Emotion recognition in multiple languages using a three layer model

By Xingfeng LI

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

March, 2016

Emotion recognition in multiple languages using a three layer model

By Xingfeng LI (1310211)

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
Associate Professor Masashi Unoki
Professor No idea
Associate Professor No idea

February, 2016 (Submitted)

Contents

Chapter 1

Introduction

1.1 Research Background

Speech is the most nature and the fastest mean of communication among human beings in our lives. And it has also became the very appealing research topic in the area of man-machine communication. Science the recent past, most existing speech studies have been presented to detect the linguistic contents of a spoken sentence, namely speech recognition. Whereas, in a nature conversation, irrespective of the message conveyed through the text, non-verbal information embedded in the utterance is crucial as well. Given that the same textual contents could have different interpretations, relying on the way that expressed [1]. Thereafter, increasing attention has been aimed at going into the investigation of affective contents of speech. Nowadays, it is generally believed that only possessing the ability to analysis underlying emotions effectively can speech systems may achieve human performance.

Speech emotion recognition (SER) system has become so widespread in day-to-day life. It can be implemented in an on-board car driving system to keep him or her alert during driving by detecting the mental state of a driver from his or her speech [2]. It is also available in call center applications for analyzing frustration or exasperation of clients, and helping to promote service quality of a call attendant [3]. Beyond these, it is particularly helpful for affective speech to speech translation (S2ST) system in which the spoken utterance in the source language is translated into the target one, and here the translated speech is rendered with the same affective states that conveyed in the original speaker's message [4]. Such kinds of systems are full of meaningful as for promoting cultural exchange activities to foster mutual understanding around the world.

1.2 Problem Statement

Humans know the emotions while they perceive them. This fact motivated researchers to investigate and identify different aspects of emotions. Most of the previous body of researchers generally agreed that emotion can be depicted as discrete categories, like joy, anger, neutral, sadness, fear, disgust, and boredom [5]. Differently, the intensity

of a certain emotion always varied with time and situation so that any small numbers of emotional categories are insufficient to reflect such rich variation of degree during conversations. For this reason, to have a nature automatical SER system, it must satisfy the required conditions of detecting degree or level of the affective contents from a uttered sentence. Hereafter, these days it is universally advocated to capture emotion as a point using dimensional description of human affects [6]. This is due to dimensional space are well suited to present the changeful intensities of expressed emotion by the use of continuous scales.

In our research, a two emotion dimensional space in valence and activation is utilized. Valence is a quality of being (subjectively) charged with positive or negative significance, which is inherent in emotional appraisal and experience; Activation quantifies how a speech is associated to an intensity level, particularly strong or weak. Such a dimensional approach has been investigated by a emotion theorist, Russel, and it was proved that valence and activation are universal primitives [7]. Obviously, to characterize the affective state in spoken utterances, dimensional method is much more appropriate for describing emotion in everyday interactions.

Disregard of the substantial advance on human underlying emotions description, automatical SER system still faces a number of serious competitions. One of them is the estimation of emotional dimensions, especially for valence. So far, it is still a challenging work to study the most relative acoustic features to distinguish and model dimensions. Valster and Schuller, in 2013 attempted to predict the emotion dimensions of valence and arousal from extracted audio features using Support Vector Machine Regressors with an intersection kernel [8]. Howbeit, obtained results of valence performed poorly. Most researchers previously try to study many speech features and their relationship between affective contents of the speech utterance, whereas, it is hardly improved the result of valence estimation [9] [10].

Another challenge is that most existing speech emotion recognition systems have focused on monolingual emotion classification. By using different kinds of feature selection strategies to collect the best feature sets, only sparse works investigate the task of bilingual or multilingual classification[11] [12]. Unfortunately, given conclusions from different studies always have different versions. Emotion classification cross cultures does not have a commonly agreed recognizer.

1.3 Objective of the Study

Generally, speech emotion perception is cross cultures, even the spoken language is incomprehensible. It turned out that speech usually have universal attributes not only among individuals in the same nation, but also among countries [13] [14]. From the point of the view of human perception, one previous study on investigating the commonalities and differences of emotion perception across languages is achieved [15]. Experimental results pointed that directions from neutral voice to other emotional states in valence-activation space are common among languages, however, the neutral position are culture-dependent. The most appealing conclusion following [15] is that, directions from neutral state to e-

motions could be adopted as features to recognize affective states in multi-languages.

Therefore, motivated by the knowledge of commonalities and differences of human perception for emotional speech among multiple languages. The aim of this study is to construct a multiple speech emotion recognition system (works on Japanese, German, and Chinese) which has the ability to precisely estimate emotion dimensions nearly the same as human responses. To achieve this goal, we study two crucial aspects in speech emotion recognition: (1) Predicting emotion dimensions accurately such as human annotated, (2) Extracting commonalities among languages to normalize cultural differences.

1.4 Structure of the Thesis

This paper is structured as follows: Chapter 1 firstly presents the background and the meaning of this investigation, hereafter, the exiting issue in the research field of emotion recognition and the objective of this study is defined. Chapter 2 addresses a related literature review on the state-of-the-art emotional research from aspects of emotional description, human emotion perception, and previous related literature review, finally, the overview of proposed system is introduced. Chapter 3 describes the collection of data of the proposed system. The utilized emotional corpus: Fujitsu database, Berlin Emo-DB, and CASIA emotional corpus. Additionally, acoustic features extraction, and experimental evaluation of semantic primitives and emotion dimensions, fuzzy inference system are discussed. Chapter 4 depicts the implementation of the proposed model, afterwards are the evaluation of system from the point of the view between human responses, different systems in bilingual scenario and in monolingual case. Chapter 5 eventually summarizes this thesis with respect to the research question and give an outlook on future work.

Chapter 2

Literature Review

2.1 Emotion Description

To investigate the relations between speech and emotion, the fist key step is choosing an appropriate method for describing emotion. In the are of speech emotion recognition, generally, there are two main methods to represent emotional states: categorical, and dimensional approach. In the following two subsections, detailed discussions of two descriptions are introduced.

2.1.1 Categorical Method

Researchers utilize categorical method agreed that there are a small number of emotions that are basic, the most occurred and recognized commonly [16]. A previous study on this idea, described by Descartes (Anscombe and Geach, 1970), supposed that the basic emotional categories are 'primary', naming, other emotions can be composed with primary affections such as the fact that any color come from a combination of three-primary color. It has been called palette theory [17]. Primary emotions are, fear, joy, anger, disgust, sadness, and surprise. Moreover, in the age of the 1990s several theorists claims that lists of basic emotions should be considered longer,in which they listed more positive and/or interpersonal emotions. Table 1 summarized important cases, from Lazarus (1999a) and Ekman (1999). Beyond them are lists that reflect different aspects. Buck (1999) presents a symbolism of what he named affective states rather than emotions per se; the table shows the states that he depicts. The list after Banse and Scherer (1996) is the most systematic in the literature concerned obviously with speech. The last column are basic emotion vocabulary developed by Cowie et al., it is a set of emotion-related categories which is small enough to be coped with, and that contains the range of emotion-related states that mostly occur.

But, some of researchers argued that giving subjects selection of just one category from a small number of basic emotions may not provide humans an sufficient level of discrimination [18]. Obviously, the intensity of a certain emotion always varied with time and situation so that any small numbers of emotional categories are insufficient to reflect

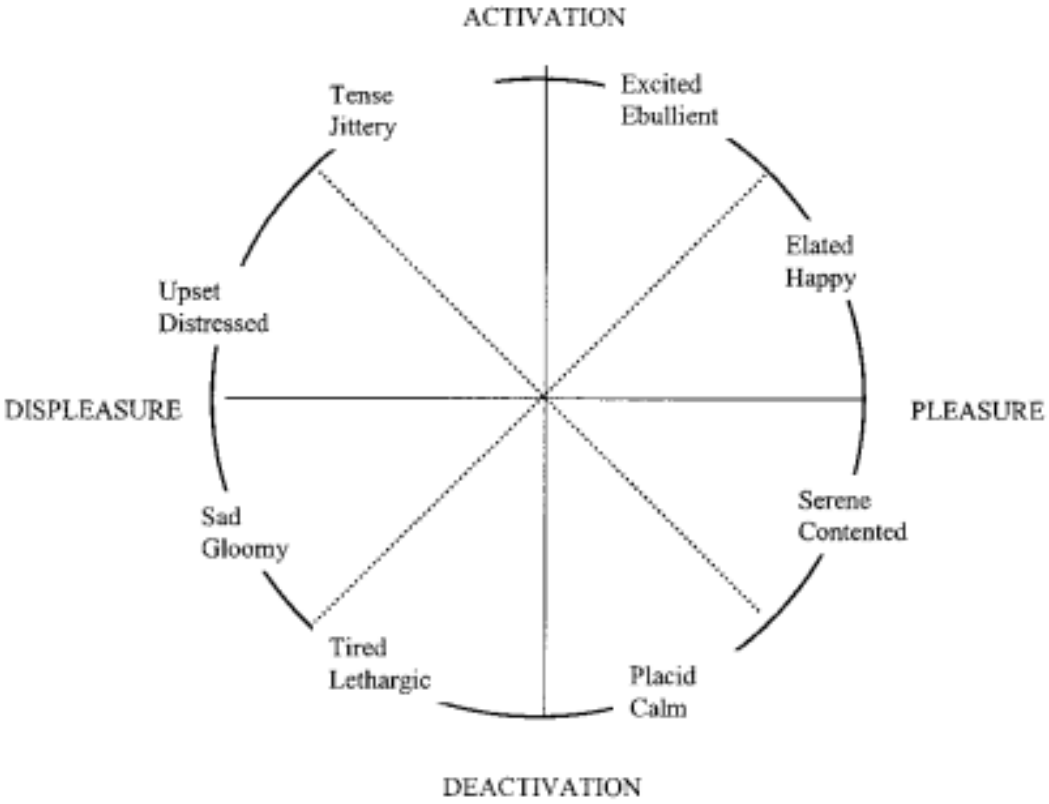


Figure 2.1: Core affect [21].

such rich variation of degree during conversations.

2.1.2 Emotion Dimensional Method

Motivated by some reported psychological literatures, many studies advocate the use of dimensional approach to describe human emotion, where affections are not binary, i.e. angry or happy, rather they are characterized as a point in a multiple dimensional space [19] [20]. Dimensional approach was long investigated by Russel [21], and it was suggested that the existence of valence (pleasure/displeasure) and arousal (activation/deactivation) are two fundamental dimensions of emotional description. Russel claimed that they were common primitives and named the impression at any position in dimensional space core affect as shown in Figure 2.1. In this dimensional space, the main axes are spanned with: one (valence) running from very pleasure to very displeasure; the other (arousal) running from very active to very passive. This emotion dimensional space has very wide support in modern emotion research.

Particularly that fear and anger are difficult to distinguish while presenting in the

Table 2.1: Lists of key emotions

Lazarus (1999a)	Ekman (1999)	Buck (1999)	Banse and Scherer (1996)	Cowie et al. (1999b)
Anger	Anger	Anger	Rage/hot anger	Angry
			Irritation/cold anger	
Fright	Fear	Fear	Fear/terror	Afraid
Sadness	Sadness/distress	Sadness	Sadness/dejection	Sad
			Grief/desperation	
Anxiety		Anxiety	Worry/anxiety	Worried
Happiness	Sensory pleasure	Happiness	Happiness	Happy
			Elation (joy)	
	Amusement			Amused
	Satisfaction			Pleased
	Contentment			Content
		Interested		Interested
		Curious		
		Surprised		
	Excitement			Excited
		Bored	Boredom/indifference	Bored
				Relaxed
		Burnt out		
Disgust	Disgust	Disgust	Disgust	
	Contempt	Scorn	Contempt/scorn	
Pride	Pride	Pride		
		Arrogance		
Jealousy		Jealousy		
Envy		Envy		
Shame	Shame	Shame	Shame/guilt	
Guilt	Guilt	Guilt		
	Embarrassment			
				Disappointed
Relief	Relief			
Hope				
				Confident
Gratitude				
Love				Loving
				Affectionate
Compassion		Pity		
		Moral rapture		
		Moral indignation		
Aesthetic				

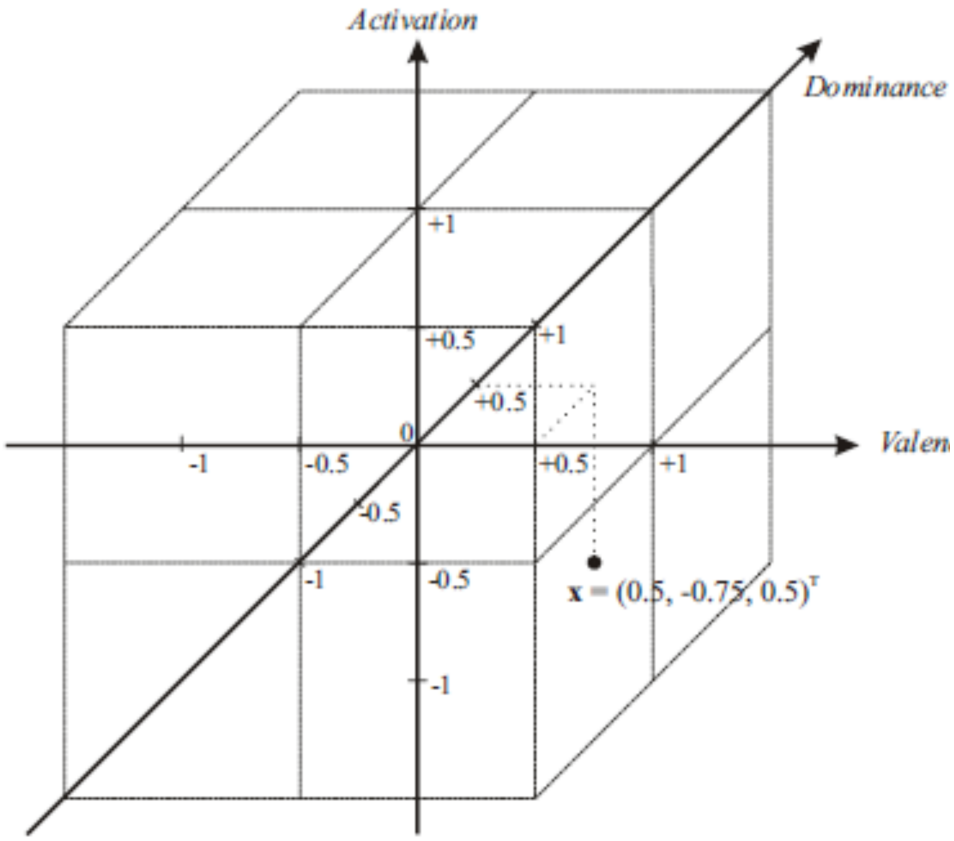


Figure 2.2: Three-dimensional emotion space, spanned by the primitives valence, activation, and dominance, with a sample emotion vector added for illustration of the component concept [10].

valence-arousal space. Other studies have found "dominance", a third emotional dimension to distinguish anger from fear in this case [10]. They proposed a generalized framework achieved by a three-dimensional emotion space. Seen from Figure 2.2, it defines emotions as point in a three-dimensional space spanned by the three basic dimensions valence (negative/positive), activation (calm/excited), dominance (weak/strong).

