochs: The second step was to train the resultant model of step 1 by Adaptive Neuro-Fuzzy technique to make a learning process. The adaptive neuro-fuzzy technique used here was a hybrid method consisting of backpropagation for the parameters associated with the input membership functions, and least squares estimation for the parameters associated with the output membership functions. The purpose of this step was to improve capability of the model by adjusting membership function parameters. A relatively short training (50 epochs) without implementing checking data was applied first here. This step generated a trained FIS model with 5 membership functions for each variable, with a 5-rule structure, and still it was a first-order Sugeno type model. Again, as in step 1, the resultant model was verified by applying the input data set of training data and input data set of checking data into this model respectively and the training error and checking error were calculated.

Step 3: To train of the resulting model of step 2 with 200 epochs: Step 3 was similar to step 2, to train the resulting model of step 2 by an Adaptive Neuro-Fuzzy technique to make a learning process, but it was a longer training process (200 epochs) and with implementing checking data . The purpose of this step was to try to improve the model further, and to detect whether the model was overfitting, which meant that the model was so fit for the training data that it no longer did a very good job of fitting the checking data. An overfitting model was “good” at only training data but not “good” at others, like checking data. The result of this step was a trained FIS model which was almost identical to the one we got from step 2, except with slight changes of parameters. Similarly, as in step 1, the resultant model was verified by applying input data set of training data and input data set of checking data, respectively, into this model and the training error and checking error were calculated.

Step 4: To refine the resulting model of Step 3: The fourth step was to refine the resultant model of step 3 by reducing the number of membership

functions from 5 to 3. The purpose of this step was to make the model to be more human-like. Through the above 3 steps, an optimal FIS with a 5-rule structure has been built. Within this model, there were 5 membership functions for each input variable, i.e. primitive features. However, taking those variables into account, it is not easy for humans to describe the intensity of these variables by 5 different levels. Hence, the number of membership functions of all variables was reduced from 5 to 3. The method used here was to delete two or three nearer and more similar curves and then add an average curve whose parameters were the average of the original two or three curves. Consequently, the result of this step was an FIS model with 5-rule structure and 3 membership functions named low, mid, and high, for each input/output variable.

Step 5: To train the resulting model of step 4 with 200 epochs : After the refined model was obtained from step 4, the fifth step was almost the same as in step 3, to apply a 200-epochs training process with checking data. The purpose of the step was to make the refined model fit the target system but not be overfitting of the training data. Eventually, a final FIS for one emotion to describe the relation of this emotion and 15 primitive features was completed. The final model was also verified by training error and checking error.

4.6 Results

5 FISes were built, one FIS for one emotion. The training errors and checking errors for each step during the training process are listed in Table 4-1 and Table 4-2 respectively. Step 4 did not have a training process.

Table 4-1 Training Error for Each Step

Step	Neutral	Joy	CA	Sadness	HA
1	5.06E-16	2.56E-16	1.02E-15	1.50E-15	1.29E-15
2	9.09E-07	1.33E-07	3.72E-06	3.93E-06	4.37E-06
3	9.09E-07	1.55E-07	3.72E-06	3.93E-06	4.37E-06
5	9.23E-07	1.67E-07	3.74E-06	3.94E-06	4.36E-06

Table 4-2 Checking Error for Each Step

Step	Neutral	Joy	CA	Sadness	HA
1	0.26529	0.13715	0.52488	0.69387	0.88355
2	0.089457	0.033491	0.22697	0.27705	0.29278
3	0.089457	0.03254	0.22697	0.27705	0.27517
5	0.09201	0.032471	0.22765	0.27677	0.27109

4.7 Examination

In order to see whether these FISes well-modeled the relationship between the emotion and the primitive feature, one examination was conducted.

4.7.1 Examination method

The relationship between the emotion and the primitive feature could be examined by exploring how each primitive feature affected the emotions. In order to see it, a method with 3 steps was developed.

Firstly, the input data was designed. A sample of the input data is shown in the right part of Table 4-3 (under the heading labeled “Input (15 primitive features)”). Each time only one of the 15 primitive features was considered as a variable and the other 14 primitive features were fixed at middle values of each maximal degree of perception of emotion that was collected in Experiment 3. The variable was changed from 0 to 3 with an interval oF0.1. The output from the FIS against one variable was treated as an ***emotional degree***, as shown in the last column (under the heading labeled “Output”) in Table 4-3.

Table 4-3 Sample of Input Data for Examination of FIS

Input (15 primitive features)												Output
Bright	Dark	High	Low	Strong	…	Heavy	Clear	Noisy	Quiet	Sharp		Normal
0	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.67448
0.1	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.66195
0.2	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.64816
0.3	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.63303
0.4	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.61648
0.5	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.59841
0.6	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.57874
0.7	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.55742
0.8	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.53438
0.9	1.417	1.167	1.292	1.458	…	1.229	1.146	1.438	1.104	1.229		0.50957
1	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		0.48295
2.3	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		-0.00096
2.4	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		-0.04387
2.5	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		-0.08666
2.6	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		-0.12916
2.7	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		-0.17121
2.8	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		-0.21266
2.9	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		-0.2534
3	1.417	1.167	1.292	1.458		1.229	1.146	1.438	1.104	1.229		-0.29334

Second, the output, i.e. emotion degree, was plotted against the input, i.e. primitive features. The relationship between the emotion and the primitive feature could be represented by the regression line of the output (see Figure 4-4, the output of FIS was the solid line and the regression line was the dotted line), the slope of the regression line indicated how significantly the primitive feature affected the emotion.

Finally, the slope of the regression line was computed. Accordingly, the effect of each primitive feature on one emotion was determined by the slope. For one emotion, if the absolute value of the slope of one primitive feature was higher, that primitive feature was considered to a greater effect on the emotion than others. And if it was lower, then it had smaller effect on the emotion. As shown in Figure 4-4, comparing the slope of the regression line in the left with that in the right sub plot, the primitive feature monotonous was affected by the emotion Neutral more than by the emotion Hot Anger.

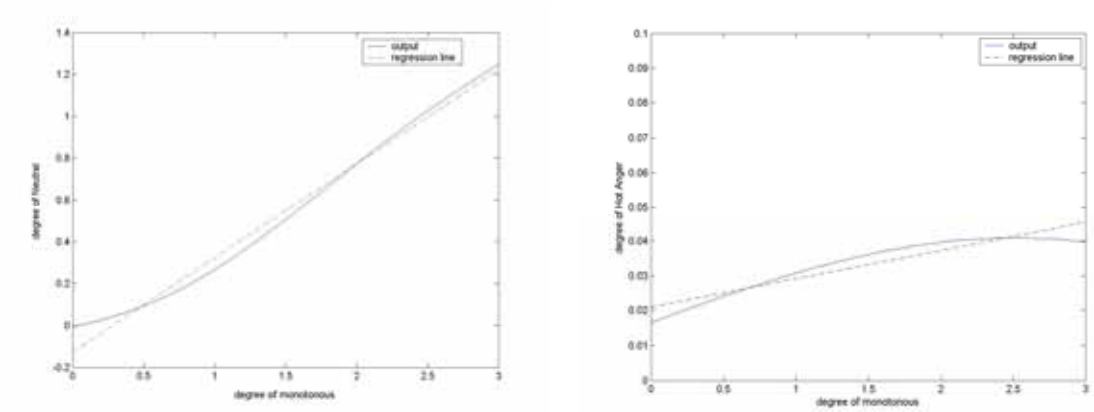


Figure 4-4 Two Plots of Monotonous against Neutral and Hot Anger

The plots indicate the relationship between Monotonous and Normal and the relationship between Monotonous and Hot Anger. The slope of regression lines indicates how significantly Monotonous was affected by the two emotions.

4.7.2 Results

Table 4-4 list the results of each FIS that addressed how each primitive feature was affected by each emotion according to the slope of regression lines. Each sub table presents the results for one emotion. By considering the absolute value of slope, the cells with shading represent the 5 primitive features that were most related to the emotion since they have higher absolute value of slope. The last two columns of each sub table were RMSE of training data (i.e. training error) and RMSE of checking data (i.e. checking error) that produced by the FISes.

By observing Table 4-4 regarding a positive correlate, Neutral was most characterized by monotonous and calm. Joy was most characterized by bright, high, and clear. Cold Anger was most characterized by heavy, low, dark and strong. Sadness was most characterized by weak, quiet, dark and low. Hot Anger was most characterized by noisy, unstable, sharp, strong and high. Regarding a negative correlate, Neutral was most characterized by bright, well-modulated and high. Joy was most characterized by heavy and low. Cold Anger was most characterized by quiet. Sadness was most characterized by strong.