According to the work of Russel (2003) and Grimm (2007), considering that four basic emotions, i.e. neutral, anger, sadness, and happiness are being studied in this work. Two powerful dimensions of valence and activation are finally utilized to represent emotions in our research. Important reason is the fact that it shows great potential to present the occurrence of emotions in our real lives, emotions are not generated in a prototypical or pure modality, but expressed with varying degrees of intensity. Hence, dimensional method provides an essential framework for describing dynamics in emotions, tracking intensities in the course of time.

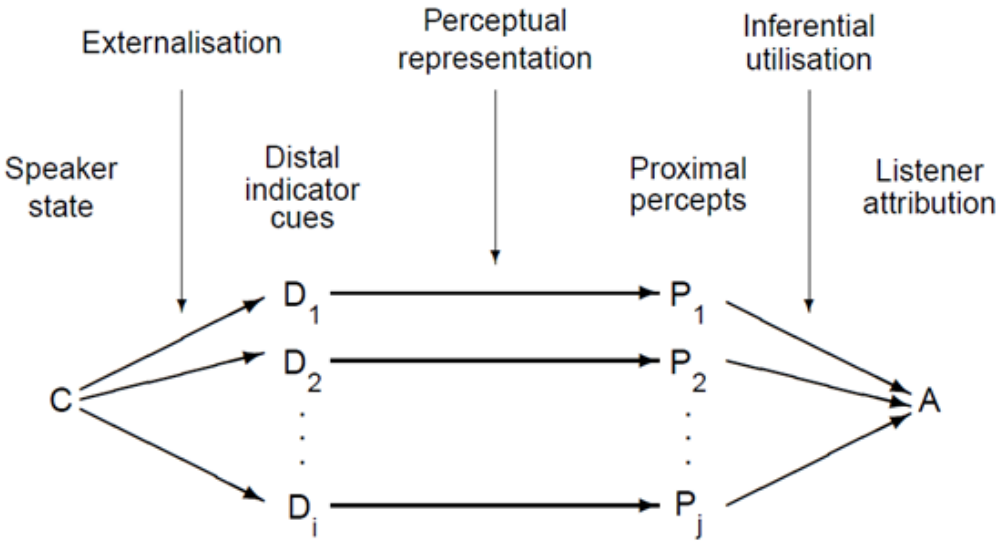


Figure 2.3: The Brunswikian lens model, adapted from Scherer (1978) [22].

2.2 Human Emotion Perception Processing

Tremendous of previous works attempted to predict emotion dimensions directly from acoustic features. However, such kinds of models can not restore the human processing on emotion perception. Scherer [22] adopted a version of Brunswik's lens model (1956) to describe human emotion perceptual processing [23], by a three-layered model as shown in Figure 2.3. The steps of human perception according to Scherer model are as follow:

- the emotional state of spoken utterances of a speaker can be decomposed into several "distal indicator cues", naming acoustic features.
- in the first place of the perceptual inference process, a listener will perceive the emotional state of a speaker from distal indicator cues indirectly by a smaller "proximal percepts".
- these percepts are used to infer the speakers's emotion (attribution) by the listener. In speech and emotion examples of proximal percepts are subjectively perceived pitch or voice quality.

Previously, Elbarougy proposed an improved Brunswik's lens model for human perception in dimensional approach after Huang and Akagi [24] [11], with the assumption that human perception for emotional voice is not directly from a change in acoustic features but rather a number of different types of smaller perception that are expressed by semantic primitives/adjectives to describe an affective utterance. The processing on human emotional perception to recognize the emotion expressed by speakers are shown in Figure 2.4. The human perception consists of two parts: the first sub-process is perception on semantic primitive, in which listeners perceive the degree of all adjectives to present the

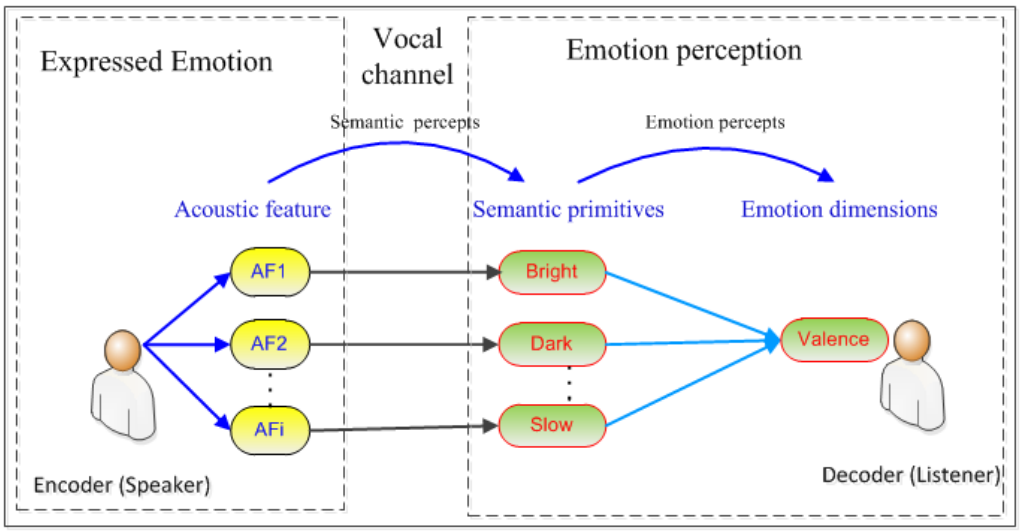


Figure 2.4: The improved Brunswikian lens model for human perception, adapted by Elbarougy after Huang and Akagi [24] [11].

emotional voice, such as very heavy, quiet slow and so on. The followed is emotional cognition by recognizing the degree of emotional state from adjectives to describe uttered speech.

This human perception-based model consists of acoustic features in the bottom layer, semantic primitives in the middle layer, and emotion dimensions in the top layer. Related works after Elbarougy and Akagi on monolingual emotion recognition have proved that this model can effectively improve the estimation of emotion dimensions, especially for valence. In this study, after this three-layered model, we will utilized it on multiple languages case.

2.3 Review of Related Research

In line with these studies, it can be concluded that two-layered perceptual model is poorly to estimate emotion dimensions directly from acoustic features [8] [9] [10]. However, the meaningful and worthwhile two findings are that,(1) dimensional approach is appropriate for representing emotion states,(2) human emotion perception modeled as multiple layers form [11], which emotions are usually described using various adjectives as a bridge not directly from acoustic features. It indicated that three-layered human perception model effectively works for imitating mono-lingual speech emotion perception. But, to construct a common emotion recognition system that be analyzed regardless of the language used, the limitation of retraining issue is still a challenge. To investigate whether emotional states can be recognized universally or not, Elbarougy proposed a bilingual speech emotion recognition system using dimensional approach [11] derived from a three-layered model adopted by Huang and Akagi [24]. Described as Figures 2.5, 2.6, 2.7. this bilingual perceptual model was constructed with combined information between two different

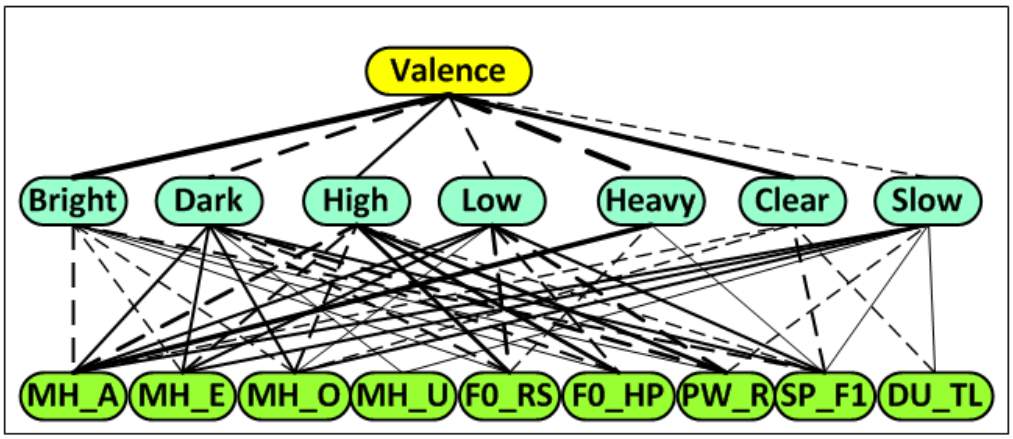


Figure 2.5: The German perceptual three-layer model for valence [11].

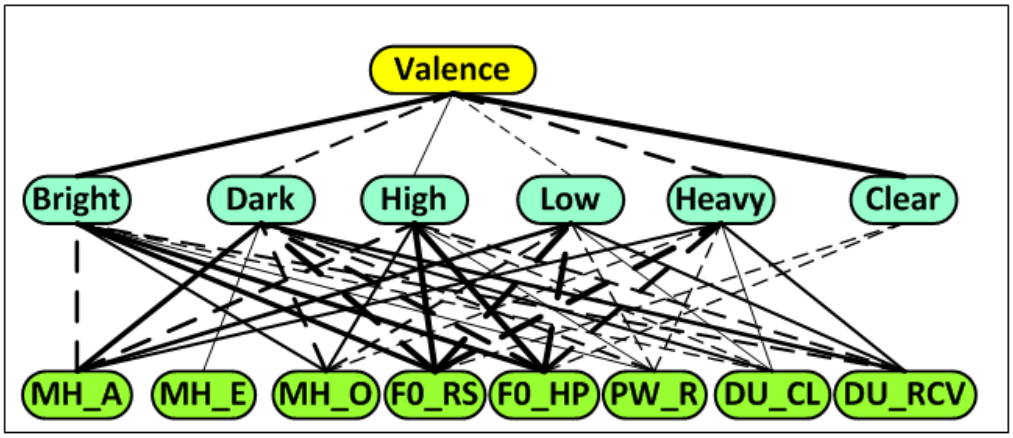


Figure 2.6: The Japanese perceptual three-layer model for valence [11].

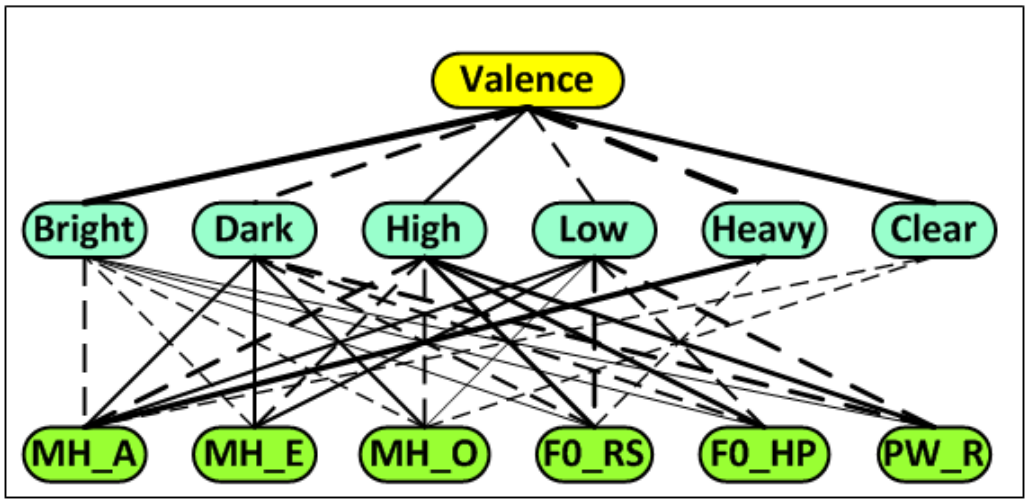


Figure 2.7: The Bilingual perceptual three-layer model for valence [11].

databases, one in Japanese and the other in German.

As for constructing this bilingual speech emotion perceptual model that works for Japanese and German languages, firstly, Elbarougy constructed a perceptual three-layered model individually for each dimensions for the two databases. Then the common acoustic features between two languages were selected to constitute the bottom layer. Moreover, the common semantic primitives between the two-languages were selected as semantic primitives for the bilingual case. But, to build a common speech emotion recognition system across languages, it is impractical to construct a three-layered perceptual model individually for each language to select common features in multiple languages cases. This is because optimal feature sets for different languages are always language-dependent. Hence there is no guarantee of numerous highly correlated features for each dimension.

The introduced bilingual SER system originally from Elbarougy's study was validated by training the system using one language, and testing using the second language. For instance, to estimate emotion dimensions for Japanese from German, the acoustic features, semantic primitives and emotion dimensions for German database were used to train this system, then the trained system is used to estimate emotion dimensions for Japanese database. To avoid language independent, the tested acoustic features from Japanese language should be normalized by dividing the values of Japanese acoustic features by the mean value of neutral utterances in German database for all Japanese acoustic features. Lastly, the Japanese normalized acoustic features are used as input to the trained system to estimate emotion dimensions for Japanese, and vice-versa. However, actually the normalization method in the acoustic features layer for avoiding language dependency are difficult to accurately predict positions in the dimensional space because elements in three layers are connected nonlinearly by FIS. Linearly normalized parameters in the input space can not ensure correct predicted values in the output space.

Unfortunately, it is found that this model can only work for several language pairs, applications of normalization and common features selection methods into the bilingual emotional speech recognition system resulted degradation of accuracy in emotional states estimation compared with mono-language cases.

2.4 Overview of Multiple Languages Speech Emotion Recognition System

People can perceive the expressive contents of a spoken utterance of one language, such as affection, even can not understand the language used. Several investigations have proved that there are some common attributes in speech not only between subjects of the same culture, but cross nations [13] [14]. In 2015, commonalities and differences of emotion perception across languages have been studied by carrying out human listening tests [15]. Obtained results from that study indicating that direction and distance from neutral to other emotions are similar among languages. Experimental results can be seen from Figure 2.8.

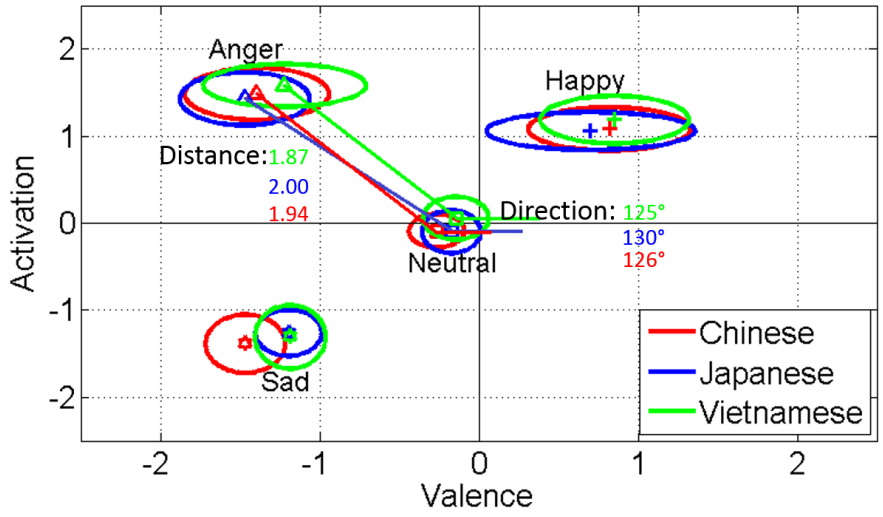


Figure 2.8: Positions of emotional states on Valence-Activation emotion space from [15]. Lines from Neutral to Anger indicate directions and distances for three subject groups.

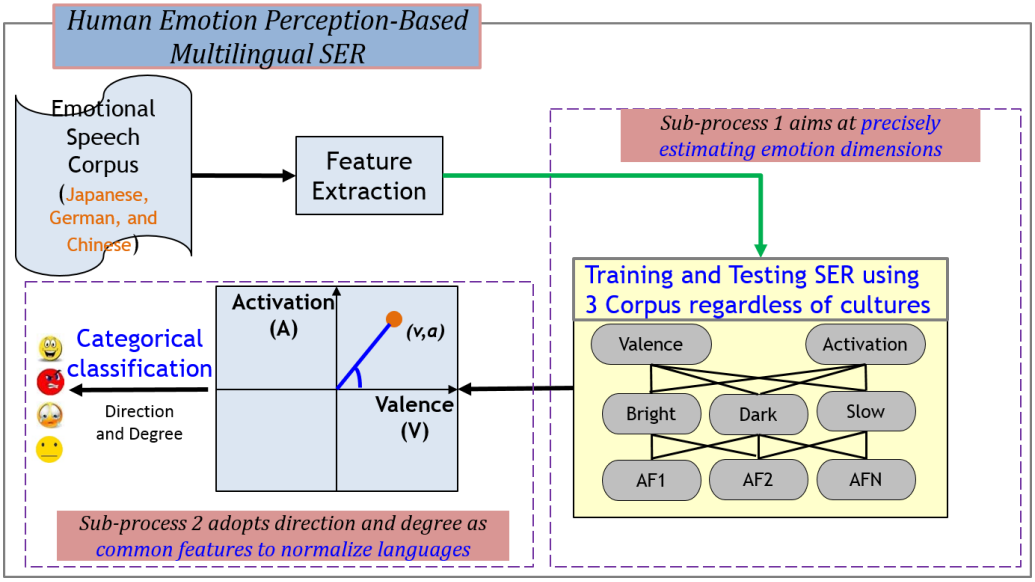


Figure 2.9: The architecture of proposed speech emotion recognition system using three-layered perceptual model.

Motivated by this new finding, regarding directions and degrees in the emotion dimensional space as distinguished features among emotions, the generalized SER system can be easily implemented with the assumption that the generalized SER system be able to precisely predict emotion dimensions in multiple language case. The block diagram of emotion recognition analysis in this study is shown in Figure 2.9.

To achieve this purpose, from existing emotional speech corpus, acoustic features which are input parameters of the SER system should be extracted in the first place. For building a generalized speech emotion recognition system, researchers previously always attempt to find out common features for different languages. Investigation on common acoustic features cross languages can be seen as the most immediate and easiest motivation. These days, most of existing investigations focus on studying numbers of speech features and their relations to the affection of utterances, simultaneously, various features selection methods are used to find the best features, unfortunately conclusions from different works are often inconsistent [9] [10] [14]. Even selecting common acoustic features between two different databases, the number of best common acoustic features is scanty and weak [11]. This revealed that it is very difficult to determine a common feature set from acoustic cues to construct such a generalized speech emotion recognition system. Motivated by this new finding summarized in Figure 2.8, based on the human processing, in this work, we are training and testing this three layered model regardless of linguistical/cultural impacts by mixing dataset from three corpus.

Since several works have fully proved that speech usually have universal attributes not only among individuals in the same nation, but also among countries [13] [14]. Additionally, the knowledge of commonalities and differences on human affective perception among languages has been found in the emotion dimensional space. It revealed that directions from neutral voice to other emotional states in valence-activation space are common among languages, and the neutral position are culture-dependent.

With this interesting finding, with the assumption that the proposed system could be able to precisely predict emotion dimensions for multiple languages just the same as human do. Afterwards, we will extract direction and degree in the dimensional space as distinguished features to classify different emotional states cross cultures.

The next chapter will state the databases, and data of elements of the proposed multiple SER system. Eventually, the procedures of predicting emotion dimensions using Fuzzy Inference System are introduced.

Chapter 3

Emotional Corpus and Elements Collection of Three-layered Model

3.1 Emotional Corpus

In this study, to achieve a multiple languages speech emotion recognition system, three emotional speech corpus involving Japanese, German, and Chinese are utilized. The characteristics of used databases are briefly presented in Table 3.1. Considering that our research aims at verifying and putting forward some new ideas which can be useful to the investigation of emotional conversation, instead of developing a real-life application. In order to have a constant testing environment without uncertainties, consequently acted emotions are selected as preferential ones. The first one is Fujitsu database, a multiple emotions single speaker produced and recorded by Fujitsu Laboratory. The second is the well known Berlin database that is broadly utilized in the field of emotion recognition. The third one is CASIA database, which is also produced by professional actors including males and females of Chinese native speakers. In the following three subsections, we will describe the used emotional corpus and chosen utterances in details.

Table 3.1: Characteristics of selected emotional speech corpus.

Corpus	Access	Language	Size	Source	Emotions
Fujitsu	Private[a]	Japanese	1 actors, 5 emotions, 179utterances	Professional actors	Neutral, joy, sadness, cold anger, hot anger.
Emo-DB	Public and free[b]	German	10 actors, 7 emotions, 800utterances	Professional actors	Anger, joy, sadness, fear, disgust, boredom, neutral
CASIA	Private[c]	Mandarin	4 actors, 6 emotions, 2400 utterances	Professional actors	Neutral, angry, happy, sad, fear, surprised.

[a] Fujitsu Laboratories Ltd., Japan
[b] Institute of Speech and Communication, Department of Communication Science, the Technical University, German
[c] Institute of Automation affiliated with Chinese Academy of Science, China

Table 3.2: Lists of uttered sentences in Fujitsu database, and translated version in English.

	in Japanese Sentence	in English Translation
1	Atarashi meru ga todoite imasu	Youve got a new mail.
2	Atama ni kuru koto nante arimasen	There is nothing frustrating.
3	Machiawase wa Aoyamarashi ndesu	I heard that we would meet in Aoyama.
4	Atarashi kuruma o kaimashita	I bought a new car.
5	Iranai meru ga attara sutete kudasai	Please delete any unwanted e-mails.
6	Sonna no furui meishindesu yo	Thats an old superstition.
7	Minna kara eru ga okura reta ndesu	Many people sent cheers.
8	Tegami ga todoita hazudesu	You should have received a letter.
9	Zutto mite imasu	I will think about you.
10	Watashi no tokoro ni wa todoite imasu	I have received it.
11	Arigatogozaimashita	Thank you.
12	Moshiwakegozaimasen	I am sorry.
13	Arigato wa iimasen	I wont say thank you.
14	Ryoko suru ni wa futari ga i nodesu	Id like to travel just the two of us.
15	Ki ga toku nari-sodeshita	I felt like fainting.
16	Kochira no techigai mogozaimashita	There were our mistakes.
17	Hanabi o miru no ni goza ga irimasu ka	Do we need a straw mat to watch fireworks.
18	Mo shinai to itta janaidesu ka	You said you would not do it again.
19	Jikandorini konai wake o oshietekudasai	Tell me the reason why you don't come on time, please.
20	Sabisueria de goryu shimashou	Meet me at the service area.

Table 3.3: The used categories in Japanese database. The second column shows the utterances id. There are two patterns for each emotion category: Joy, Cold Anger, Hot Anger, and Sadness. And only one pattern for Neutral.

Emotional speech category	Utterances id
Neutral	a001-a020
Joy	b001-b020 c001-c020
Cold anger	d001-d020 e001-e020
Sadness	f001-f020 g001-g020
Hot anger	h001-h020 i001-i020

3.1.1 Fujitsu Database

The Japanese database is developed by Fujitsu Laboratory with one professional female speaker. The professional female actress was asked for acting the uttered sentence with 5 emotional speech categories involving joy, cold anger, hot anger, neutral, and sadness. In the acted corpus, there are totally 20 difference Japanese utterances as listed in Table 3.2.

The actress speaker produced each utterance nine times: once for neutral state, and twice for each of the rest 4 emotions (joy, hot anger, cold anger and sadness) as illustrated in Table 3.3. Totaly there are nine acted utterances for each sentence and 180 utterances for 20 different sentences. Because of one missed utterance in cold anger, eventually the number of utterances in Fujitsu dataset is 179.

To be consistent in the use of categories among three emotional speech databases, 140 utterances in Japanese from 4 emotional categories of neutral, joy, hot anger, and sadness are finally utilized exclude cold anger.

3.1.2 Berlin Emo-DB

The German database is Berlin database of emotional speech, which is published by the Technical University of Berlin [25]. Berlin Emo-DB is one of the most popular corpus utilized by researchers in the field of emotion recognition. In this database, ten actors (5 females and 5 males) each uttered 10 German sentences (5 longer and 5 short, typically from 1.5 to 4 seconds) to act 7 emotions. The numbers of utterances for these emotional categories presented in Berlin Emo-DB are: boredom (81), fear (69), anger (127), joy (71), disgust (46), neutral (79), and sadness (62). The corpus is produced in 16 bit, 16KHz under studio noise conditions. To chose categories similar to those in the Japanese database (joy, hot anger, sadness, and neutral). Eventually, 50 joy, 50 anger, 50 sadness, and 50 neutral, totally 200 utterances are collected from Berlin dataset. 100 utterances were spoken by 5 males and the rest were spoken by 5 females. Table 3.4 and Table 3.5 describe the number of sentences which chosen from the male and female separately. Additionally, the spoken contents of German emotional data can be seen in Table 3.6.

Table 3.4: Used male utterances in Berlin database

Speaker ID	Neutral	Joy	Anger	Sad	Total
M03	6	7	4	7	24
M10	4	3	4	3	14
M11	5	7	6	7	25
M12	3	2	6	4	15
M15	7	6	5	4	22
Total	25	25	25	25	100

Table 3.5: Used female utterances in Berlin database

Speaker ID	Neutral	Joy	Anger	Sad	Total
F08	6	9	4	8	27
F09	4	1	6	4	15
F13	8	7	3	2	20
F14	2	6	5	4	17
F16	5	2	7	7	21
Total	25	25	25	25	100

Table 3.6: Lists of uttered sentences in Emo-DB database, and translated version in English.

	in German sentence	in English translation
a01	Der Lappen liegt auf dem Eisschrank.	The tablecloth is lying on the frigde.
a02	Das will sie am Mittwoch abgeben.	She will hand it in on Wednesday.
a04	Heute abend knnte ich es ihm sagen.	Tonight I could tell him.
a05	Das schwarze Stck Papier befindet sich da oben neben dem Holzstck.	The black sheet of paper is located up there besides the piece of timber.
a07	In sieben Stunden wird es soweit sein.	In seven hours it will be.
b01	Was sind denn das fr Tten, die da unter dem Tisch stehen?	What about the bags standing there under the table?
b02	Sie haben es gerade hochgetragen und jetzt gehen sie wieder runter.	They just carried it upstairs and now they are going down again.
b03	An den Wochenenden bin ich jetzt immer nach Hause gefahren und habe Agnes besucht.	Currently at the weekends I always went home and saw Agnes.
b09	Ich will das eben wegbringen und dann mit Karl was trinken gehen.	I will just discard this and then go for a drink with Kar.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

3.1.3 CASIA Database

The CASIA emotional speech dataset is a Mandarin emotional speech dataset developed by the Institute of Automation affiliated with Chinese Academy of Sciences. It was recorded by 2 male and 2 female speakers with neutral and 5 categories of acting emotions, including angry, happy, sad, fear and surprise. The contents consist of two parts, naming dominant and spontaneous. The utterances of the dominant contents have at least one dominant word, e.g. anger or annoyed for angry, pleased or joyful for happy, and sad for sad, etc. There are 100 utterances for each emotion. The utterances of the spontaneous contents are picked from news articles, conversations and essays without emotional-rich words. There are 300 utterances in this part. Each speaker utters $(100 + 300)\times6 = 2400$ sentences in total. Among them, the spontaneous speech can be taken as content aligned data and the dominant speech are content discriminative. Between these two categories,

Table 3.7: Lists of uttered sentences in CASIA database, and translated version in English.

ID	in Chinese	in English translation
406	shi fen fu za de mi mi dou zheng.	A completely complex and secret struggle.
418	du shen yi ren zai na biao yan.	He/She acts there alone.
425	gu niang zai xi ju zhong he bie de nan ren tan hua shi.	While the girl talked to the other man in the theater.
427	ta jie shu le zhe chang hao wu yi yi de lian ai.	He/She ended the meaningless love
431	zhi jie liao dang di dui ta biao bai.	Expressing your love to her openly.
432	xiao shi hou ting zu mu jiang guo yi ge gu shi.	When I was a little girl, my grandma told me a story.
440	shan dian de chang du ke neng zhi you shu bai qian mi.	The length of lightning maybe few thousand kilometre.
441	mai gei na xie chi you zheng shu de fa guo wu qi zhi zao shang.	Sell them to the France certificated weapons companies.
447	ji zhu zhe ge qi ji bei fa xian de shi jian.	Do remember the moment while discovering the wonders.
455	dian tai zuo xian chang guang bo.	Radio station is doing a live broadcast.
466	zhi li yu ke xue yan jiu de ren.	The person identified himself strongly with researching.
472	mei ge ren dou ying gai jiang yi ge gu shi.	Everyone should tell a story.
492	dui kang sai jiu zhe yang jie shu le	The battle is over.

the neutral speech can be used as a complete dataset of 400 utterances because there is no dominant word for neutral. The speech waveforms were recorded utterance by utterance, sampled at 16 kHz, digitized using 16-bit data and stored in single channel files. The speakers are denoted as lch, wzh, zqy and zzx.

In this work, we select the emotional speech from the spontaneous part which are picked from news articles, conversations and essays without emotional-rich words. The linguistic contents of these used utterances are shown in the Table 3.7. In order to have same number of utterances for each category for both male and female, we finally selected 50 utterances for each emotion, 200 chosen utterances in total. Table 3.8 illustrate the number of each category from males and females.

Differently from Fujitsu dataset and Berlin Emo-DB, acted emotions in CASIA database are not typical ones, all produced emotions in spontaneous part are picked from news articles, conversations and essays. Consequently, it may do not well enough to produce emotions in a clear and natural way. Further that the proposed system in this work is implemented by imitating the human response processing. Therefore, to have a standard baseline for emotion classification in mandarin case. The selected 200 utterances were firstly annotated again using the categorical method by 11 Chinese native speakers (5

Table 3.8: Initial numbers of each category for males and females from CASIA database

	Male		Female		
	wzh	zzx	lch	zqy	Total
Neutral	12	13	12	13	50
Happy	12	13	12	13	50
Angery	13	12	13	12	50
Sad	12	13	13	13	50
Total	49	51	50	50	200

Table 3.9: Numbers of each category for males and females from CASIA database after Human Categorical Label

	Male		Female		
	wzh	zzx	lch	zqy	Total
Neutral	17	23	10	18	68
Happy	8	1	12	9	30
Angery	13	13	13	11	50
Sad	11	14	14	11	50
Total	49	51	49	49	198

females and 6 males). Experimental results finally show that 68 utterances were recognized as neutral speech, 30, 50, and 50 utterances were grouped as happy, angery, and sad respectively. 2 spoken utterances can not be classified into any one of the above four emotional categories. Hence, 198 human-annotated utterances are used in our study ultimately. The human perception based category labeling results can be seen in Table 3.9.