Table 4-4 Raw Results of the Fuzzy Inference Systems

Normal	
PF	Slope
明るい bright	-0.336
抑揚のある well- modulated	-0.301
声の高い high	-0.286
落ち着きのない unstable	-0.169
うるさい noisy	-0.121
強い strong	-0.102
鋭い sharp	-0.101
弱い weak	-0.094
暗い dark	-0.008
重い heavy	0.058
声の低い low	0.107
静かな quiet	0.156
明らかな clear	0.184
落ち着いた calm	0.222
単調な monotonous	0.420
RMSE of Checking data	0.097
RMSE of Training data	0.022

Joy	
PF	Slope
重い heavy	-0.202
声の低い low	-0.130
暗い dark	-0.086
うるさい noisy	-0.074
鋭い sharp	-0.066
静かな quiet	-0.060
単調な monotonous	-0.012
弱い weak	-0.011
落ち着いた calm	-0.011
強い strong	0.004
落ち着きのない unstable	0.059
抑揚のある well- modulated	0.107
明らかな clear	0.187
声の高い high	0.191
明るい bright	0.254
RMSE of Checking data	0.034
RMSE of Training data	0.012

Cold Anger	
PF	Slope
静かな quiet	-0.326
単調な monotonous	-0.203
落ち着いた calm	-0.143
弱い weak	-0.120
明るい bright	-0.079
声の高い high	-0.041
明らかな clear	-0.005
鋭い sharp	0.067
落ち着きのない unstable	0.088
うるさい noisy	0.141
抑揚のある well- modulated	0.170
強い strong	0.265
暗い dark	0.308
声の低い low	0.405
重い heavy	0.408
RMSE of Checking data	0.051
RMSE of Training data	0.030

Sadness	
PF	Slope
強い strong	-0.068
声の高い high	-0.060
明るい bright	-0.047
落ち着きのない unstable	-0.037
明らかな clear	-0.033
うるさい noisy	-0.022
単調な monotonous	-0.017
鋭い sharp	-0.011
抑揚のある well- modulated	0.026
重い heavy	0.028
落ち着いた calm	0.067
声の低い low	0.083
暗い dark	0.136
静かな quiet	0.239
弱い weak	0.273
RMSE of Checking data	0.016
RMSE of Training data	0.031

Hot Anger	
PF	Slope
明るい bright	-0.020
静かな quiet	-0.015
明らかな clear	-0.014
落ち着いた calm	-0.014
弱い weak	-0.009
声の低い low	-0.002
暗い dark	-0.001
単調な monotonous	0.007
重い heavy	0.032
抑揚のある well- modulated	0.038
声の高い high	0.065
強い strong	0.077
鋭い sharp	0.077
落ち着きのない unstable	0.100
うるさい noisy	0.102
RMSE of Checking data	0.022
RMSE of Training data	0.038

4.8 Discussion

In this chapter, fuzzy inference system (FIS) was proposed to build the relationship between the emotion and the primitive feature. As Table 4-4 showing, five FISes well-modeled the relationship between the emotions and the primitive features,

corresponding to human ability to perceive emotional speech. Giving it an example, when we heard a voice and felt it is joy, we would not say the voice sounds heavy, but conversely, we probably would say the voice sounds bright or clear. Such observations match the result of FIS of Joy. Therefore, it was concluded that FIS well-modeled the relationship between the emotion and the primitive feature.

Chapter 5

Acoustic Features Analysis

Chapter 5 describes acoustic features analysis. It includes descriptions of what acoustic features are measured and how to measure and analyze these acoustic features in order to see what acoustic features affect primitive features most. Acoustic features are concerned in terms of three attributes of sound, pitch, loudness, and timbre. Mainly, this chapter focuses on pitch and loudness, since they are considered more related to primitive features. Regarding timbre, although the analysis of timbre is not completely done yet since the time limit, this chapter also discusses two values measured from power spectrum, because they provide useful clues to find those acoustic features in terms of timbre that affect primitive features.

5.1 Relationship between Acoustic Features and Primitive Features

As described in Chapter 1, primitive features were defined as those adjectives used to describe emotional speech, i.e. primitive features described some of the attributes of sound. Since pitch, loudness, and timbre are the main three attributes of sound, and furthermore since pitch and loudness are especially relevant to the perception of emotional speech, acoustic features were analyzed in terms of these two categories first. Pitch effect was mainly caused by the changes of F0. Loudness effect was mainly caused by the changes of power. Therefore, F0 contour and power were essential elements in the analysis stage. They were calculated by STRAIGHT [24] with FFT length 1024 point, frame rate 1ms, and shift frame 1ms.

5.1.1 F0

In this paper, the terms F0 and pitch are used interchangeably, although strictly, speaking “pitch” is the perception by the human ear of the fundamental frequency (F0) of vibration of the vocal folds. The F0 contour changes during an utterance indicate a variety of linguistic meanings, including denoting accented words or accentual phrases. Also, the perception by listeners of a difference in the emotional speech is strongly related to the variation of F0 in the accentual phrases and the overall intonation of the utterance. F0 was calculated by applying STRAIGHT, with FFT length 1024 point and

frame rate 1ms. The following 3 steps were executed to analyze all utterances that were used in this study.

Step 1: It was to find the accentual phrases for each utterance. When a Japanese speaker produces an utterance, he/she will produce it in separate phrases, often indicated by a pause as well as a drop in F0. A Japanese utterance can be divided into several accentual phrases. For example, the sentence, /a ta ra shi i ku ru ma o ka i ma shi ta/, was always spoken with pauses in such a way, / a ta ra shi i ku ru ma o ka i ma shi ta/. There are 3 accentual phases here.Accentual phrase is considered as, usually, when one wishes to say a sentence, he/she always presents it with a separation of several phases, which was called accentual phase. Utterance of Japanese could be divided into several accentual phrases. For example, a sentence, /a ta ra shi i ku ru ma o ka i ma shi ta/, was always spoken with pauses in such a way, / a ta ra shi i ku ru ma o ka i ma shi ta/. There were 3 accentual phases here. F0 contours and accentual phases of this utterance are shown in Figure 5-1.

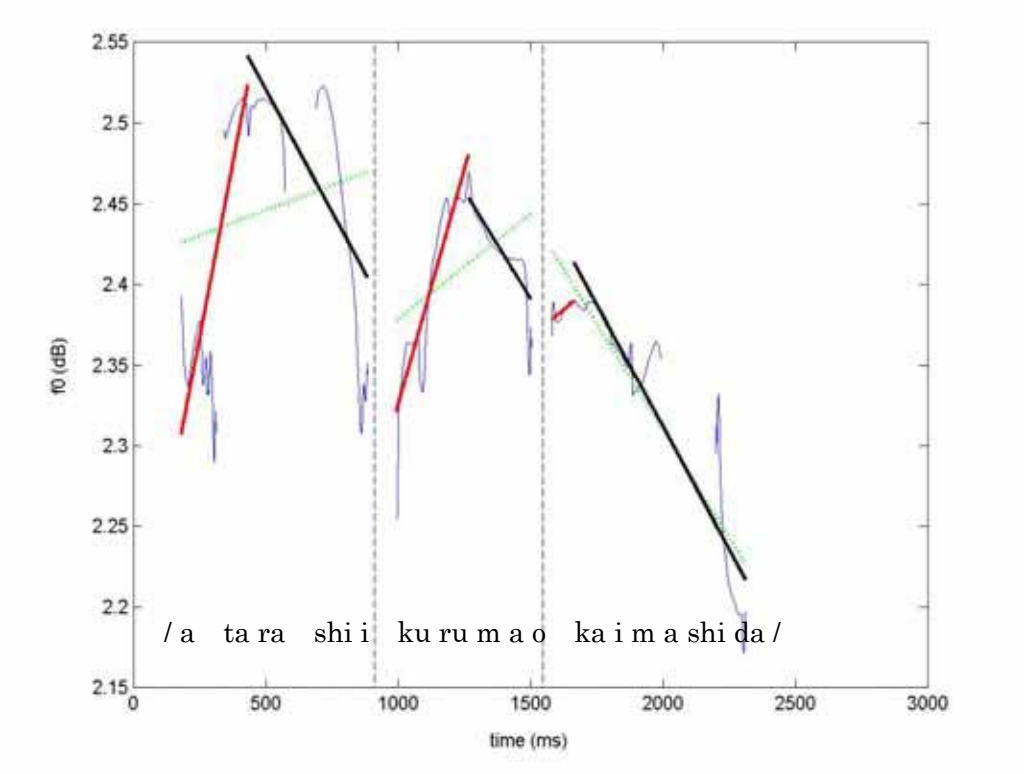


Figure 5-1 F0 Contour and Accentual Phrases of the Utterance /a ta ra shi i ku ru ma o ka i ma shi da/. Its intended emotion was Neutral.

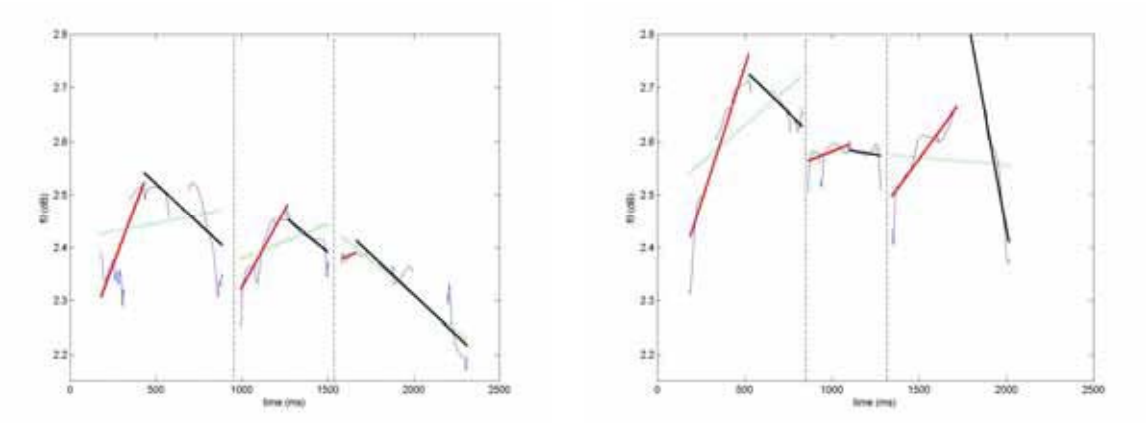


Figure 5-2 F0 Contour and Accentual Phrases of the Utterance /a ta ra shi i ku ru ma o ka i ma shi da/ spoken in different intended emotions. The left plot represents Neutral utterance and the right plot represents Joy utterance.

Step 2: It was to measure the acoustic features. Figure 5-2 presented the F0 contour and accentual phrases of the sentence / a ta ra shi I ku ru ma o

ka i ma shi ta/ spoken in Neutral and Joy. By comparing those two sub plots in Figure 5-2, the variation of F0 in both the accentual phrase and the overall utterance was evident. Due to such observations from Figure 5-2, the variation of F0 in accentual phrase and in the overall utterance, the following 9 acoustic features were measured. 5 of the 9 acoustic features were measured from each accentual phrase, and the other 4 acoustic features were measured from each utterance.

For each accentual phrase, the regression lines during the rising areas and the falling areas were calculated. Regression lines of the utterance /a ta ra shi i ku ru ma o ka i ma shi ta/ are shown in Figure 5-1. According to the regression lines, the four acoustic features, rising slope (RS), rising duration (RD), falling slope (FS), and falling duration (FD) for each accentual phrase were measured. Furthermore, pitch range in each accentual phrase (PRAP) was measured, where *pitch range* is the bandwidth of the range bounded by the lowest and highest F0 of each phrase.