3.2 Acoustic Features Extraction

In this research, as inputting parameters, acoustic features are a very crucial part to be studied. Hereafter, 21 related acoustic features successfully used in previous works were studied. 16 acoustic features are originally come from F0, power envelope, power spectrum, and duration were selected from the work by Huang ans Akagi [13]. Beyond these, 5 voice quality features investigated by Elabarougy [11] were also utilized. Duration related acoustic features are extracted by segmentation mannually, and the remained ones analyzed by STRAIGHT [27]. These used acoustic features can be grouped into five subgroups: 4 F0-related features, 4 Power envelop-related features, 5 power spectrum-related features, 3 duration-related features, and 5 voice quality-related features as shown in Table 3.10.

Table 3.10: Acoustic features groups

Group	Features
F0	F0 mean value of Rising Slope (F0 RS), F0 Highest Pitch (F0 HP), F0 Average Pitch (F0 AP), F0 Rising Slope of the 1st accentual phrase (F0 RS1)
Power envelope	mean value of Power Range in Accentual Phrase(PW RAP), Power Range (PW R), Rising Slope of the 1st accentual phrase (PW RS1), the Ratio between the average power in High frequency portion (over 3 kHz) and the Total average power (PW RHT)
Spectrum	1st Formant frequency (SP F1), 2nd Formant frequency (SP F2), 3rd Formant frequency (SP F3), Spectral Tilt (SP Ti), Spectral Balance (SP SB)
Duration	Total Length (DU TL), consonant length (DU CL), Ratio between Consonant length and Vowel length (DU RCV).
Voice quality	the mean value of the difference between the first harmonic and the second harmonic H1-H2 for vowel /a/,/e/,/i/,/o/, and /u/ per utterance, MH A, MH E, MH I, MH O, and MH U, respectively.

3.3 Experimental Evaluation of Semantic Primitives and Emotion Dimensions

Experiments on semantic primitive and emotion dimension is carried out to obtain the elements of the middle layer and the top layer of the proposed system. Emotion dimensions are the output parameters of the system, and semantic primitives are used as a bridge to map acoustic features to emotion dimensions. These adjectives can be utilized to recognize emotions of the speaker. These adjectives are: Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow, which originally from the work [24]. Actually the evaluation of semantic primitives and emotion dimensions of Japanese and German language have been done in [11]. Consequently, the Chinese part is needed to be furnished.

For the evaluation of CASIA dataset, human listening test is carried out by 10 Chinese native speakers (5 females and 5 males). Each emotional speech was evaluated 17 times by subjects, once for each semantic primitive. Subjects rated each of the 17 semantic primitives on a five-point scale: "1-Does not feel at all", "2-Seldom feels", "3-Feels a little", "4-feels", "5-Feels very much".

Additionally, the CASIA emotional database is annotated using the dimensional approach by doing human listening test with the same subjects in semantic primitives experiment. For evaluating emotion dimensions, a five-point scale -2, -1, 0, 1, 2 was utilized: valence (from -2 very negative to +2 very positive) and activation (from -2 very calm to +2 very excited). This experiments consists of two sessions, one for each emotion dimension. Subjects evaluated emotion dimension in valence and activation respectively for the whole database in two individual sessions.

Before starting the two experiments, the basic concept of semantic primitive and emotion dimension are explained to listeners. Hereafter they took a training part to listen to an example set contains 20 utterances, which over five-point scales. Such a training set aims at making listeners understand these adjectives. All stimulus were played randomly by binaural headphones with a comfortable sound pressure level in a soundproof room. Listening subjects evaluate their first perceived impression from the manner of speaking, nut not from the linguistic content, and then select a value on the five point scale for each adjective. Moreover, the inter-rater agreement was evaluated by pairwise Person's correlation between each two subjects's rating, for each semantic primitive respectively. Subject's rating with correlation value greater than 0.8 will be used. The average value of the subjects's scaling for each adjective and emotion dimension was calculated for each utterance.

3.4 Fuzzing Inference System

As so far, introduction of elements collections for three-layered model is completely done. The following key point is about estimating emotion dimensions. Estimation of emotion dimensions can be achieved by employing various of estimators, such as Support Vector Regression (SVR), K-nearest neighborhood (KNN), or Fuzzy Inference System FIS. In related previous work after [28] has indicated that FIS can give the best results and to be suited well for emotion dimensions estimation yielding small errors. This is due to that: first of all fuzzy logic is a human knowledge based mathematical model implemented by If-Then rules, and it is exactly what the model addresses to do in dealing with the speech affective perception. Secondly, fuzzy logic works well on non-linear functional modelling, and the relationship between emotion dimensions and acoustic features are certainly complex and non-linear. Hence, fuzzy logic is suitable to map these relationships. Additionally, fuzzy logic is derived from natural language, and the natural language used in this investigation is in the form of semantic primitives. On the whole FIS is an optimal connecting among human-based three-layered model.

3.5 Normalization

There is a obviously difference of values of acoustic features in comparison to affective speech and neutral ones, for instance, the vowel triangles form and their position of emotional speeches and neutral samples in F1-/F2-dimensional space are totally different.

Another issue that can be faced is the fact that vocal tracts are different among human beings, namely, formant frequency are emotion-dependent and speaker-dependent.

Extracted 21 acoustic features of Fujitsu, Berlin, and CASIA databases are normalized in order to avoid speaker-dependency and emotion-dependency on the acoustic features. Such an normalization method is after[11], in which all acoustic feature values are normalized by those of the neutral voice. It was performed by dividing the values of acoustic features by the mean value of neutral utterances for all acoustic features of all speakers.

Let $f = \{f_i\}(i = 1, 2, \ldots, K, \ldots, N)$ be a sequence of values of one acoustic feature, where N is the number of utterances in the used databases, and let the first K values of this sequence are calculated for neutral utterances, and the rest values calculated for the other emotional states. Then every element $\hat{f}_i$ in the the normalized acoustic feature $\hat{f} = \{\hat{f}_i\}(i = 1, 2, \ldots, K, \ldots, N)$ can be calculated by the following equation:

$$\hat{f}_i = \frac{f_i}{(\sum_{i=1}^{K} f_i/K)} \tag{3.1}$$

Additionally, since FIS is trained with (0,1) values, all training data from three emotional corpus are normalized to 0 to 1 after the above normalization.

Chapter 4

Multiple Speech Emotion Recognition System

4.1 System Implementation

The utilized elements in each layer for the proposed multiple language speech emotion recognition system are shown in Table 4.1 as following: valence dimension in the top layer, 17 semantic primitives in the middle layer and 21 acoustic features in the bottom layer. Emotion dimensions predicted with various of estimators, such as FIS, KNN, or SVM. As discussed in Chapter 3, most of statistical methodologies are mainly utilizing a linear and precise relationship to characterize the input and the output space, but the relationship among acoustic features, semantic primitives, and emotion dimensions are non-linearly. Hence, fuzzy logic is a more suitable mathematical approach to describing this nonlinear relationship.

To achieve this proposed model, a two folds approach is addressed to estimate emotion dimensions from acoustic features as following:

- **Step 1:** Estimation of Semantic Primitive: during this stage each semantic primitives are predicted from acoustic features. Each of semantic primitives is predicted from 21 acoustic features individually by different FIS; acoustic features are utilized as input parameters, and semantic primitives are output parameters.
- **Step 2:** Estimation of Emotion Dimensions: the values of predicted semantic primitives are used to estimate emotion dimension. One FIS system is needed for valence and activation respectively. The input of each FIS are the predicted semantic primitives and the output is the estimated emotion dimension.

In the stage of multinational emotions modelling, all three databases are used simultaneously with 10-fold cross validation.

4.2 Evaluation Measures

The ultimate purpose of this investigation is to achieve a multiple languages SER system, which can accurately predict emotion dimensions as well as human responses. With

Table 4.1: Elements of multilingual perceptual model for valence

Layer	Elements	Numbers
Top layer	Valence	1
Middle layer	Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow.	17
Bottom layer	F0_RS, F0_HP, F0_AP, F0_RS1, PW_RAP, PW_R, PW_RS1, PW_RHT, SP_F1, SP_F2, SP_F3, SP_Ti, SP_SB, DU_TL, DU_CL, DU_RCV, MH A, MH E, MH I, MH O, MH U.	21

such an assumption, we can make full use of the distinguished features, i.e. direction and degree in emotion dimensional space to classify multinational emotions. Since the output of proposed system are not directly emotional categories, but estimated values in dimensional space. Hence, estimated values from this model will be compared with human responses.

Most of previous researches advocated the use of Mean Absolute Error (MAE) to measure the performance of the emotion dimensions which is calculated by the predicted values of systems and the evaluated values by human subjects. The MAE is the most common means to evaluate the machine learning algorithms performance on emotion studies, as in this scenario.

To obtain the estimated emotion dimensions by restoring human processing, two subprocedures are carried out. The first stage is to evaluate semantic primitives from acoustic features, and the second stage is to predict emotion dimensions from estimated semantic primitives. Therefore, in order to compared with human responses, the MAE are analysed from the point of the view of semantic primitives and emotion dimensions respectively. The MAE is used to measure the distance between the estimated values by the proposed system and the annotated values by human beings. The smaller the MAE closer estimated values to the human responses.

In terms of each semantic primitive and emotion dimensions, MAE is analyzed as following: $\widehat{x} = \{\widehat{x}_i\}(i = 1, 2, \ldots, N)$ is the series of the predicted values of one semantic primitive or one emotion dimension from the proposed model, additionally, $x = \{x_i\}(i = 1, 2, \ldots, N)$ is the series of predicted values by human subjects for the corresponding semantic primitive or emotion dimension. Where N is the number of speeches in used database. Then, the mean absolute error MAE is calculated according to the following equation:

$$MAE = \frac{\sum_{i=1}^{N} \left| \widehat{x}_N^{(i)} - x_N^{(i)} \right|}{N} \tag{4.1}$$

Simultaneously, to obtain a comparative baseline for evaluating the estimating precision of proposed system, the evaluation of semantic primitives and emotion dimensions of

Japanese and Chinese corpus are evaluated three times by one Japanese native speaker and one Chinese native speaker respectively. With the limitation of German native speaker, the semantic primitives and emotion dimensions of Berlin Emo-DB are evaluated by another Japanese native speaker. Afterwards, the mean standard deviation (MSTDEV) of semantic primitives and emotion dimensions among human responses from subjects in the listening experiments is seemed as baseline, which calculated following from the following equation.

$$MSTDEV^{(j)} = \frac{\sum_{i=1}^{N} \sqrt{\frac{\sum_{m=1}^{N1} (x_m - \overline{\mu})^2}{N1}}}{N} \tag{4.2}$$

where $j \in \{Valence, Activation\}$ or $j \in \{Bright, Dark, High, \ldots, Slow\}$ and N have the same definition as Eq. 4.3, $N1$ is the number of times of listening tests ($N1$=3), and $\overline{\mu}$ is the average value of one utterance from per subject for 3 times.

4.3 System Evaluation

Our study aims at constructing a speech emotion recognition system (SER) which can work well for multiple languages. This section introduces the performance of estimated results for the proposed system. In order to explicitly study whether the proposed system accurately predict emotion dimensions as humans do. The obtained values of emotion dimensions from the proposed model will be compared with human responses in the dimensional space.

Additionally, for richly discussing and validating the availability and effectiveness of our philosophy. Based on the existing three speech emotional corpus, beyond the achieved multiple languages SER system, we construct six more emotion systems as shown in Table 4.2.

In summary, the performance of the proposed multiple languages SER system is compared with human responses firstly. In the following section, comparisons of multilingual SER, bilingual SER, and monolingual SER are carried out to discuss the effectiveness of proposed philosophy in this study. Implementations of these six emotion system will be introduced in Section 4.4.2.

4.3.1 Evaluation of Multilingual SER and Human Subjects

In the case of implementation of multilingual SER system, all used elements of three layers haven been described in Section 4.1. This system is trained and tested using three emotional corpus simultaneously with 10-fold cross validation regardless of linguistical/cultural impacts.

The distribution of output of Japanese, German, and Chinese from the achieved multilingual SER as well as the human responses are presented in scatter-plot of Valence-Activation in Figure 4.1, Figure 4.2, and Figure 4.3 respectively.There are two panels (a), (b) in each space; the left panel presents the distribution of human evaluation, and

Table 4.2: List of SER systems for bilingual and monolingual cases

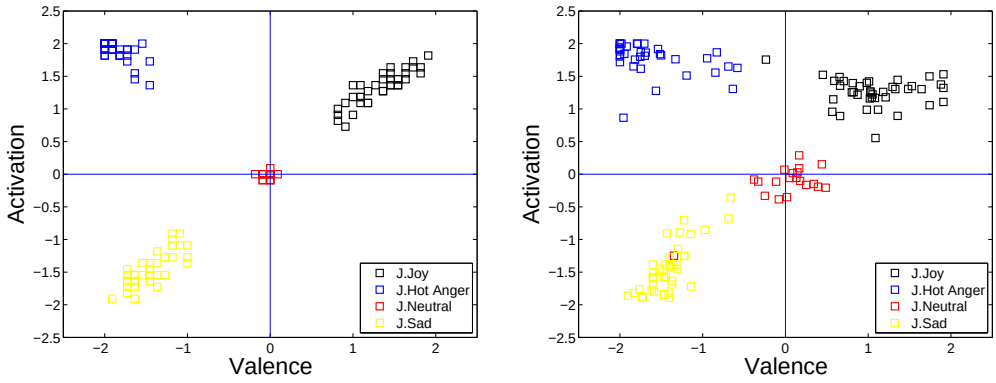
Language	Japanese	German	Chinese
Japanese	Monolingual SER for Japanese, shorten as Mono_J	Bilingual SER for Japanese and German, shorten as, Bi_JG	Bilingual SER for Japanese and Chinese, shorten as, Bi_JC
German		Monolingual SER for German, shorten as Mono_G	Bilingual SER for German and Chinese, shorten as Bi_GC
Chinese			Monolingual SER for Chinese, shorten as Mono_C

the right panel shows the estimation of the multilingual SER, The affective state of each speech is characterized by one point in the emotion dimensional space.