For each utterance, average pitch (AP), pitch range (PR), and highest pitch (HP) were measured. A*verage pitch* is the average F0 value of the contour, *pitch range* is the bandwidth of the range bounded by the lowest and highest F0 of the utterance and highest pitch is the highest F0 value. Finally rising slope of the first accentual phrase (RS1st) was measured.

Step 3: It was to explore what acoustic features affected the primitive features by calculating the correlation coefficients of them. The result is shown in Table 5-1 (on page 48).

5.1.2 Power

Loudness is a perceptual or subjective quality of a sound. Changes in power can cause changes of loudness. In this thesis, power contours were calculated by STRAIGHT. Concerning the variation of power in accentual phrase, the analysis method of the power contour was similar to the method used for pitch analysis. There were 3 steps for analyzing the power contour.

Step 1: To find the accentual phrase.

Step 2: To measure the acoustic features. By using the same method, for each accentual phrase, rising slope (RS), rising duration (RD), falling slope (FS), falling duration (FD), and mean value of power range in accentual phrase (PRAP) was measured. For each utterance, power range (PWR), rising slope of the first accentual phrase power (RS1st), and the ratio between the average power in high frequency portion (over then 3 kHz) and the average power (RHT) was measured.

Step 3: To see what value affected primitive features by calculating the

correlation coefficients. The result is also shown in Table 5-2 (on page 49)

5.1.3 Discussion

Correlation coefficients are shown in Table 5-1 and Table 5-2. Dark cells indicate that the correlation coefficients were over 0.6. For pitch (see Table 5-1), those acoustic features that were most related to primitive features were highest pitch (HP), average pitch (AP), mean value of rising slope (RS), and rising slope of the first accentual phrase (RS1st). For power (see Table 5-2), those acoustic features that were most related to primitive features were mean value of power range in accentual phrase (PRAP), power range (PWR), and rising slope of the first accentual phrase (RS1st). The ratio between the mean value of power in high frequency and the mean value of power through the whole frequency domain (RHT) affects 4 primitive features, however their correlation coefficient are relatively low.

From a different point of view, we discuss how a pair of opposite primitive features is affected by the same acoustic features. The results showed that most of the opposite primitive features have opposite correlation coefficients. Detailed observations are shown as follows:

For F0 (see Table 5-1)

- *Bright* and *dark* are affected by HP and AP
- *High* and *low* are affected by RS, HP, AP and RS1st.
- *Calm* and *unstable* are affected by RS, HP and RS1st.
- *Noisy* and *quiet* are affected by RS and HP.
- *Strong* and *weak* are not affected by any acoustic features of pitch.
- Although *Heavy* and *clear* are not an opposite pair, they are affected by HP, AP and RS1st in an opposite way.

For power (see Table 5-2)

- *Bright* and *dark* is not affected by same acoustic features. *Dark* is affected by PRAP, PWR, and RS1st. However, there is no notable acoustic feature that affects *bright*.
- *High* and *low is* affected by PWR and RS1st.
- Three opposite pairs are affected by PRAP and PWR. They are *strong* and *weak*, *calm* and *unstable*, and *noisy* and *quiet.*
- Two opposite pairs are not affected by acoustic features. They are well modulated and monotonous, heavy and clear.

By observing each primitive feature individually, the following was concluded. The easier the primitive feature is perceived, the higher is the correlation coefficient. Also, the easier the primitive feature is perceived, the more acoustic features can be found. For example, *high*, *low*, *dark* and *quiet* are easy to perceive in terms of primitive

features, and they have higher correlation coefficients and are affected by more acoustic features. Conversely, *well modulated* and *monotonous* are not easy to be perceived, hence they have lower correlation coefficients and are affected by fewer acoustic features.

Table 5-1 Correlation Coefficients between Primitive Features and Acoustic Features Measured from Pitch

The acoustic features were mean value of rising slope (RS), mean value of rising duration (RD), mean value of falling slope (FS), mean value of falling duration (FD), mean value of pitch range in accentual phrase (PRAP), average pitch (AP), pitch range (PR), highest pitch (HP), and rising slope of the first accentual phrase (RS1st). The correlation coefficient over 0.6 is marked in gray which indicated which acoustic features had higher effect on which primitive features.

PF	bright	dark	high	low	strong	weak	calm	unstable	well mo-dulated	mono-tonous	heavy	clear	noisy	quiet	sharp
RS	0.45	-0.64	0.70	-0.60	0.56	-0.54	-0.72	0.65	0.54	-0.32	-0.39	0.44	0.63	-0.68	0.57
RD	0.01	0.25	0.01	0.23	-0.08	0.28	0.12	0.00	0.09	-0.11	0.20	-0.17	-0.03	0.25	-0.10
FS	0.06	-0.09	0.33	-0.05	0.40	-0.13	-0.35	0.46	0.39	-0.28	0.07	0.00	0.43	-0.20	0.41
FD	-0.21	0.08	-0.38	0.07	-0.41	0.11	0.45	-0.45	-0.50	0.50	-0.02	-0.01	-0.37	0.26	-0.39
PRAP	0.06	-0.05	0.35	0.00	0.47	-0.13	-0.43	0.52	0.50	-0.39	0.15	0.00	0.49	-0.25	0.49
AP	0.72	-0.89	0.87	-0.91	0.33	-0.54	-0.63	0.58	0.41	-0.11	-0.78	0.76	0.52	-0.72	0.32
PR	0.02	0.00	0.26	0.06	0.40	-0.11	-0.37	0.44	0.41	-0.35	0.16	-0.02	0.38	-0.21	0.42
HP	0.69	-0.88	0.90	-0.89	0.42	-0.56	-0.70	0.66	0.50	-0.18	-0.73	0.74	0.60	-0.74	0.42
RS1st	0.50	-0.79	0.77	-0.78	0.45	-0.58	-0.66	0.60	0.42	-0.10	-0.61	0.66	0.57	-0.72	0.46

Table 5-2 Correlation Coefficients between Primitive Features and Acoustic Features Measured from Power

The acoustic features were mean value of rising slope (RS), mean value of rising duration (RD), mean value of falling slope (FS), mean value of falling duration (FD), mean value of power range in accentual phrase (PRAP), power range (PWR), rising slope of the first accentual phrase (RS1st), and the ratio between the average power in high frequency portion (over then 3 kHz) and the average power (RHT). The correlation coefficient over 0.6 is marked in gray which indicated which acoustic features had higher effect on which primitive features.

PF	bright	dark	high	low	strong	weak	calm	unstable	well mo dulated	mono- tonous	heavy	clear	noisy	quiet	sharp
RS	0.02	-0.11	0.06	-0.08	0.11	-0.14	-0.15	0.22	0.10	0.00	-0.07	0.07	0.13	-0.14	0.10
RD	0.22	-0.35	0.30	-0.46	-0.12	-0.01	-0.05	0.07	-0.09	0.19	-0.43	0.26	0.09	-0.06	-0.11
FS	-0.10	0.14	-0.25	0.12	-0.21	0.09	0.20	-0.26	-0.16	0.12	0.02	-0.06	-0.22	0.15	-0.22
FD	-0.22	0.09	-0.38	0.07	-0.41	0.12	0.45	-0.45	-0.49	0.49	-0.02	-0.02	-0.37	0.26	-0.39
PRAP	0.31	-0.67	0.62	-0.56	0.73	-0.68	-0.76	0.72	0.55	-0.26	-0.30	0.47	0.76	-0.79	0.72
PWR	0.43	-0.74	0.75	-0.65	0.69	-0.67	-0.78	0.76	0.59	-0.27	-0.41	0.57	0.76	-0.81	0.67
RS1st	0.48	-0.80	0.64	-0.70	0.45	-0.78	-0.60	0.51	0.27	-0.01	-0.56	0.64	0.44	-0.79	0.41
RHT	-0.10	-0.05	0.29	0.00	0.67	-0.14	-0.54	0.66	0.52	-0.41	0.24	-0.10	0.72	-0.30	0.67

5.2 Power Spectrum

In Section 5.1, several acoustic features in terms of pitch and loudness were investigated and discussed. However, only these acoustic features were not enough to build the relationship between the primitive feature and the acoustic feature for the perceptual model. More were needed. The analysis of acoustic features in terms of timbre has been doing. Although the analysis is not completely done yet since the time limit, part of the analytic results provided very useful clues to find the acoustic features in terms of timbre that affect the primitive features. This section discusses two values measured from power spectrum. They are average power spectrum and variance in vowel power spectrum.

5.2.1 Average power spectrum

Power spectra were calculated by STRAIGHT, with FFT length 1024 point and frame rate 1ms. Average power spectrum was the average of power spectra crossing time domain. The average power spectra of 10 utterances that had the same emotion were collected. Their results for the emotions, Neutral, Joy, Cold Anger, Sadness and Hot Anger, are shown in Figure 5-3, Figure 5-4, Figure 5-5, Figure 5-6, and Figure 5-7, respectively. Two aspects, the aspect of the frequency domain and the aspect of the time domain, were considered.

From the aspect of the frequency domain, by comparing the average power spectrum in the high frequency portion and in the low frequency portion, it was noticed that for Hot Anger, average power spectra in the high frequency portion were not as low as the other emotions. This observation provided a useful clue to find the acoustic features that affected those primitive features related to Hot Anger. It is necessary to do more detailed analysis of power spectrum in the frequency domain.

Regarding the aspect of the time domain, by comparing Figure 5-3 and the other figures, the average power spectra of the 10 Neutral utterances had a relatively similar contour compared to other emotions. This observation provided a motivation for exploring the variance in vowel power spectra through the time domain. The detailed results are described in the next section.

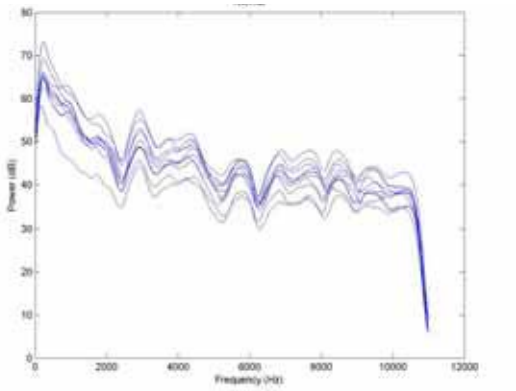


Figure 5-3 Average Power Spectrum of 10 Neutral Utterances

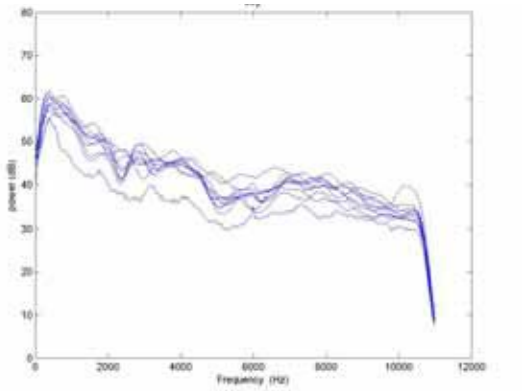


Figure 5-4 Average Power Spectrum of 10 Joy Utterances

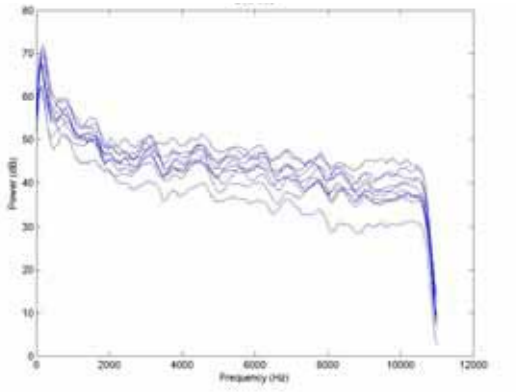


Figure 5-5 Average Power Spectrum of 10 Sadness Utterances

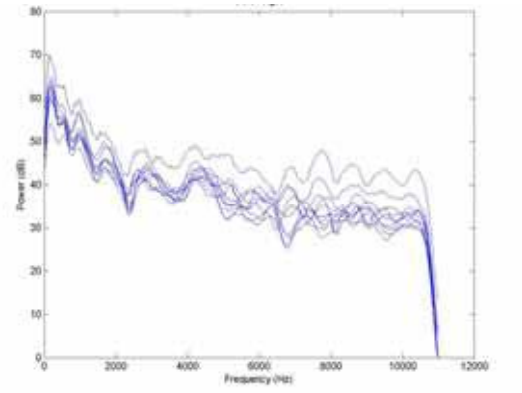


Figure 5-6 Average Power Spectrum of 10 Cold Anger utterances

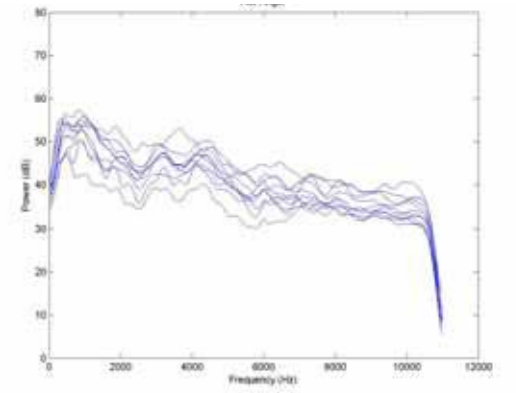


Figure 5-7 Average Power Spectrum of 10 Hot Anger utterances

5.2.2 Variance in vowel power spectrum

Power spectra of vowels were averagely extracted through the time domain, including vowels at beginning, middle, and ending of each utterance. The variance in vowel power spectrum is shown in Figure 5-8 and the data are listed in Table 5-3. The results showed that for every sentence, Neutral has the smallest variance in vowel power spectrum through the time domain. And it gave an explanation why Neutral utterances had relatively similar contour compared to the other emotions discussed in the previous section.

Moreover, by calculating the regression lines of vowel power spectrum against the time domain for each utterance, regression lines of 10 utterances with the same emotion were plotted in Figure 5-9, Figure 5-10, Figure 5-11, Figure 5-12, and Figure 5-13, for Neutral, Joy, Cold Anger, Sadness and Hot Anger respectively. In this way, the regression line indicated the trend of the vowel power spectrum in one utterance. Regarding Neutral, in Figure 5-9, only one utterance had a trend upward, but others' had a trend downward. Regarding Joy, in Figure 5-10, it seemed the trend depended on the particular sentence. Regarding Cold Anger and Sadness, all utterances had trends downward. Finally, regarding Hot Anger, most utterances had an upward trend except one.

In order to compare the variance of the upward or downward trend in different emotions, an average line for 10 regression lines of one emotion was made, and the average lines for 5 emotions are depicted Figure 5-14. When the time period was 2500ms, the most noticeable feature was the trend of Cold Anger of a downward slope to about 11 dB, and Sadness showed a downward slope to about 6 dB. Neutral showed a downward slope to about 3 dB. On the other hand, Hot Anger was the only one that showed an upward slope to about 3 dB. It seems that the slope for Joy decreased to about 2 dB. However, this should be ignored since we see in Figure 5-10 that the slope of Joy both increases and decrease. The downward slope to about 2 dB does not actually represent what Joy utterances.

5.3 Concluding remarks

In this chapter, in order to find the relationship between the primitive feature layer and the acoustic feature layer for the perceptual model, the acoustic features were analyzed in terms of the three attributes of sound, pitch, loudness and timbre. The relationships between the primitive features and the acoustic features in terms of pitch and loudness have been constructed. But, the relationship between the primitive features and the acoustic features in terms of timbre is not completely constructed yet. The following are concluded.

For pitch, a number of acoustic features were measured from F0 contour. According to their correlation coefficients, the acoustic features that are most related to primitive feature are highest pitch (HP), average pitch (AP), mean value of rising slope (RS), and rising slope of the first accentual phrase (RS1st).

For loudness, a number of acoustic features were measured from power contour. According to their correlation coefficients, the acoustic features that are most related to primitive feature are mean value of pitch range in accentual phrase (PRAP), power range (PWR), rising slope of the first accentual phrase (RS1st), and the ratio between the mean value of power in high frequency and the mean value of power through the whole frequency domain (RHT).

Regarding timbre, the uncompleted results suggest that more acoustic features that affect the primitive feature can be found, by further analysis of power spectrum. For example, considering frequency domain, it could be the ratio of the average power spectrum in low frequency and that in high frequency, and considering time domain, it could be the variance in vowel power spectrum.

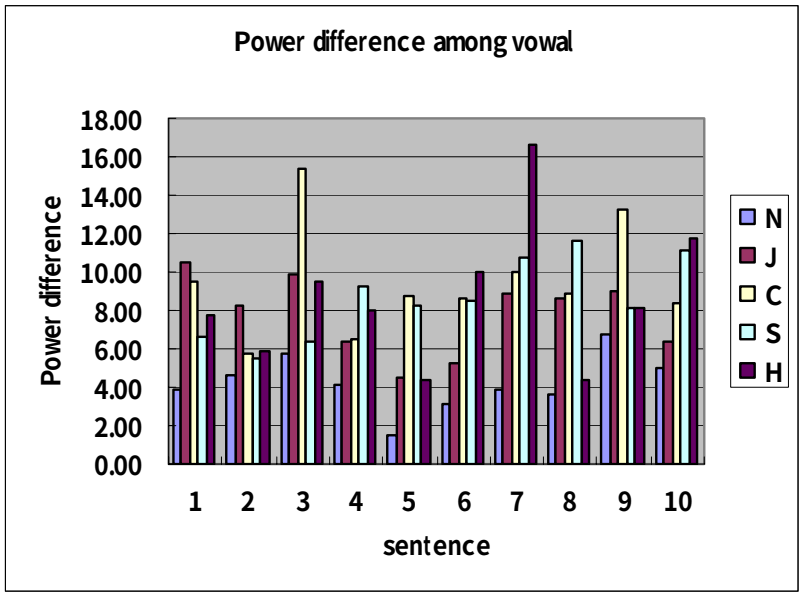


Figure 5-8 Variance in Vowel Power Spectrum in Time Domain

Table 5-3 Detailed Data of Variance in Vowel Power Spectrum in Time Domain

Emotions	1	2	3	4	5	6	7	8	9	10
N	3.92	4.66	5.72	4.13	1.46	3.10	3.92	3.57	6.72	4.99
J	10.46	8.26	9.84	6.36	4.52	5.21	8.84	8.65	8.96	6.35
C	9.44	5.70	15.33	6.54	8.76	8.59	10.05	8.91	13.23	8.39
S	6.58	5.48	6.39	9.21	8.20	8.51	10.75	11.65	8.12	11.11
H	7.73	5.88	9.52	7.95	4.42	9.95	16.66	4.36	8.09	11.73