Given that the proposed multilingual SER is derived from the three-layered model. Therefore, the comparison of MAEs and MSTDEVs are addressed for the predicted semantic primitives and the estimated emotion dimensions, to imitate the human error for estimating semantic primitives and emotion dimensions, respectively. Figure 4.4, Figure 4.5, Figure 4.6 show the comparison of MAEs and MSTDEVs for all 17 semantic primitives corresponding to the Japanese, German, and Chinese cases. Seen from Figures 4.4, 4.5 and 4.6, we can conclude that all MAEs of 17 semantic primitives of three emotional corpus from the Multilingual SER are pretty nearly with MSTDEVs of those from listening experiments by human subjects. The maximum differences of MAEs and MSTDEVs, that beyond the the scope of MSTDEV of human responses are achieved by Strong with 0.18 in Japanese, Clear with 0.08 in German, and Clear with 0.19 in Chinese. But, considering that each semantic primitive can be evaluated from 1 to 5, by comparison, these differences are small and acceptable.

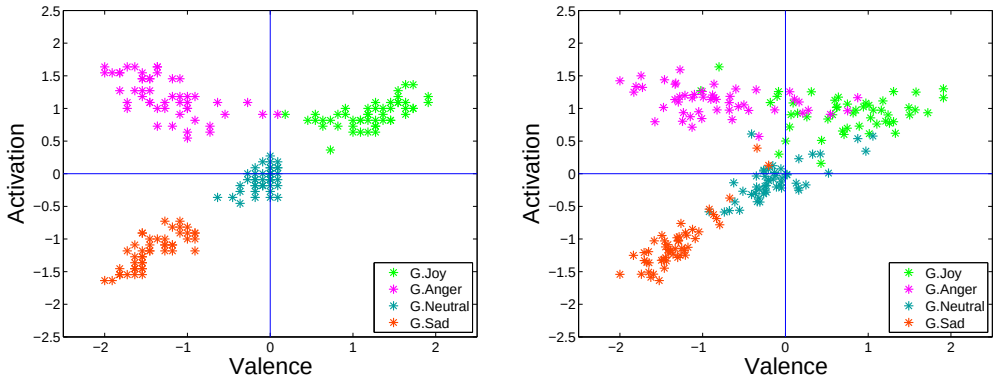
To evaluate the performance of this system, the comparison of the final output for human responses and SER system are followed. The results are shown in Figure 4.7, Figure 4.8, and Figure 4.9 for Japanese, German, and Chinese. MAEs of valence and activation of Japanese, German, and Chinese are: 0.30 and 0.22; 0.44 and 0.23; 0.46 and 0.31 respectively. The maximum differences of MAEs and MSTDEVs, that beyond the the scope of MSTDEV of human responses are all achieved by valence, with 0.13 in Japanese, 0.18 in German, and 0.18 in Chinese respectively. Seen from the MAEs of emotion dimensions of three datasets, the MAEs comes greater while dealing with multiple speakers imitating. However, such differences in valence are all below 0.2, which are also small and insignificant compared with the scale of emotion dimensions (from -2 to 2).

So far, these evaluation of semantic primitives and emotion dimensions indicate that estimated values of SP and ED from proposed Multilingual SER are very closed to human responses. Consequently, based on the accurate predicted emotion dimensions, we can effectively apply the knowledge of commonalities and differences on human emotion perception among languages to classify the multinational emotions in Section 4.5.



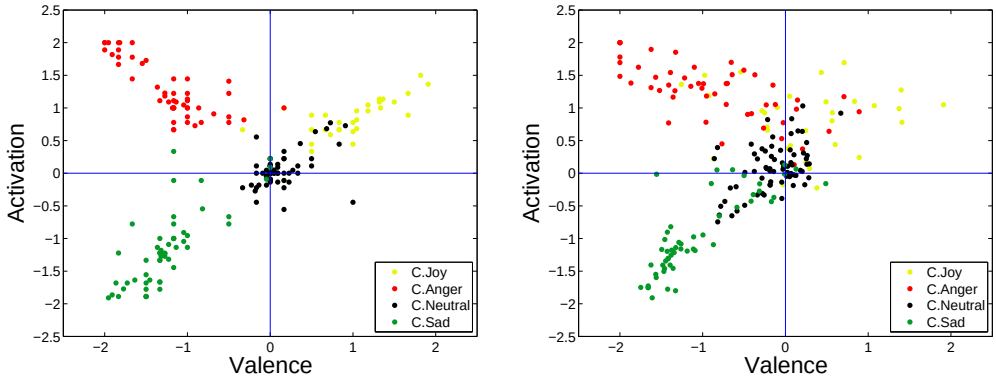
(a) Manually labeled by Human. (b) Estimated using Multilingual SER.

Figure 4.1: Scatter Plot of Estimated Positions of Japanese database in V-A space.



(a) Manually labeled by Human. (b) Estimated using Multilingual SER.

Figure 4.2: Scatter Plot of Estimated Positions of German database in V-A space.



(a) Manually labeled by Human. (b) Estimated using Multilingual SER.

Figure 4.3: Scatter Plot of Estimated Positions of Chinese database in V-A space.

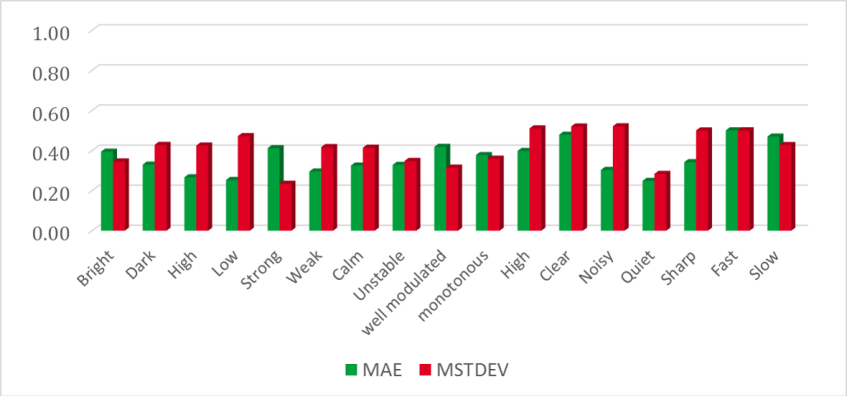


Figure 4.4: Comparison of MAE and MSTDEV for each Semantic Primitive (SP) in Japanese database

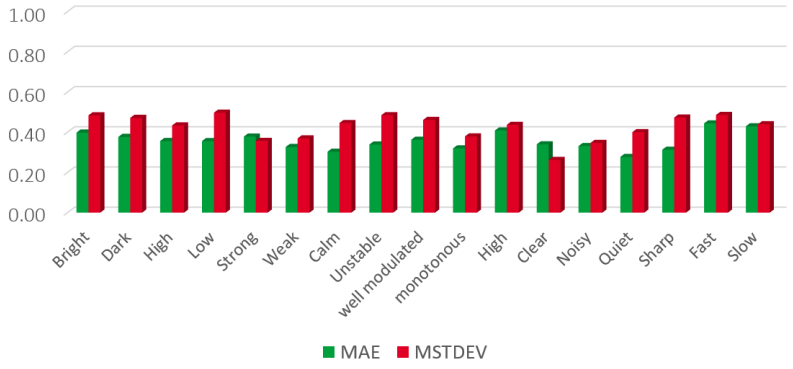


Figure 4.5: Comparison of MAE and MSTDEV for each Semantic Primitive (SP) in German database

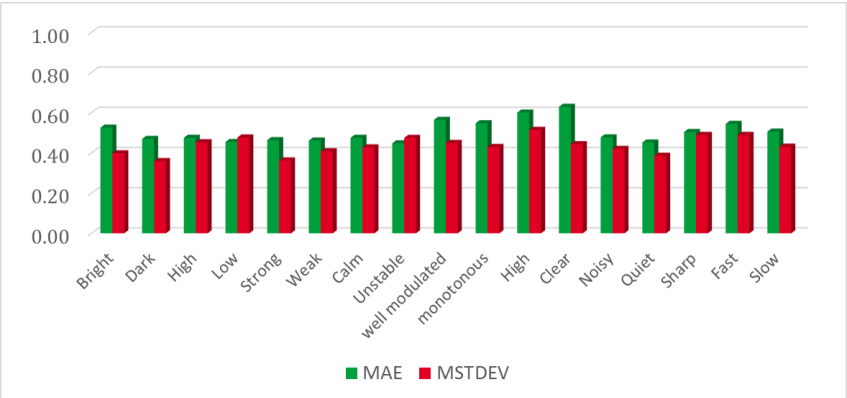


Figure 4.6: Comparison of MAE and MSTDEV for each Semantic Primitive (SP) in Chinese database

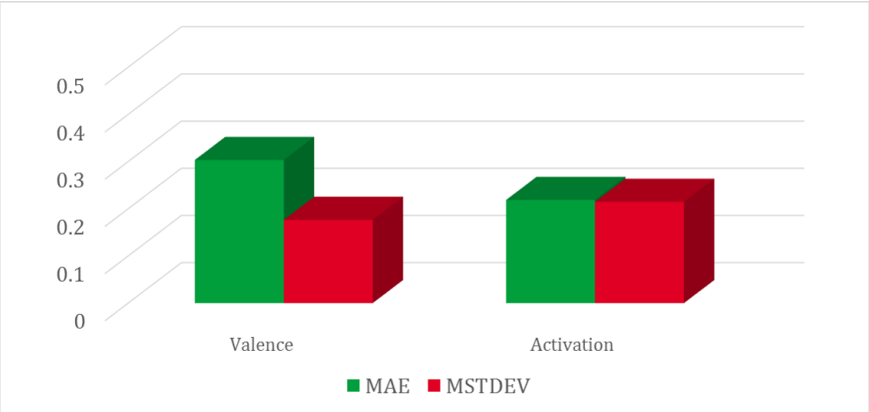


Figure 4.7: Comparison of MAE and MSTDEV for each Emotion Dimension (ED) in Japanese database

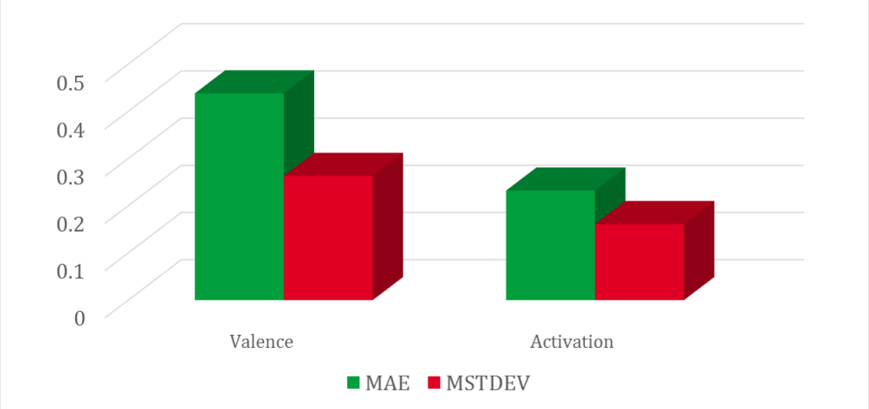


Figure 4.8: Comparison of MAE and MSTDEV for each Emotion Dimension (ED) in German database

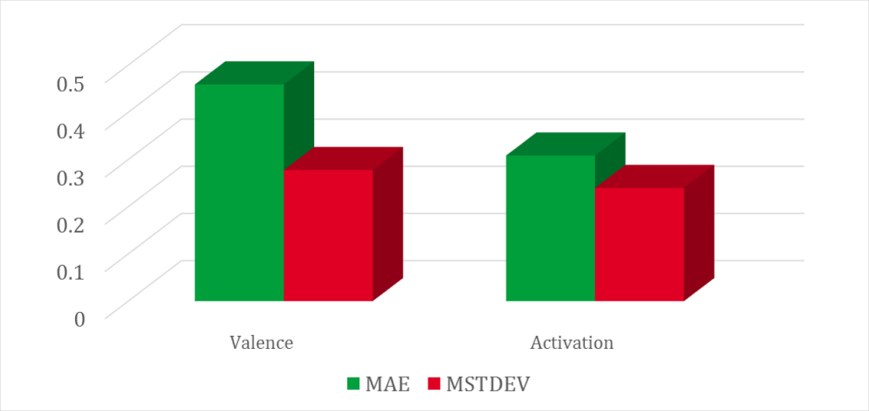


Figure 4.9: Comparison of MAE and MSTDEV for each Emotion Dimension (ED) in Chinese database

4.3.2 Evaluation of Proposed Philosophy

This subsection aims at investigating whether the proposed philosophy in this study can be effectively utilized to restore the human emotion perceptual processing regardless of cultures. Beyond the multilingual SER, six more emotion systems are implemented for a comparative analysis of the performance of the proposed multilingual SER in the view of bilingual and monolingual scenarios.

With the correlation based correlated feature selection method, the selected elements of the three layered model of Mono_J, Mono_G, Mono_C, Bi_JG, Bi_JC, and Bi_GC are listed in the following Tables. Sharing similar way to model multinational emotions, all these six emotion systems are trained and tested regardless of cultures using FIS. The estimated results of these systems will be presented in this section as well. To have a comparative analysis of our proposed philosophy, consequently, comparison with related works on emotion recognition using the same three corpus will be discussed after classification in Section 4.5.

Table 4.3: Mono_J: Each of 17 Semantic Primitives is estimated from 10 acoustic features firstly, followed by, the emotion dimension valence (activation) is predicted from 17 estimated Semantic Primitives.

Layer	Elements	Numbers
Top layer	Valence/Activation	1
Middle layer	Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow	17
Bottom layer	MH A, MH E, MH_O, F0_RS, F0_HP, PW_R, PW_RHT, DU_TL, DU_CL, DU_RCV	10

Table 4.4: Mono_G: All semantic primitives share the same 11 correlated acoustic features, exclude bright and heavy. Each of Semantic Primitives is estimated from correlated acoustic features firstly, followed by, the emotion dimension valence (activation) is predicted from 17 estimated Semantic Primitives.