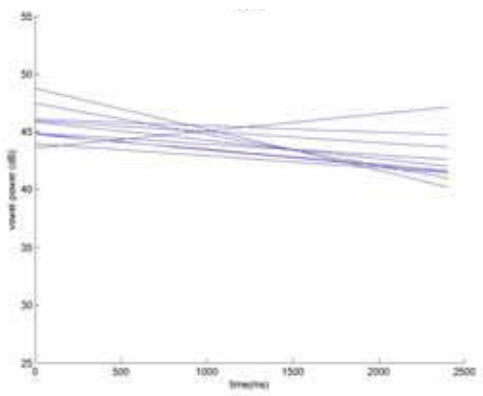


Figure 5-9 Vowel Power Spectrum in Time Domain of 10 Neutral Utterances

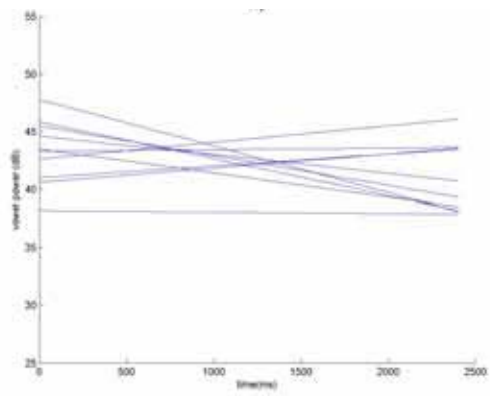


Figure 5-10 Vowel Power Spectrum in Time Domain of 10 Joy Utterances

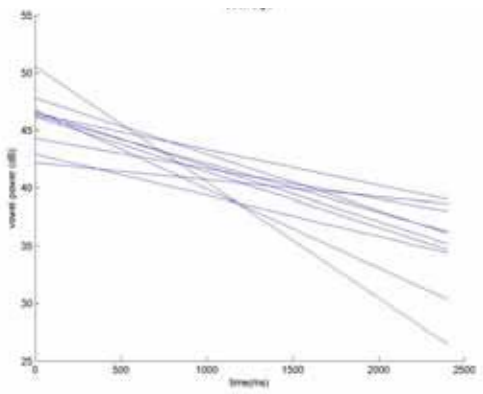


Figure 5-11 Vowel Power Spectrum in Time Domain of 10 Anger Utterances

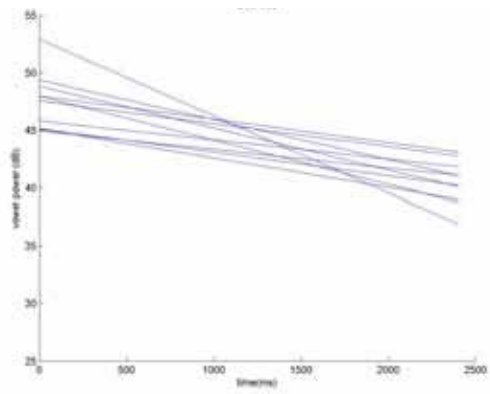


Figure 5-12 Vowel Power Spectrum in Time Domain of 10 Sadness Utterances

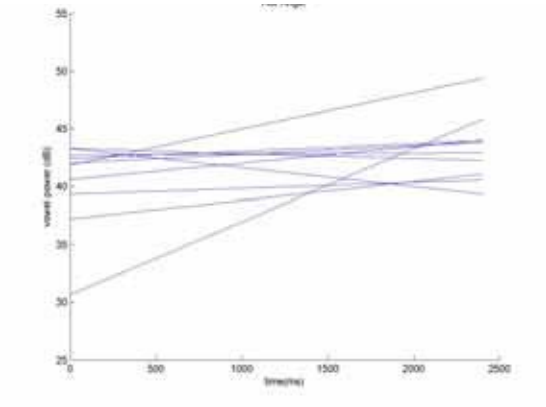


Figure 5-13 Vowel Power Spectrum in Time Domain of 10 Hot Anger Utterances

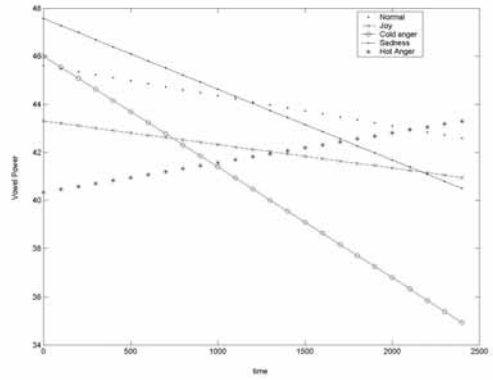


Figure 5-14 Trend of Vowel Power Spectrum of 5 Emotions

Chapter 6

Conclusion

This chapter summarizes what objectives have been achieved in this thesis and then concludes with a summary of what this study has contributions to this field of research. Finally, it discusses future work of the research.

6.1 Summary

Concerning the perception of emotional speech, a three-layered perceptual model is proposed in this thesis. The three layers are the emotion, the primitive feature, and the acoustic feature. To complete the perceptual model, it should be done within two stages. The first stage is to construct the model by the top-down approach. The second stage is to verify the constructed model by the bottom-up approach. For this thesis, the focus is on the first stage. In order to construct the perceptual model by the top-down approach, this thesis established two relationships of the model. The first is the relationship between the emotion and the primitive feature, and the second is the relationship between the primitive feature and the acoustic layer.

To establish the relationship between the emotion and the primitive feature, three perceptual experiments were conducted. The purpose of the first experiment was to examine all of the utterances in terms of 5 emotions in the voice database. The purposes of the other two experiments were selection of the primitive features. One was to build the psychological distance model and the other was to rate 34 adjectives in terms of how appropriate each adjective was to describe the 15 utterances, where 34 adjectives are chosen from 60 adjectives by the pre-experiment.

Summarizing the information gained from these three experiments, there are three main results:

1. The perceptual ratings of each utterance against each of 5 emotions are obtained form the first experiment
2. The psychological distance model is build by the second experiment
3. Perceptual ratings of 34 adjectives are given by the third experiment. 15 adjective are selected as the members of the primitive features for the perceptual model according to the perceptual ratings and the direction that

each adjective which are superimposed onto the psychological distance model.

4. The relationship, which is represented as one fuzzy inference systems for each of the 5 emotions, between the emotion and the primitive feature is established according to the perceptual ratings of the emotions and the primitive features by applying fuzzy logic.

To find the relationship between the primitive feature and the acoustic feature, some acoustic features of pitch and power are measured from the speech signals since pitch and loudness are considered the most significant attributes of sound. By calculating correlation coefficients of measured acoustic features and 15 primitive features, it is found that:

1. With regard to pitch, 4 acoustic features, highest pitch (HP), average pitch (AP), mean value of rising slope (RS), and rising slope of the first accentual phrase (RS1st) affect primitive features deeply.
2. With regard to the power characteristics, there are three acoustic features, mean value of power range in the accentual phrase (PRAP), power range (PWR), and rising slope of the first accentual phrase (RS1st) which affect primitive features most.
3. The easier a primitive feature is perceived the higher is the correlation coefficient. Also, the easier a primitive feature is perceived, the more acoustic features can be found.

Combing the two relationships that have been investigated in this study, the following five figures (see Figure 6-1, Figure 6-2, Figure 6-3, Figure 6-4, and Figure 6-5) show conceptual models for each emotion respectively. The black arrow indicates the relation is positive correlate. The red arrow indicates the relation is negative correlate. The line is thicker, the correlation is higher.

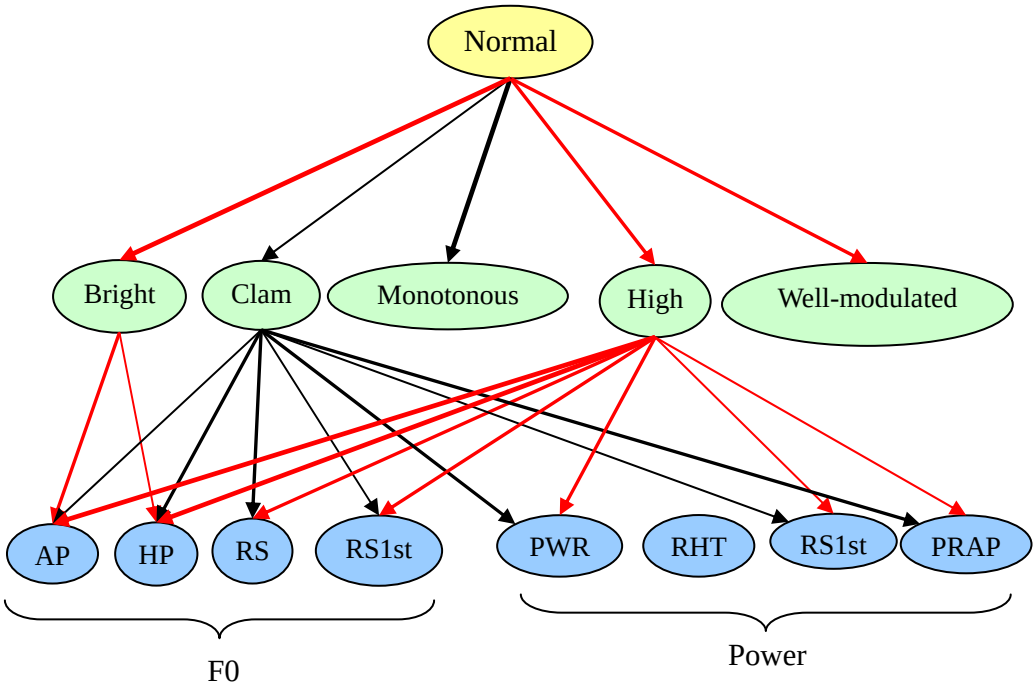


Figure 6-1 Conceptual Model of Emotion Neutral

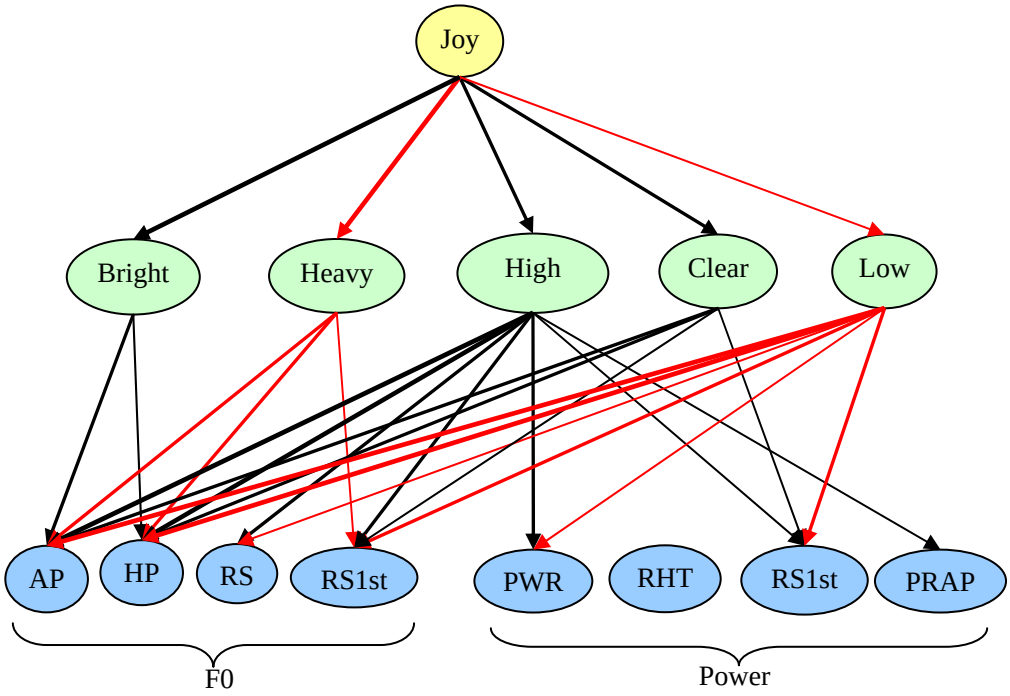


Figure 6-2 Conceptual Model of Emotion Joy

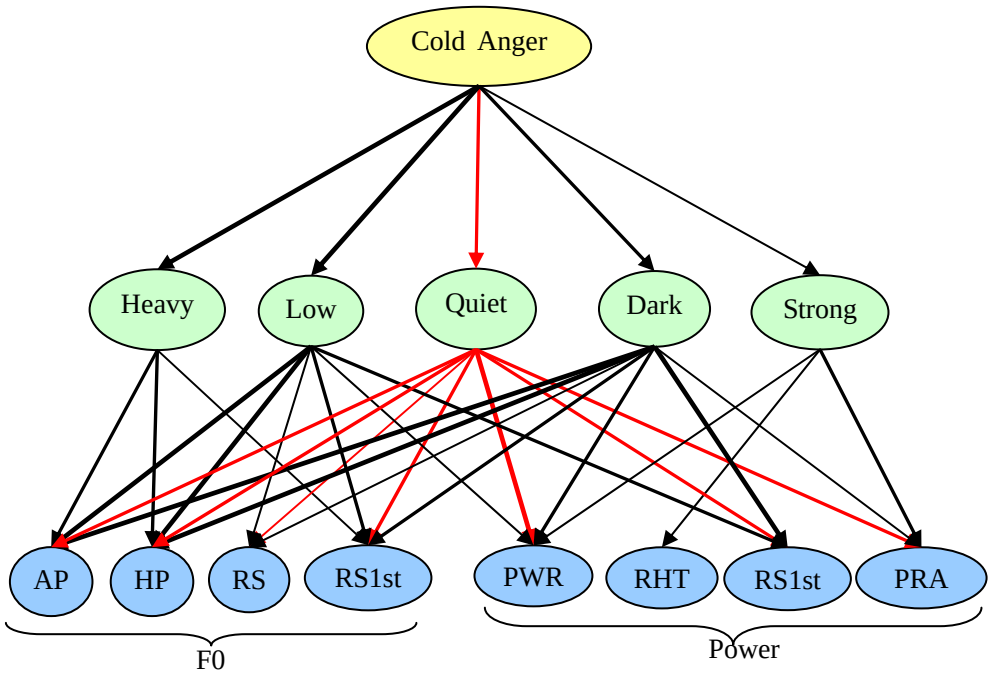


Figure 6-3 Conceptual Model of Emotion Cold Anger

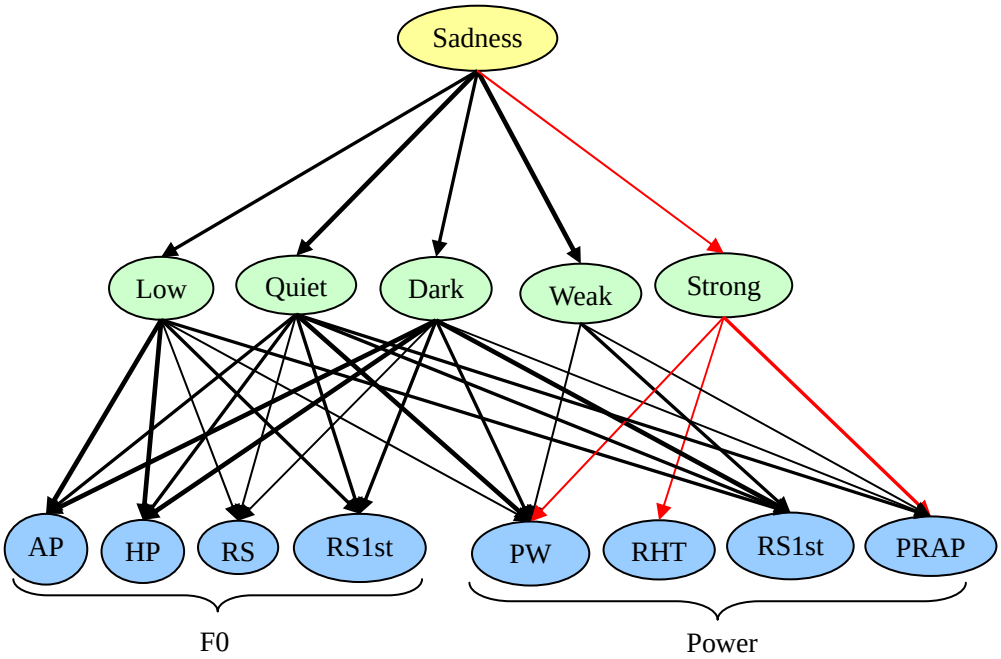


Figure 6-4 Conceptual Model of Emotion Sadness

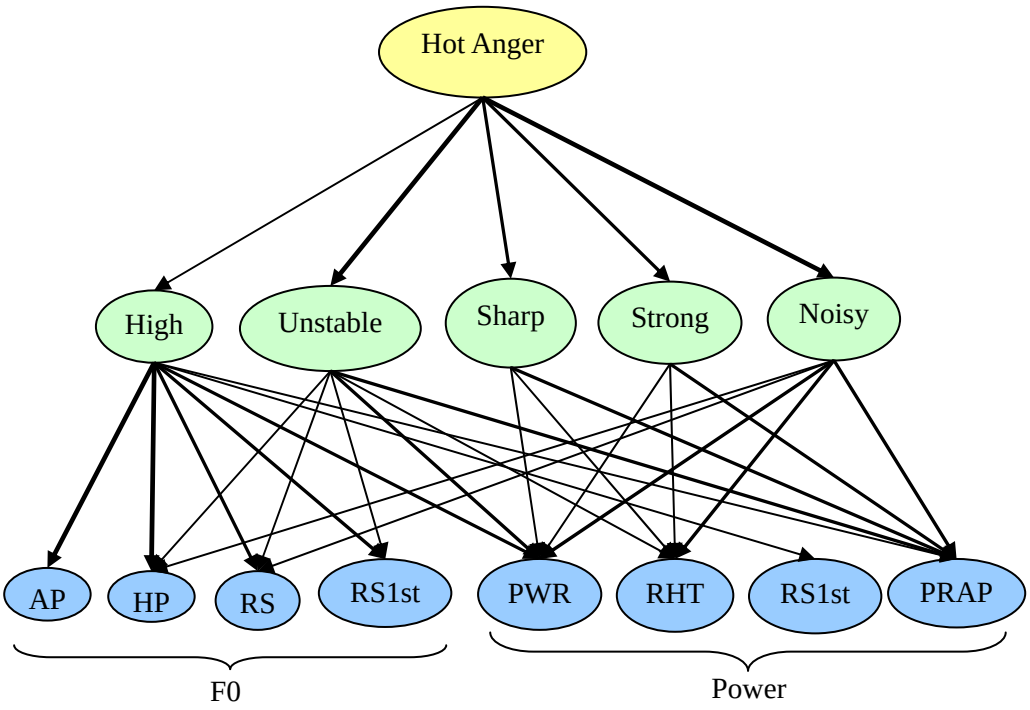


Figure 6-5 Conceptual Model of Emotion Hot Anger

6.2 Contributions

The purpose of this thesis, which is to construct a framework for the perceptual model by the top-down approach, is achieved. The proposed perceptual model explains the perception of emotional speech from a new perspective. That is, emotions are not directly relative to acoustic features. A new layer, the primitive feature, is inserted between the emotion and the acoustic feature.

The experiment results showed significance of the relationship between the emotion and the primitive feature. This thesis also reports that fuzzy logic appropriately deals with the relationship between the emotion and the primitive features in the perpetual model. It indicates that fuzzy logic can also be used in audible perception. The analyzed results also showed significance of the relationship between the primitive feature and the acoustic feature. They showed some acoustic features that have positive or negative correlated with primitive features. This thesis fills the gap of the previous work.

The work of this thesis is important since it creates the framework from which

the proposed perceptual model can be developed for future researches. The model not only provides an engineering point of view, but also covers the very nature of human of vagueness by a “precise” mathematical model. These two significant characteristics of the perceptual model open many possibilities for future extension and multiple fields of application of emotional speech processing, such as text-to-speech processing, emotional morphing speech processing, or creation of “humane interface” of human-machine interactions.

6.3 Future Work

In order to complete the construction of the proposed perceptual model, there are two major tasks that should be done. One is to conduct more acoustic features analyses. In this thesis, the acoustic features in terms of two perception units, pitch and loudness, have been analyzed; however, more acoustic features associated with other perception, for example duration and voice quality need to be examined. In addition, the relationship between the primitive feature and the acoustic feature should then be built by fuzzy logic according to the acoustic analysis results.

After the framework of the perceptual model has been built by the top-down approach, the other necessary work is to enforce the model by verifying the framework with a bottom- up approach. That is, according to what acoustic features have been found as well as those that will be found in the framework, it is important to synthesize emotional speech. In addition, it is necessary to conduct perceptual experiments to verify whether synthesized speech has perceptual ratings that corresponding to the propose model simulates in terms of variance in primitive features and variance in emotions.

Appendix-A.