Layer	Elements	Numbers
Top layer	Valence/Activation	1
Middle layer	Bright*, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy*, Clear, Noisy, Quiet, Sharp, Fast, and Slow	17
Bottom layer	MH A, MH E, MH_O, MH_U, F0_RS, F0_HP, PW_R, PW_RHT, PW_RAP, SP_F1, DU_TL	11
Especially, Bright* and Heavy* are estimated from 8 acoustic features: MH A, MH E, MH_O, MH_U, F0_RS, F0_HP, PW_R, SP_F1		

In the case of each monolingual SER, it obviously indicates that the best feature set is language dependent just as we mentioned previously. Compared relevant acoustic

Table 4.5: Mono_C: All 17 semantic primitives are of critical importance to valence and activation. Whereas, for each semantic primitive, the best optimal feature set is adjective-dependent. This table has elucidated 13 acoustic features, which are regarded as highly correlated features to semantic primitives. Those with circled symbols represent selected and used acoustic features, differently, acoustic features with horizontal lines are helpless ones. To obtain estimated emotion dimensions, first of all, each of the 17 semantic primitives in the middle of three layered model should be predicted separately from specified circled acoustic features using one FIS. Beyond that, the estimation of emotion dimension can be done from 17 estimated adjectives in the previous part with another FIS (○: used; −: not used)

Feature	MH_A	MH_I	MH_O	MH_U	F0_RS	F0_HP	PW_RHT	SP_F2	SP_F3	SP_TL	SP_SB	DU_TL	DU_CL
Bright	○	−	○	−	○	○	○	○	−	−	○	○	○
Dark	○	−	○	−	○	○	○	○	−	−	○	○	○
High	○	−	○	−	○	○	○	○	−	○	○	○	○
Low	○	−	○	−	○	○	○	○	−	−	○	○	○
Strong	○	−	○	−	○	○	○	○	○	○	○	○	○
Weak	○	−	○	−	○	○	○	○	○	○	○	○	○
Calm	○	−	○	−	○	○	○	○	○	○	○	○	○
Unstable	○	○	○	○	○	○	○	○	○	○	○	○	○
Wellmodulated	○	○	○	○	○	○	○	○	○	○	−	○	○
Monotonous	−	○	−	○	○	○	○	−	○	○	−	○	○
Heavy	○	−	○	−	○	○	○	−	○	○	○	○	○
Clear	○	−	○	−	−	○	○	−	○	○	−	○	○
Noisy	○	−	○	−	○	○	○	○	○	○	○	○	○
Quiet	○	−	○	−	○	○	○	○	−	−	○	○	○
Sharp	○	−	○	−	○	○	○	−	○	○	○	○	○
Fast	○	−	○	−	○	○	○	−	−	○	○	○	○
Slow	○	−	○	−	○	○	○	−	−	−	○	○	○

features among three databases, we can find that all 17 semantic primitives share the same correlated acoustic features in the bottom layer in Japanese. As for German scenario, bright and heavy have dependent acoustic features differently from others. Moreover, in the view of Chinese corpus, the best acoustic feature set of each semantic primitive is adjective dependent. This is always related to the properties of different emotional corpus, such as cultural background, speaker dependent/independent, emotional types, etc. Different from the other two emotional datasets, the Chinese emotional speeches are not typical ones, they are picked up from news, articles, conversations and essays, additionally, as we know, Chinese emotions are often widely range like a roller coaster.

With the strategy of correlation based feature selection method, three bilingual SER systems can be easily achieved as presented in Table 4.6, Table 4.7, and Table 4.8 for

Table 4.6: Bi_JG: Each of 17 Semantic Primitives is estimated from 9 acoustic features firstly, followed by, the emotion dimension valence (activation) is predicted from 17 estimated Semantic Primitives [29].

Layer	Elements	Numbers
Top layer	Valence/Activation	1
Middle layer	Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow	17
Bottom layer	MH A, MH E, F0_RS, F0_HP, PW_R, PW_RHT, DU_TL, DU_CL, DU_RCV	9

Table 4.7: Bi_GC: As for the Bilingual (German and Chiense) SER, it is found that all 17 semantic primitives are significant to activation, and 10 correlated semantic primitives below the table are seemed as relevant ones to valence. In addition, the best acoustic feature set of each semantic primitive is adjective dependent, this table elucidated 15 acoustic features, which are regarded as highly correlated features to 7 semantic primitives listed here below. All other 10 semantic primitives are estimated with 21 acoustic features. To obtain estimated emotion dimensions, first of all, each of the 17 semantic primitives in the middle of three layered model should be predicted separately from specified circled acoustic features using one FIS. Beyond that, the estimation of emotion dimension can be done from 17 estimated adjectives in the previous part with another FIS (∘: used; −: not used)

Features	MH_A	MH_E	MH_O	MH_U	F0_RS	F0_HP	PW_R	PR_RHT	PW_RAP	SP_F1	SP_F3	SP_TL	DU_TL	DU_CL	DU_RCV
Bright	∘	∘	∘	∘	∘	∘	∘	∘	−	∘	−	−	∘	∘	∘
Calm	∘	∘	∘	∘	∘	∘	∘	∘	−	∘	−	−	∘	−	∘
Monotonous	∘	∘	∘	∘	∘	∘	∘	∘	∘	−	∘	∘	−	−	−
Quiet	∘	∘	∘	∘	∘	∘	∘	∘	−	∘	−	−	∘	∘	∘
Sharp	∘	∘	∘	∘	∘	∘	∘	∘	−	∘	−	−	∘	∘	∘
Fast	∘	∘	−	∘	∘	∘	∘	∘	−	∘	−	−	∘	∘	−
Slow	∘	∘	∘	∘	∘	∘	∘	−	−	∘	−	−	∘	∘	−
The other 10 semantic primitives are estimated using 21 acoustic features. Additionally, valence is predicted from 10 primitives(Bright, Dark, High, Low, Weak, Heavy, Clear, Quiet, Fast, Slow.)															

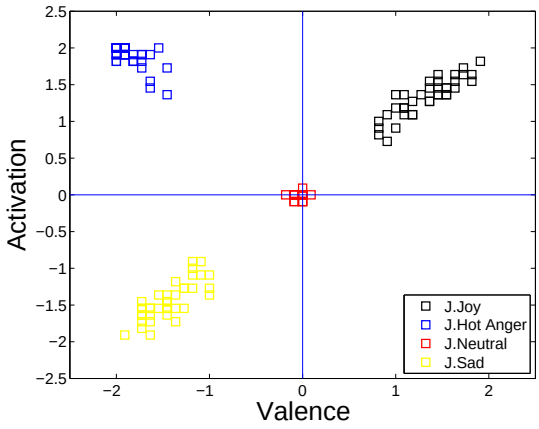
Japanese and German, German and Chinese, Japanese and Chinese, respectively. This means that our proposed feature selection method can be successfully used to select powerful features for languages regardless of cultures.

Table 4.8: Bi_JC: As for the Bilingual (Japanese and Chiense) SER, it is found that all 17 semantic primitives are significant to valence and activation. In addition, the best acoustic feature set of each semantic primitive is adjective dependent, this table elucidated 13 acoustic features, which are regarded as highly correlated features to 14 semantic primitives listed here below. All other 3 semantic primitives are estimated with 21 acoustic features. To obtain estimated emotion dimensions, first of all, each of the 17 semantic primitives in the middle of three layered model should be predicted separately from specified circled acoustic features using one FIS. Beyond that, the estimation of emotion dimension can be done from 17 estimated adjectives in the previous part with another FIS (○: used; −: not used)

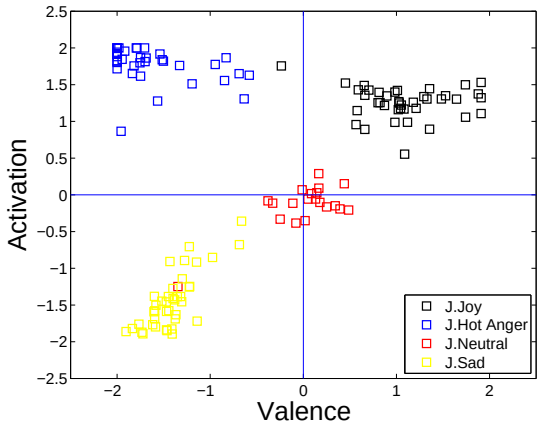
Features	MH_A	MH_I	MH_O	F0_RS	F0_HP	PW_R	PW_RHT	SP_F3	SP_TL	SP_SB	DU_TL	DU_CL	DU_RCV
Bright	−	−	○	○	○	−	○	−	−	−	○	○	−
Dark	○	−	○	○	○	○	−	−	−	−	○	○	○
High	○	○	○	○	○	○	−	−	−	−	○	○	○
Low	−	○	○	○	○	○	−	−	−	○	○	○	○
Strong	○	−	○	○	○	−	○	−	−	−	○	○	○
Weak	○	−	○	○	○	−	○	−	−	−	○	○	○
Calm	○	−	○	○	○	−	○	−	−	−	○	○	○
Unstable	○	−	○	○	○	−	○	−	−	−	○	○	○
Well modulated	○	−	○	○	○	○	○	−	○	−	−	○	○
Monotonous	○	−	○	○	○	○	○	−	○	−	−	○	−
Clear	−	○	○	○	○	○	○	○	○	○	○	○	○
Noisy	○	−	○	○	○	○	○	−	−	−	○	○	○
Quiet	○	−	○	○	○	−	○	−	−	○	○	○	○
Sharp	○	−	○	○	○	○	○	−	−	−	○	○	○
Semantic Primitives: Heavy, Fast, and Slow are estimated using 21 acoustic features.													

Hereafter, the comparative analysis among multilingual SER, bilingual SER, and Monolingual SER will be addressed.

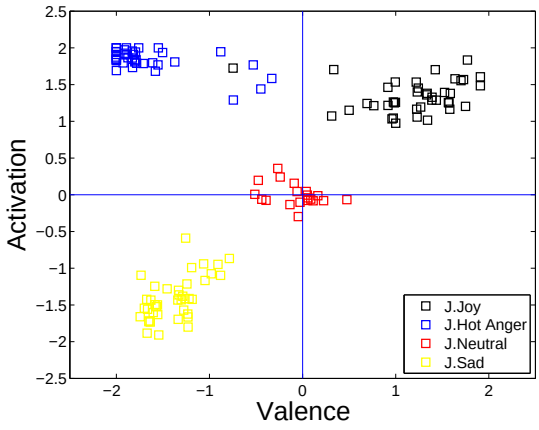
- **Distributions of Japanese emotional corpus**



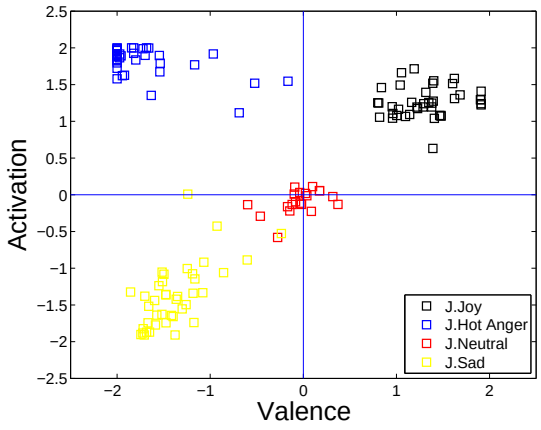
(a) Manually labeled by Human.



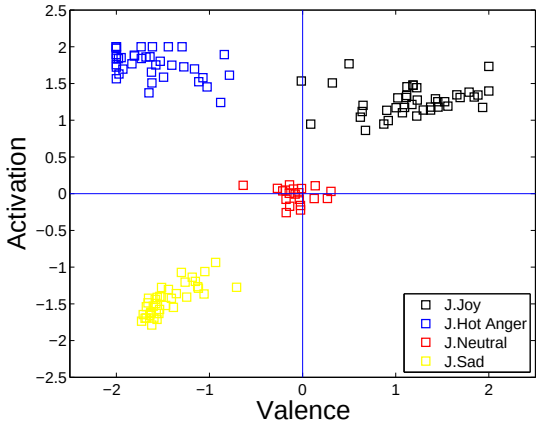
(b) Japanese from Multilingual SER.



(c) Japanese from Mono_J.



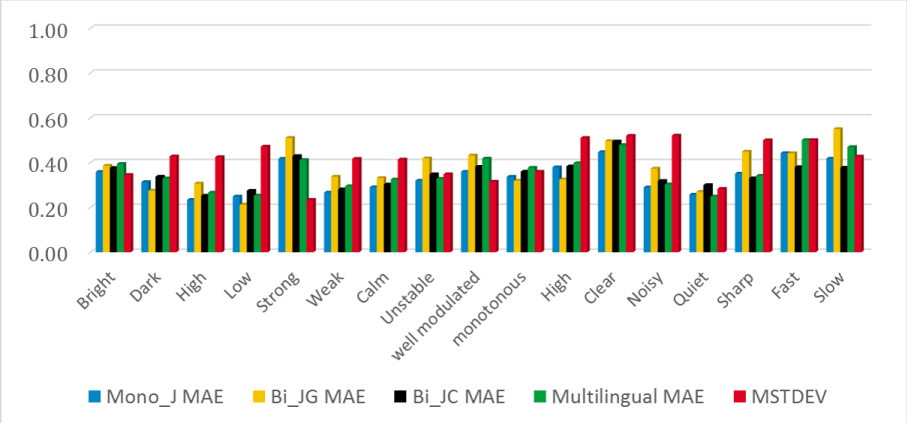
(d) Japanese from Bi_JC SER.



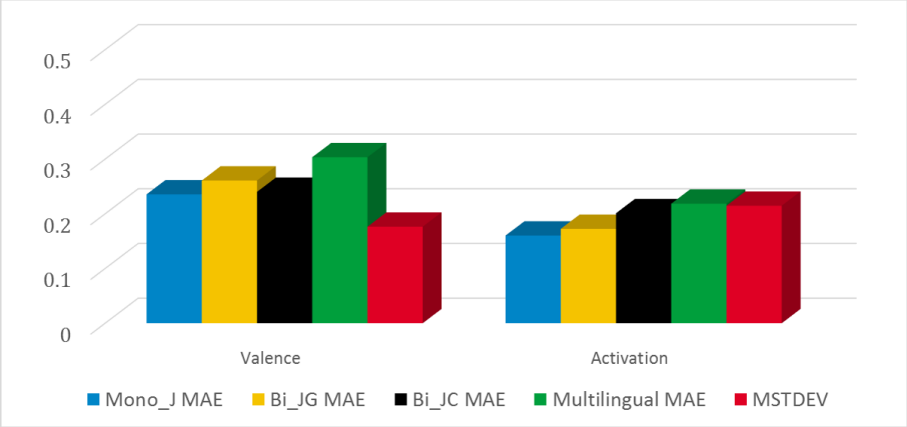
(e) Japanese from Bi_JG SER.

Figure 4.10: Scatter Plot of Estimated Positions of Japanese database in V-A space.

- **Comparative analysis of Japanese emotional corpus**



(a) Comparison of MAE and MSTDEV for each Semantic Primitive (SP) in Japanese database.