Supplement Data of Experiment Results

A-1. Rating Result of Experiment 1

The table listed the percentage of ratings of each utterance of 171 utterances against each of 5 emotions.

UID	Neutral	Joy	Cold Anger	Sad	Hot Anger
a001	1	0	0	0	0
a002	0.945455	0	0.05	0.004545	0
a003	0.931818	0	0.068182	0	0
a004	0.995455	0	0.004545	0	0
a005	0.972727	0	0.027273	0	0
a006	1	0	0	0	0
a007	0.995455	0	0.004545	0	0
a008	0.95	0.022727	0.027273	0	0
a009	0.972727	0	0.027273	0	0
a010	0.954545	0.022727	0.022727	0	0
a011	0.977273	0	0.022727	0	0
a012	0.972727	0	0.004545	0.022727	0
a013	0.95	0	0.027273	0.022727	0
a015	0.977273	0	0.022727	0	0
a016	0.995455	0	0.004545	0	0
a017	0.972727	0	0.027273	0	0
a018	0.972727	0	0.027273	0	0
a019	0.990909	0	0.009091	0	0
a020	1	0	0	0	0
b001	0	1	0	0	0

UID	Neutral	Joy	Cold Anger	Sad	Hot Anger
b002	0.05	0.95	0	0	0
b003	0.05	0.95	0	0	0
b004	0.031818	0.968182	0	0	0
b005	0.190909	0.809091	0	0	0
b006	0.131818	0.822727	0.022727	0.022727	0
b007	0.018182	0.981818	0	0	0
b008	0.159091	0.840909	0	0	0
b009	0.154545	0.845455	0	0	0
b010	0.118182	0.881818	0	0	0
b011	0.275556	0.715556	0	0	0.008889
b012	0.086364	0.886364	0.022727	0.004545	0
b013	0.395455	0.604545	0	0	0
b015	0.377273	0.622727	0	0	0
b016	0.104545	0.895455	0	0	0
b017	0.168182	0.831818	0	0	0
b018	0.145455	0.809091	0.045455	0	0
b019	0.336364	0.640909	0.022727	0	0
b020	0.372727	0.627273	0	0	0
c001	0.009091	0.968182	0.022727	0	0
c002	0.004545	0.995455	0	0	0
c003	0.140909	0.859091	0	0	0
c004	0.022727	0.977273	0	0	0
c005	0.136364	0.863636	0	0	0
c006	0.136364	0.863636	0	0	0
c007	0.05	0.95	0	0	0
c008	0.131818	0.868182	0	0	0
c009	0.036364	0.963636	0	0	0
c010	0.022727	0.977273	0	0	0
c011	0.009091	0.990909	0	0	0
c012	0.154545	0.845455	0	0	0
c013	0.145455	0.85	0	0.004545	0

UID	Neutral	Joy	Cold Anger	Sad	Hot Anger
c015	0.022727	0.963636	0	0.013636	0
c016	0.154545	0.845455	0	0	0
c017	0.113636	0.863636	0.022727	0	0
c018	0.068182	0.909091	0.022727	0	0
c019	0.068182	0.886364	0.045455	0	0
c020	0.045455	0.954545	0	0	0
d001	0.109091	0	0.845455	0.031818	0.013636
d002	0.045455	0	0.927273	0.009091	0.018182
d003	0.245455	0	0.686364	0.068182	0
d004	0.245455	0	0.672727	0.077273	0.004545
d005	0.054545	0	0.936364	0.009091	0
d006	0.077273	0	0.872727	0.040909	0.009091
d007	0.027273	0	0.968182	0	0.004545
d008	0.1	0	0.845455	0.045455	0.009091
d009	0.109091	0	0.713636	0.154545	0.022727
d010	0.1	0	0.890909	0.004545	0.004545
d011	0.113636	0	0.813636	0.063636	0.009091
d012	0.072727	0	0.927273	0	0
d013	0.063636	0	0.922727	0.009091	0.004545
d015	0.113636	0	0.877273	0.009091	0
d016	0.25	0	0.772727	0	0
d017	0.031818	0	0.959091	0	0
d018	0.090909	0	0.872727	0.027273	0.009091
d019	0.072727	0	0.927273	0	0
d020	0.063636	0	0.927273	0	0.009091
e001	0.072727	0	0.904545	0.013636	0.009091
e002	0.090909	0	0.872727	0.018182	0.018182
e003	0.2	0	0.704545	0.095455	0
e004	0.231818	0	0.754545	0.013636	0
e005	0.104545	0	0.854545	0.031818	0.009091
e006	0.031818	0	0.931818	0.036364	0

UID	Neutral	Joy	Cold Anger	Sad	Hot Anger
e007	0.1	0	0.768182	0.131818	0
e008	0.045455	0	0.804545	0.145455	0.004545
e009	0.063636	0.022727	0.881818	0.022727	0.009091
e010	0.027273	0	0.859091	0.109091	0.004545
e011	0.045455	0	0.85	0.095455	0.009091
e012	0.104545	0	0.840909	0.05	0.004545
e013	0.068182	0	0.913636	0.018182	0
e015	0.059091	0	0.913636	0.022727	0.004545
e016	0.054545	0	0.9	0.040909	0.004545
e017	0.122727	0	0.863636	0.013636	0
e018	0.031818	0	0.959091	0	0.009091
e019	0.077273	0	0.909091	0.009091	0.004545
e020	0.136364	0	0.831818	0.027273	0.004545
F001	0.104545	0	0.05	0.845455	0
F002	0.045455	0	0.054545	0.9	0
F003	0.027273	0	0.018182	0.954545	0
F004	0.040909	0	0.05	0.909091	0
F005	0.036364	0	0.027273	0.936364	0
F006	0.027273	0	0.054545	0.918182	0
F007	0.059091	0	0.05	0.890909	0
F008	0.013636	0	0.054545	0.931818	0
F009	0.013636	0	0.113636	0.872727	0
F010	0.013636	0	0.05	0.936364	0
F011	0.059091	0	0.059091	0.881818	0
F012	0.004545	0	0.022727	0.972727	0
F013	0.045455	0	0.059091	0.895455	0
F015	0.018182	0	0.068182	0.913636	0
F016	0.068182	0	0.068182	0.863636	0
F017	0.072727	0	0.086364	0.840909	0
F018	0.013636	0	0.059091	0.927273	0
F019	0.036364	0	0.072727	0.890909	0

UID	Neutral	Joy	Cold Anger	Sad	Hot Anger
F020	0.013636	0	0.045455	0.940909	0
g001	0.159091	0	0	0.845455	0
g002	0.081818	0	0.009091	0.909091	0
g003	0.05	0	0	0.95	0
g004	0.018182	0	0	0.981818	0
g005	0.018182	0	0	0.981818	0
g006	0.036364	0	0	0.963636	0
g007	0.068182	0	0	0.931818	0
g008	0.063636	0	0	0.936364	0
g009	0.040909	0	0.004545	0.954545	0
g010	0.090909	0	0	0.909091	0
g011	0.059091	0	0	0.940909	0
g012	0.045455	0	0.009091	0.945455	0
g013	0.027273	0	0.004545	0.968182	0
g015	0.027273	0	0	0.972727	0
g016	0.013636	0	0	0.986364	0
g017	0.25	0	0	0.772727	0
g018	0.031818	0	0	0.968182	0
g019	0.127273	0	0	0.872727	0
g020	0.131818	0	0.022727	0.845455	0
h001	0.004545	0	0.068182	0	0.927273
h002	0	0	0	0	1
h003	0.027273	0	0.031818	0	0.940909
h004	0.059091	0	0.063636	0	0.763636
h005	0.018182	0	0.095455	0	0.886364
h006	0.004545	0	0.05	0	0.945455
h007	0	0	0	0	1
h008	0.004545	0	0.022727	0	0.972727
h009	0.018182	0	0.004545	0	0.977273
h010	0.004545	0	0.004545	0	0.990909
h011	0.009091	0	0	0	0.990909

UID	Neutral	Joy	Cold Anger	Sad	Hot Anger
h012	0.05	0.004545	0	0	0.945455
h013	0	0	0.004545	0	0.995455
h015	0.004545	0	0	0	0.995455
h016	0.004545	0	0	0	0.995455
h017	0	0	0	0	1
h018	0	0	0	0	1
h019	0	0	0	0	1
h020	0.022727	0.027273	0	0	0.95
i001	0.004545	0	0.036364	0	0.959091
i002	0	0	0	0	1
i003	0.027273	0	0	0	0.972727
i004	0	0	0.05	0	0.95
i005	0	0	0	0	1
i006	0.004545	0	0.022727	0	0.972727
i007	0.004545	0.022727	0.004545	0	0.968182
i008	0.004545	0	0.045455	0	0.95
i009	0.009091	0	0	0	0.990909
i010	0	0	0	0	1
i011	0.009091	0.018182	0	0	0.972727
i012	0.009091	0	0	0.004545	0.986364
i013	0	0	0	0	1
i015	0	0	0	0	1
i017	0.004545	0	0.022727	0.004545	0.968182
i018	0	0	0	0	1
i019	0	0	0	0	1
i020	0.068182	0.018182	0.013636	0	0.9

A-2. Rating Result of Experiment 2

The table listed the ratings of 15 utterances against 34 adjectives in terms of how appropriate the utterance was described by the adjective.

UID	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17
N1	0.625	0.208	0.500	0.583	0.250	0.042	0.375	0.208	1.125	0.250	0.625	0.125	0.208	0.500	0.208	0.958	0.208
N2	0.292	0.417	0.292	0.708	0.167	0.500	0.417	0.250	1.167	0.333	0.542	0.417	0.167	0.542	0.208	0.833	0.167
N3	0.458	0.125	0.417	0.458	0.333	0.083	0.375	0.083	1.333	0.042	0.625	0.167	0.292	0.250	0.375	1.000	0.083
J1	3.000	0.000	1.833	0.083	0.708	0.125	0.292	0.417	0.125	1.917	0.000	1.708	0.458	0.083	0.042	2.250	0.000
J2	1.875	0.000	1.625	0.208	0.542	0.125	0.167	0.500	0.667	0.708	0.167	1.000	0.542	0.458	0.417	1.083	0.125
J3	2.708	0.000	2.000	0.125	0.625	0.042	0.042	0.667	0.125	1.667	0.000	1.958	0.542	0.125	0.333	2.000	0.083
C1	0.042	1.667	0.167	1.875	1.583	0.000	1.542	0.083	1.542	0.042	1.917	0.042	0.958	0.417	0.750	0.750	0.917
C2	0.000	2.167	0.125	2.125	1.667	0.083	1.583	0.042	1.667	0.000	2.083	0.000	1.167	0.375	0.875	0.667	1.167
C3	0.167	2.000	0.167	2.208	1.667	0.083	1.458	0.042	1.542	0.000	2.167	0.000	1.167	0.458	0.833	0.875	1.250
S1	0.083	2.750	0.042	2.375	0.125	1.875	1.125	0.583	0.458	0.833	2.333	0.042	0.000	1.625	0.417	0.458	0.042
S2	0.042	2.583	0.208	1.833	0.000	2.542	0.167	1.833	0.417	1.125	1.708	0.250	0.000	1.375	0.292	0.667	0.083
S3	0.208	2.083	0.333	1.167	0.083	2.125	0.083	1.542	0.167	1.375	1.042	0.333	0.167	0.917	0.083	1.083	0.000
H1	0.667	0.375	1.458	0.375	2.500	0.000	1.333	0.167	1.792	0.042	1.375	0.167	2.333	0.000	2.167	0.625	2.542
H2	0.750	0.208	1.417	0.583	2.875	0.000	1.417	0.000	1.583	0.042	1.167	0.333	2.417	0.125	2.167	0.750	2.750
H3	0.917	0.208	2.167	0.208	2.792	0.000	1.083	0.167	1.250	0.000	0.792	0.167	2.417	0.083	2.042	0.792	2.667

UID	A18	A19	A20	A21	A22	A23	A24	A25	A26	A27	A28	A29	A30	A31	A32	A33	A34
N1	0.667	0.167	0.833	0.000	1.958	0.125	0.500	0.042	0.125	1.500	0.333	1.625	0.125	0.083	2.542	0.000	1.250
N2	0.875	0.167	0.875	0.083	2.375	0.042	0.625	0.042	0.042	1.458	0.125	1.917	0.083	0.125	2.667	0.083	1.083
N3	0.708	0.208	0.625	0.000	1.875	0.000	0.458	0.208	0.042	1.375	0.208	1.750	0.167	0.042	2.667	0.167	0.625
J1	2.083	0.125	0.167	0.083	0.875	0.333	2.042	0.000	0.000	2.042	0.000	2.167	0.000	1.708	0.125	0.625	0.375
J2	1.000	0.375	0.333	0.167	1.292	0.375	1.000	0.042	0.000	1.667	0.000	2.083	0.000	0.500	2.083	0.208	0.875
J3	1.792	0.417	0.083	0.292	0.708	0.542	1.708	0.000	0.000	2.000	0.042	2.167	0.000	1.667	0.125	1.083	0.125
C1	0.625	0.667	0.333	0.250	1.417	0.208	0.208	0.667	0.750	0.625	0.167	1.042	0.542	1.333	0.417	0.125	1.000
C2	0.750	0.708	0.625	0.500	1.125	0.583	0.167	0.708	0.875	0.833	0.208	0.833	0.708	1.125	0.417	1.167	0.250
C3	0.542	0.625	0.875	0.333	1.375	0.292	0.208	0.583	0.875	0.500	0.125	0.583	0.667	1.000	0.417	0.542	0.583
S1	0.750	0.083	2.042	0.042	2.333	0.000	0.167	0.583	1.250	0.125	0.667	0.250	1.208	0.625	0.667	0.000	2.208
S2	0.875	0.083	2.083	0.125	2.208	0.042	0.167	0.500	1.042	0.375	0.542	0.542	1.708	0.917	0.625	0.125	2.167
S3	1.375	0.042	2.042	0.042	2.458	0.000	0.875	0.042	0.167	0.625	0.458	1.042	0.583	1.292	0.208	0.167	1.667
H1	0.375	2.375	0.000	1.333	0.000	1.583	0.167	1.292	0.917	1.083	0.000	1.417	0.500	2.167	0.042	1.333	0.000
H2	0.458	2.708	0.000	1.833	0.000	2.042	0.042	1.542	1.250	1.083	0.000	1.875	0.417	2.542	0.083	1.375	0.083
H3	0.542	2.667	0.042	1.417	0.000	1.500	0.125	1.250	0.833	1.292	0.083	1.875	0.250	1.792	0.167	1.417	0.083

A-3. Detailed data of Regression Coefficient of

Analysis of a1, a2, a3, and correlation coefficient CC of each adjective were described Section 3.2.5. Each column indicated the adjective by id number (ID) as appeared in Table 3-3 (on page 22).

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
a1	0.402	-0.299	0.058	-0.296	-0.711	0.121	-0.452	0.156	-0.350	0.346	-0.462	0.302	-0.563	0.063	-0.496	0.235	-0.724
a2	0.566	-0.935	0.570	-0.635	0.505	-0.752	0.034	-0.348	0.184	-0.123	-0.481	0.252	0.453	-0.418	0.303	0.196	0.422
a3	-0.658	-0.069	-0.539	0.151	-0.183	-0.270	0.116	-0.300	0.431	-0.563	0.137	-0.451	-0.181	0.000	-0.106	-0.291	-0.113
CC	0.979	0.968	0.950	0.897	0.989	0.936	0.927	0.872	0.977	0.966	0.950	0.961	0.962	0.930	0.906	0.866	0.944

ID	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
a1	0.264	-0.574	0.121	-0.347	0.370	-0.336	0.357	-0.382	-0.338	0.265	0.019	0.227	-0.148	-0.370	0.511	-0.265	-0.296
a2	0.024	0.472	-0.695	0.262	-0.619	0.354	0.178	0.085	-0.159	0.452	-0.187	0.530	-0.379	0.238	0.082	0.284	-0.635
a3	-0.343	-0.208	0.032	-0.181	0.427	-0.247	-0.336	-0.076	-0.090	-0.067	0.021	-0.057	-0.086	-0.652	0.984	-0.305	0.151
CC	0.885	0.895	0.984	0.881	0.981	0.890	0.911	0.910	0.890	0.975	0.931	0.952	0.898	0.956	0.971	0.904	0.605

REFERENCES

[1] Jef Raskin, *The Humane Interface: New Directions for Designing Interactive System*, Pearson Education, 2000.

[2] Carl E. Williams and Kenneth N. Stevens, *On Determing the Emotional State of Pilots during Flight: An Exploratory Study. Aerospace Medicine*, 40(12):1369-1372, Dec 1969.

[3] Carl E. Williams and Kenneth N. Stevens, *Emotions and Speech: Some Acoustical Correlates.* JASA, Vol.52, Number 4, Part 2, pp.1238-1250, 1972

[4] ; . 2002. "怒りの感情音声における音響特徴量と感情知覚との関係に関する研究"、北陸先端科学技術大学院大学修士論文.

[5] Janet E. Cahn, *Generating Expression in Synthesized Speech, Master thesis,* MIT, Media Laboratory, 1990.

[6] K. Ueda, M. Akagi, Sharpness and amplitude envelopes of broadband noise, JASA, Vol.87, Number 2, pp.814-819.1990

[7] E. Zwicker, *Psychoacostics*, Springer, 1982.

[8] D. W. Massaro. *Perceiving Talking Faces from Speech Perception to a Behavioral Principle*. MIT Press, 1997.

[9] M. Kuriyagawa, H. Yahiro, S. Kashiwagi, *Seven attributes in tone quality evaluation*, J. Acost. Soc. Jpn., 34, 493-500, autumn 1978 (in Japanese).

[10] J. M. Grey, *Multidimensional perceptual scaling of musical timbres*, J. Acoust. Sco. Am. 61, 1270-1277, 1977.

[11] J. M. Grey, J. W. Gordon, *Perceptual effects of spectral modifications on musical timbres*, J. Acoust. Sco. Am. 63, 1493-1500, 1978.

[12] R. Plomp, L. C. W. Pols, J. P. van de Geer, *Dimensional analysis of vowel spectra*, J. Acoust. Soc. Am. 41 707-712, 1967.

[13] K. Ohgushi, *Physical and psychological factors governing timbre of*

complex tones, J. Acoust. Soc. Jpn. 36, 253-259, 1980 (in Japanese).

[14] 岡太彬訓, 今泉忠, 『パソコン多次元尺度構成法』, 共立出版, 1994.

[15] 林知己夫, 飽戸弘, 『多次元尺度解析法：その有効性と問題点』, サイエンス社, 1976.

[16] J.B. Kruskal, Mmultidimensional Scaling, SAGE Publications, 1978.

[17] K. Ueda, *Should We Assume a Hierarchical Structure for Adjectives Describing Timbre?*, J. Acoust. Soc. Jpn. 44, pp. 102-107, 1988 (in Japanese).

[18]J. S. R. Jang, C. T. Sun, E. Mizutani. *Neuro-Fuzzy and Soft Computing*, Prentice Hall, 1996.

[19] Mathworks, *MATLAB Fuzzy Logic Toolbox Manual.*

[20] V. Kecman, *Learning and Soft Computing: Support Vector Machines, Neural Network, and Fuzzy Logic Models*, MIT, 2001.

[21] O. Wolkenhauer, *Data Engineering Fuzzy Mathematics in Systems Theory and Data Analysis*, John Wiley & Sons, 2001.

[22] S. Chiu, *Fuzzy Model Identification Based on Cluster Estimation*, Journal of Intelligent & Fuzzy Systems, Vol.2, No.3, Sept. 1994.

[23] Sugeno, M., *Industrial Applications of Fuzzy Control*, Elsevier Science Pub. Co., 1985.

[24] H. kawahara., I. Masuda-Katsuse, A. Cheveigne, *Resturcturing 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, Vol.27, pp. 187-207, 1999.