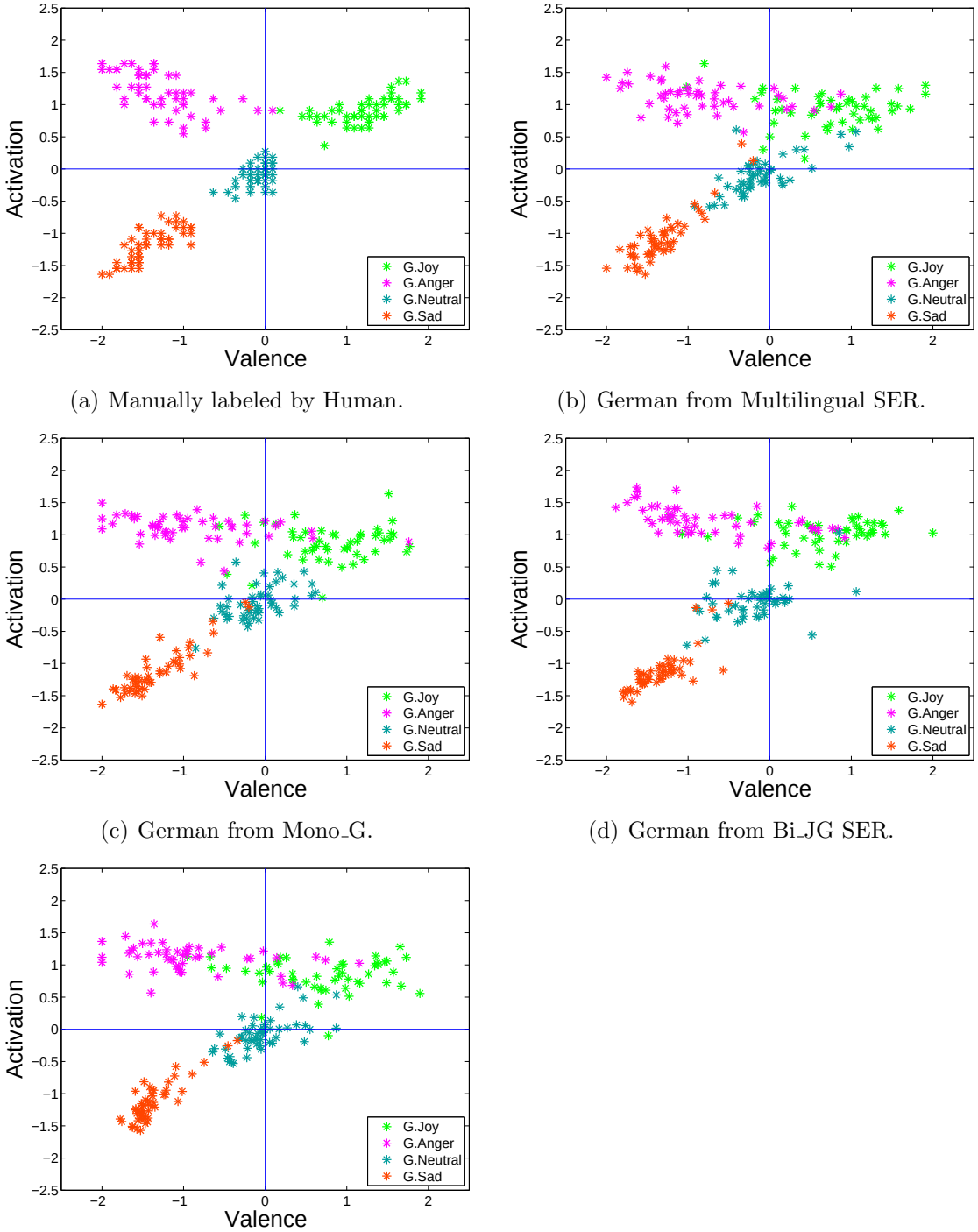
(b) Comparison of MAE and MSTDEV for each Emotion Dimension (ED) in Japanese database.

Figure 4.11: Comparative analysis of Japanese emotional corpus in Monolingual, Bilingual, and Multilingual.

Figure 4.11 (a) and (b) respectively show the comparisons of MAEs and MDTDEVs of Japanese corpus for semantic primitives and emotion dimensions in the terms of Mono_J, Bi_JG, Bi_JC, Multilingual SERs. The results of MAEs of Japanese semantic primitives/emotion dimensions from each SER indicate that there are insignificant differences among these systems. Moreover, it proves that our proposed multilingual SER trained using 3 databases has the ability to precisely estimated emotion dimensions with only a little bit higher error than the estimation using Monolingual or Bilingual SERs.

Therefore, we can conclude that our proposed philosophy of emotion dimensions estimation using three layered model effectively works regardless of cultures. The same conclusions can be obviously obtained for German and Chinese emotional corpus as illustrated in Figures 4.12 and 4.13, Figures 4.14 and 4.15.

- **Distributions of German emotional corpus**



(a) Manually labeled by Human.

(b) German from Multilingual SER.

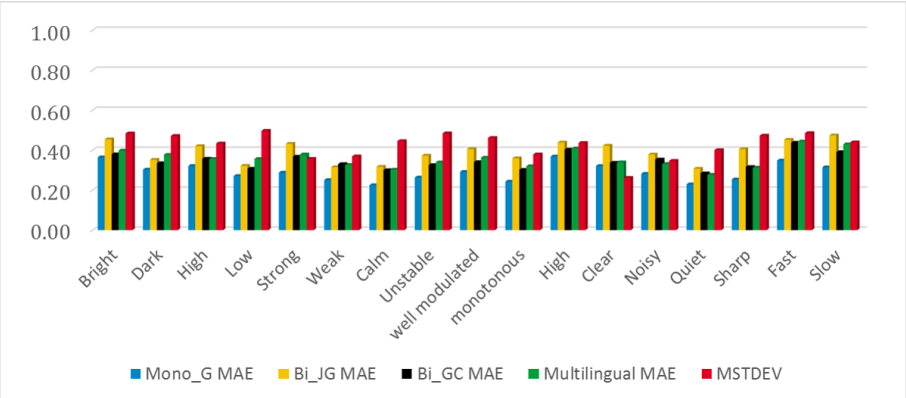
(c) German from Mono_G.

(d) German from Bi_JG SER.

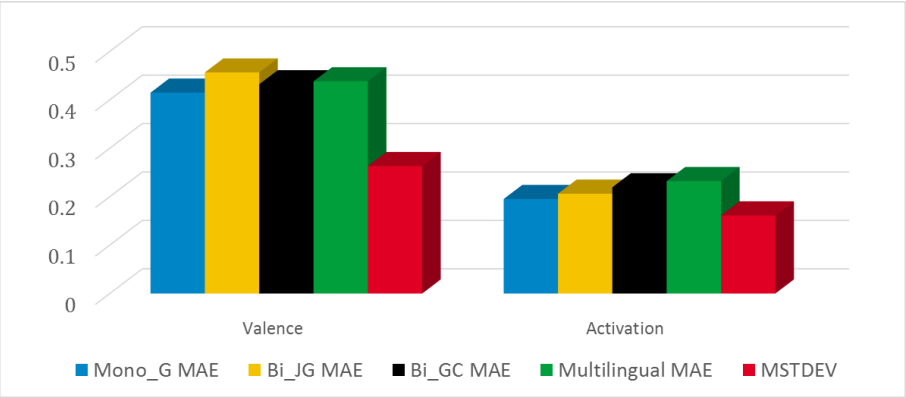
(e) German from Bi_GC SER.

Figure 4.12: Scatter Plot of Estimated Positions of German database in V-A space.

- **Comparative analysis of German emotional corpus**



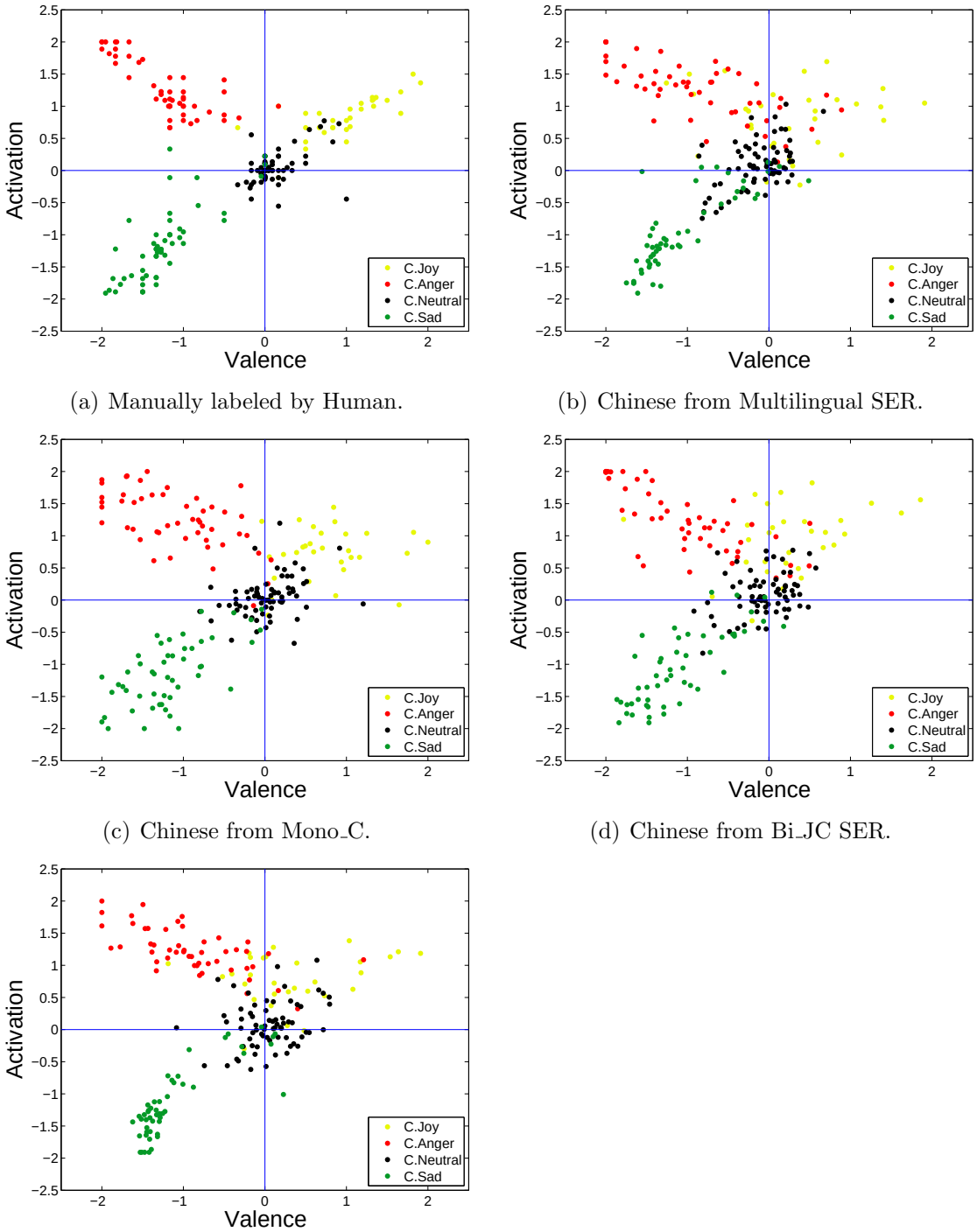
(a) Comparison of MAE and MSTDEV for each Semantic Primitive (SP) in German database.



(b) Comparison of MAE and MSTDEV for each Emotion Dimension (ED) in German database.

Figure 4.13: Comparative analysis of German emotional corpus in Monolingual, Bilingual, and Multilingual.

- **Distributions of Chinese emotional corpus**



(a) Manually labeled by Human.

(b) Chinese from Multilingual SER.

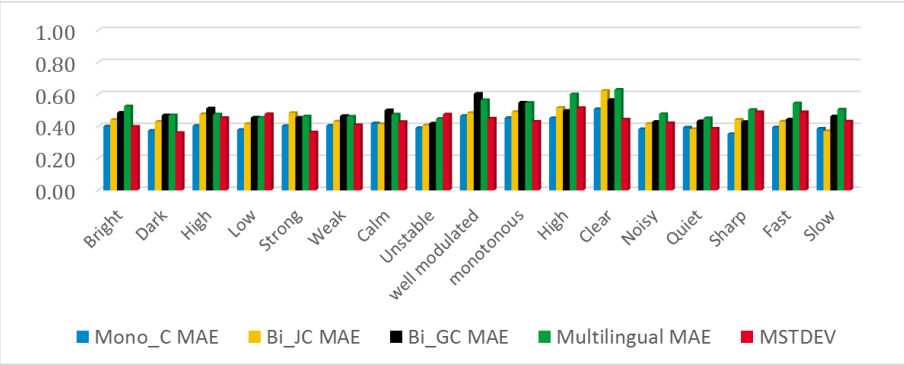
(c) Chinese from Mono_C.

(d) Chinese from Bi_JC SER.

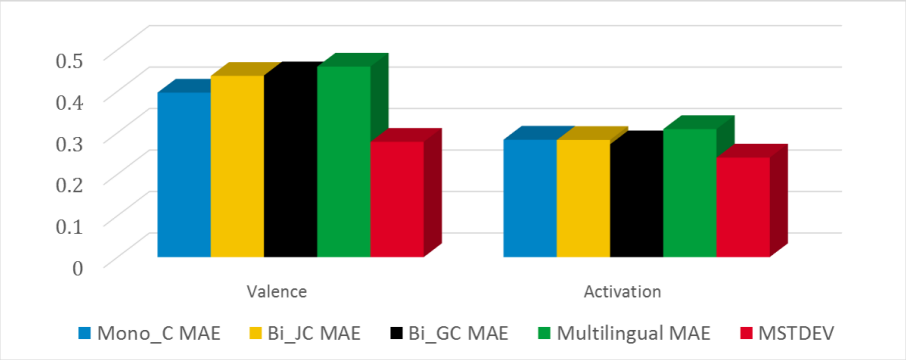
(e) Chinese from Bi_GC SER.

Figure 4.14: Scatter Plot of Estimated Positions of Chinese database in V-A space.

- **Comparative analysis of Chinese emotional corpus**



(a) Comparison of MAE and MSTDEV for each Semantic Primitive (SP) in Chinese database.



(b) Comparison of MAE and MSTDEV for each Emotion Dimension (ED) in Chinese database.

Figure 4.15: Comparative analysis of Chinese emotional corpus in Monolingual, Bilingual, and Multilingual.

4.4 Classification into Emotion Categories

This study aims at constructing a multilingual SER with the assumption that it possess the ability to precisely estimate emotion dimensions as humans do. To achieve this, we proposed three strategies from correlation based feature selection, over multinational emotion modelling, to multilingual emotion classification. Until now, the first two parts have been fully proved, obtained results from these two parts effectively claimed that, using proposed approaches, the expected system has been achieved with the assumption came true.

Hence, based on the knowledge of commonalities and differences on human emotion perception among languages introduced in Chapter 2. In this section, we will apply this new finding to classify multinational emotions. Commonalities and differences of human perception among multiple languages have been studied in V-A space by doing listening tests using subjects with different nations in [15]. It indicated that directions and distances from neutral to other emotions in the emotion dimensional space are similar among languages. Derived from the highly precision estimated values in dimensional space from the proposed three-layered model, in this study, a novel normalization method is presented to realistically imitate human emotion perception by adopting the direction and distance features to recognize emotional states for multiple languages. Firstly, the angle between the direction from neutral to emotional state and the horizontal directions towards positive valence are extracted using following Eq. 4.5.

$$angle = arctan(\frac{y_E - y_N}{x_E - x_N}) \tag{4.3}$$

where (x_E, y_E) is the center position of the emotional state E, and (x_N, y_N) is that of the neutral state N. Secondly, the degree from neutral to emotional state in dimensional space is computed by Euclidean-distance in Eq. 4.6.

$$d(E, N) = \sqrt{(x_E - x_N)^2 + (y_E - y_N)^2} \tag{4.4}$$

The definition of the parameters are the same as in Eq. 4.5.

Support Vector Machine (SVM) with 10-fold cross validation is used for classifying emotional categories by mapping direction and degree to emotional categories. In this part, the direction and degree were used as input to train the SVM classifier to classify three emotional corpus (Japanese, German, and Chinese) into four categories: Neutral, Joy, Anger, and Sadness. The Confusion matrix of the results is shown in Table 4.9 for Japanese, Table 4.10 for German, and Table 4.11 for Chinese.

In these tables, the numbers are the percentages of recognized utterances of the category. Table 4.9 respectively show the classification rate for Japanese from Bi_JC SER, Japanese from Bi_JG SER, Japanese from Multilingual SER, Japanese from Japanese SER, and one related study on emotion recognition using the same Japanese database after [31]. From this table, we can easily see that the highest average recognition rates of Japanese can be achieved by using Monolingual Japanese SER. Moreover, we attained 58.4% noisy reduction rate of emotional classification of related work after Elbarougy on Japanese

SER [31] using the monolingual Japanese system, which means our proposed philosophy can effectively work on speech emotion recognition. Additionally, the highest performance in German and Chinese are achieved by using Monolingual SER as well. And 27.5% and 23% noisy reduction rate are respectively obtained of emotional classification of previous study after Schuller [32] in German and after Zhou [33] in Chinese.

Especially, another valuable conclusion can also be summarized from the comparisons of recognition rates from the Bi_JC SER, Bi_JG SER, Multilingual SER, and Multilingual SER, it is obviously that the overall recognition rates from these systems are pretty nearly (same to German and Chinese Corpus), which has indirectly indicated that the estimated emotion dimensions from monolingual, bilingual, and multilingual SERs are almost the same. It further proves that our proposed system can work well on multinational emotion dimensions estimation regardless of cultures.

Table 4.9: Recognition Rates (R.R.) for Japanese (JPN) database

R.R.	JPN from Bi_JC	JPN from Bi_JG	JPN from Multilingual	JPN from Monolingual	Reda Elbarougy (Acoust. Sci. & Tech., 2014[31])
Neutral	100.00%	95.00%	95.00%	100.00%	80.00%
Joy	98.00%	95.00%	93.00%	97.50%	97.50%
Anger	90.00%	95.00%	95.00%	95.00%	92.50%
Sad	95.00%	100.00%	100.00%	95.00%	100.00%
Average	***95.75%***	***96.25%***	***95.75%***	***96.88%***	***92.50%***

Table 4.10: Recognition Rates (R.R.) for German(GER) database

R.R.	GER from Bi_GC	GER from Bi_GJ	GER from Multilingual	GER from Monolingual	Bjorn Schuller [32] (Speech Prosody, 2006)
Neutral	96.00%	92.00%	90.00%	98.00%	88.50%
Joy	86.00%	84.00%	82.00%	86.00%	69.00%
Anger	78.00%	72.00%	82.00%	86.00%	93.70%
Sad	100.00%	90.00%	96.00%	90.00%	94.30%
Average	***90.00%***	***84.50%***	***87.50%***	***90.00%***	***86.38%***

Table 4.11: Recognition Rates (R.R.) for Chinese (CHN) database

R.R.	CHN from Bi_JC	CHN from Bi_GC	CHN from Multilingual	CHN from Monolingual	Yu Zhou (IEICE, 2010)[33]
Neutral	93.00%	87.00%	89.00%	96.00%	98.00%
Joy	70.00%	67.00%	77.00%	97.00%	82.25%
Anger	80.00%	82.00%	73.00%	88.00%	90.25%
Sad	90.00%	84.00%	92.00%	92.00%	94.50%
Average	***83.25%***	***80.00%***	***82.75%***	***93.25%***	***91.25%***

Table 4.12: Emotion classification of three corpus without training

R.R.	JPN from Bi_GC	GER from Bi_JC	CHN from JG
Neutral	65.00%	28%	97.05%
Joy	47.50%	42%	13.33%
Anger	92.50%	52%	50.00%
Sad	55.00%	84%	72.00%
Average	***65.00%***	***51.50%***	***58.10%***

- **Discussion**

Actually, all expected results has been furnished by using the proposed multilingual speech emotion recognition system. Additionally, in the view of acoustic features discussion, based on the existing bilingual speech emotion recognition systems, we will try to recognize Chinese emotions from Bi_JG, classify Japanese emotions from Bi_GC, and evaluated German emotion from Bi_JC. Classification results without training can be seen from Table 4.12. The overall classification rate of Japanese, German, and Chinese were 65%, 51.5%, and 58.1% respectively, which has greatly decreased compared with that from SER with training. And it particularly show that the optimal feature set are always languages dependent.

Chapter 5

Summary and Future Work

5.1 Summary

This study is motivated by the long-term purpose to achieve a speech emotion recognition system that possess the ability to precisely estimate emotion dimensions for multinational languages. In most of previous speech emotion recognition studies, it greatly show that implemented strategies can work well for each monolingual language only. As for constructing a generalized SER system for multiple languages is always a challenging topic. This is due to the lack of powerful features to estimate emotion dimensions in multilingual scenario, and also the difficulty in designing model cross cultures, etc.

Our investigation is achieved from the view of mimicking human emotion perceptual processing among languages by using a three layered model with the assumption that the proposed SER possess the ability to estimate emotion dimensions as humans do. Here below are the procedures of constructing the proposed multiple languages speech emotion recognition system.

The elements of three layered model in three languages (Japanese, German, Chinese) are collected in Chapter 3, from acoustic features, over semantic primitives, to emotion dimensions. As input parameters, 21 acoustic features related to F0, power envelop, power spectrum, duration, and voice quality are extracted using STRAIGHT and manually segmentation. Semantic primitives are the elements in the middle layer which are used for describing emotional voice. Additionally, emotion dimensions are utilized to present emotions in this study. Two experiments are carried out for collection of semantic primitives and emotion dimensions respectively.

Afterwards, to precisely estimate emotion dimensions, three layered model is studied, and human knowledge based Fuzzy Inference System is applied to connect three layers. The semantic primitives are predicted using FISs firstly from acoustic feature in the bottom layer, and then the emotion dimensions can be predicted from the estimated semantic primitives. In this stage, our proposed system will be validated by training and testing the system using three languages in Japanese, German, and Chinese simultaneously to model multinational emotions as introduced in Section 4.2.

And then, with obtain emotion dimensions, the multilingual SER was evaluated how

closed the estimated results from system to the evaluation from human subjects by listening tests with Mean absolute error. To furnish the evaluation, the MAE have been calculated both for semantic primitives and emotion dimensions. Especially, to fully discuss the effectiveness and powerful of our proposed philosophy, using the same strategy and existing databases, we have constructed six more emotion systems in bilingual and monolingual cases. These systems helps to comparative analysis. The comparisons among these systems concluded that proposed strategies are effectively work for estimating emotion dimensions regardless of cultures.

Subsequently, motivated by the commonalities and differences among multilingual emotional perceptions, we extracted direction and degree in the dimensional space as distinguished features to classify different emotional states cross cultures as addressed in Section 4.5.

From the classification results, we can see that the overall recognition rates from these monolingual, bilingual, and multilingual SERs are pretty nearly, which has indirectly indicated that the estimated emotion dimensions from monolingual, bilingual, and multilingual SERs are almost the same. It further proves that our proposed system can work well on multinational emotion dimensions estimation regardless of cultures. Particularly, compared with related works on emotion recognition using the same emotional corpus [31] [32] [33], 58.4%, 27.5%, and 23% noisy reduction rate are respectively obtained in Japanese, German, and Chinese while using the Monolingual SER following our proposed methods, which means our proposed philosophy can effectively work on speech emotion recognition.

Finally, to discuss the acoustic features strength, we have evaluated each emotional database cross languages, as recognizing Chinese emotions from Bi_JG, classify Japanese emotions from Bi_GC, and evaluated German emotion from Bi_JC without training. Obtained results indicated again that the best feature sets are always languages dependent.

With such investigation on emotion dimensions estimation in multiple languages, we have achieved a multinational emotion recognition system. Especially, the proposed methods take into account human perception can help us, to find the best acoustic feature set for emotion dimensions in multiple languages, and also improve the recognition rates, more meaningful to help us to construct a speech to speech translation system in multiple languages, which will promote cultural exchange activities to foster mutual understanding around the world.

5.2 Future Work

In this work, our proposed philosophy has been successfully proved. However, to achieve a real-life application on speech emotion recognition is still a challenging work. With the investigation, we can found that the number of acoustic features are increased while increasing the number of acoustic features. We need to investigate more effective acoustic features for distinguish various emotions beyond Big Six emotions.

Additionally, in our investigation, only the acoustic features are used as input parameters, However, linguistic contents of the spoken utterances is also an important part of

the conveyed emotion in our daily life. In the future, we will try to combine acoustic and linguistic information to improve the performance of speech emotion recognition, and apply it to the application in real life Speech to Speech translation system.

Bibliography

[1] Koolagudi, G. Shashidhar, and K. Sreenivasa Rao. "Emotion recognition from speech: a review." International journal of speech technology 15.2 (2012): 99-117.

[2] B. Schuller, G. Rigoll, and M. Lang. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture." in: Proceedings of the ICASSO 2004, Vol. 1, 2004, pp. 577-580.

[3] C. M. Lee, and S. S. Narayanan. "Toward detecting emotions in spoken dialogs." Speech and Audio Processing, IEEE Transactions on 13.2 (2005): 293-303.

[4] M. E. Ayadi, M. S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition 44.3 (2011): 572-587.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. "Emotion recognition in human-computer interaction." Signal Processing Magazine, IEEE, 18(1), 32-80.

[6] M. Schrder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. C. Gielen. " Acoustic correlates of emotion dimensions in view of speech synthesis. " In INTERSPEECH, pp. 87-90, 2001.

[7] J. A. Russell. "Core affect and the psychological construction of emotion." Psychological Review, 110:145172, 2003

[8] M. Valstar, and B. Schuller et.al. "AVEC 2013: the continuous audio/visual emotion and depression recognition challenge." in Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. ACM, pp. 310, 2013.

[9] D. Wu, T. Parsons, E. Mower, and S. Narayanan. "Speech emotion estimation in 3D space." Multimedia and Expo (ICME), IEEE International Conference on. IEEE, 2010.

[10] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan. "Primitives-based evaluation and estimation of emotions in speech." Speech Communication, 49(10), 787-800, 2007.

[11] R. Elbarougy, and M. Akagi. "Cross-lingual speech emotion recognition system based on a three-layer model for human perception." Signal and Information Processing Association Annual Summit and Conference (APSIPA), Kaohsiung, Taiwan, 2013.

[12] Hozjan, Vladimir, and Zdravko Kacic. "Context-independent multilingual emotion recognition from speech signals." International Journal of Speech Technology 6.3 (2003): 311-320.

[13] C.F. Huang, D. Erickson, and M.Akagi. "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners." Acoustic 2008, Pairs, pp. 2317-2322, 2008.

[14] R. Banse, and K.R. Scherer. "Acoustic profiles in vocal emotion expression." Journal of personality and social psychology, 70(3), pp. 614-636, March (1996).

[15] X. Han, R. Elbarougy, M. Akagi, J. Li, T.D. Ngo, and T.D. Bui. "A study on perception of emotional states in multiple languages on valence and activation approach." Pro NCSP2015, Kuala Lumpur, Malaysia (2015).

[16] P. Ekman, and W. V. Friesen. "Unmasking the face: A guide to recognizing emotions from facial clues." Prentice Hall, New Jersey, 1975.

[17] K.R. Scherer. "On the nature and function of emotion: A component process approach." Approaches to emotion 2293 (1984): 317.

[18] P. Greasley, C. Sherrard, and M. Waterman. "Emotion in language and speech: Methodological issues in naturalistic approaches." Language and Speech 43.4 (2000): 355-375.

[19] J.A. Russell. "A circumplex model of affect." Journal of personality and social psychology 39.6 (1980): 1161.

[20] K.R. Scherer, A. Schorr, and T. Johnstone, eds. " Appraisal processes in emotion: Theory, methods, research." Oxford University Press, 2001.

[21] J.A. Russell. "Core affect and the psychological construction of emotion." Psychological Review, 110:145172

[22] K.R. Scherer. "Personality inference from voice quality: The loud voice of extroversion[J]." European Journal of Social Psychology, 1978, 8(4): 467-487.

[23] E. Brunswik. "Historical and thematic relations of psychology to other sciences." The Scientific Monthly 83 (1956): 151-161.

[24] C.F. Huang, and Masato Akagi. "A three-layered model for expressive speech perception." Speech Communication 50.10 (2008): 810-828.

[25] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss. "A database of German emotional speech," in: Proceedings of the Interspeech 2005, Lissabon, Portugal, 2005, pp.15171520.

[26] "Mandarin emotional speech corpus," http://www.chineseldc.org/doc/CLDC-SPC-2005-010/intro.htm, 2005. Institute of Automation, Chinese Academy of Sciences.

[27] H. Kawahara. STRAIGHT, Exploration of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds, Acoust. Sci and Tech., 27(6), pp. 349-353 (2006).

[28] R. Elbarougy, and M. Akagi. Comparison of Methods for Emotion Dimensions Estimation in Speech Using a Three-Layered Model, Proc. IEICE technical report. Speech, (June 2012).

[29] X. LI, and M. Akagi, "Toward Improving Estimated Accuracy of Emotion Dimensions in Bilingual Scenario based on Three-layered Model," "Oriental COCOSDA/CASLRE," Paper85, Shanghai, China, 2015.

[30] X. LI, and M. Akagi, "Automatic Speech Emotion Recognition in Chinese Using a Three-layered Model in Dimensional Approach," Pro NCSP2016, Hawaii, USA (2016).

[31] R. Elbarougy, and M. Akagi. "Improving speech emotion dimensions estimation using a three-layer model for human perception," Acoustical Science and Technology, 2014.

[32] B. Schuller, et.al. "Emotion recognition in the noise applying large acoustic feature sets." Speech Prosody, Dresden (2006): 276-289.

[33] Y. Zhou, J. Li, Y. Sun, J. Zhang, Y. Yan, and M. Akagi "A hybrid speech emotion recognition system based on spectral and prosodic features." IEICE Transactions on Information and Systems 93.10 (2010): 2813–2821.

Publications

International Conferences

[1] Li, X. and Akagi, M. "Toward improving estimation accuracy of emotion dimensions in bilingual scenario based on three-layered model," *Proceedings of International Conference (O-COCOSDA/CASLRE)*, IEEE, pp. 21-26, Shanghai, China, September, 2015.

[2] Li, X. and Akagi, M. "Automatic Speech Emotion Recognition in Chinese Using a Three-layered Model in Dimensional Approach," *Proceedings of International Conference (NCSP'16)*, Honolulu, Hawaii, March, 2016.

Domestic Conferences

[3] Li, X. and Akagi, M. "Study on Estimation of Bilingual Speech Emotion Dimensions using a Three-layered Model," *Proceedings of the Autumn Meeting of the ASJ*, Aizu, Fukusima, September, 2015.

[4] Li, X. and Akagi, M. "Improving Estimation Accuracy of Dimensions Values for Speech Emotion in Bilingual Cases Using a Three-layered Model," *Proceedings of The Taiwan/Japan Joint Research Meeting on Psychological & Physiological Acoustics and Electroacoustics*,Taiwan, October, 2015.