A Study on Constructing an Automatic Speech Emotion Recognition System based on a Three-Layer Model for Human Perception

by

Reda El-Said Mohamed El-Sayed El-Barougy

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Professor Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology

September, 2013

Abstract

The voice is an extraordinary human instrument. Every time we speak, our voice reveals our gender, age, culture background, level of education, native birth, emotional state, and our relationship with the person spoken to. All these clues are contained in even small speech segment, and other people can read our voices with remarkable accuracy. When we speak, we "encode" important information about ourselves; when we listen to others, we can "decode" important information about them. One of the goals of human-computer interaction (HCI) is the improvement of the user experience, trying to make this interaction closer to human-human communication. Inclusion of speech emotion recognition was one of the key points to include "perception" to multi-media devices. This improved their user interfaces. However, the analysis of emotional states by the study of the implicit channel of communication (i.e. the recognition of not only what is said but also how it is said) may improve HCI making these applications more usable and friendly.

We can communicate using speech from which various information can be perceived. Emotion is an especial element that does not depend on the content of the utterance and is useful in communications that reflects the speaker's intention. Most previous techniques for automatic speech emotion recognition focus only on the classification of emotional states as discrete categories such as happy, sad, angry, fearful, surprised, and disgusted. However, emotions are usually gradually change from weak to high degree. Therefore, an automatic speech emotion recognition system should be able to detect the degree or the level of the emotional state form the voice. Hence, in this study we adopt the dimensional descriptions of human emotion, where emotional states are estimated as a point in a three-dimensional space. These dimensions are suitable for representing the gradient nature of emotional state.

This research is concerned with the automatic speech emotion recognition system based on the dimensional model. In this model, human emotional state is represented as a point in a space consists of three dimensions: valence, activation, and dominance. Valence is used to describe emotion in terms of positive and negative assessments (e.g. happy and encouraging have positive-valence whereas angry and sad have negative-valence). Activation is used to define emotion in terms of arousal or excitation (e.g. happy and angry have positive-activation while sad and bored have negative-activation). The dominance dimension indicating the degree of weakness or strength of an expression, this dimension used to distinguish between the close neighborhood of anger and fear in the valence-activation space. The input for the automatic system are acoustic features extracted from speech signal and the output are the estimated values of valence, activation, and dominance. These estimated values for the three dimensions not only identify the emotional state but also the degree of the emotional state such as "low happy", "happy", "very happy".

Conventional speech emotion recognition methods using the dimensional approach are mainly focused on investigating the relationship between acoustic features and emotion dimensions as a two-layer model, i.e. acoustic feature layer and emotion dimension layer. However, using this model has the following problems: (i) we do not know what acoustic features are related to each emotion dimension (ii) the acoustic features that correlate to the valence dimension are less numerous, less strong, and more inconsistent, and (iii) the values of emotion dimensions are difficult to estimate precisely only on the basis of acoustic information. Due to these limitations, values of the valence dimension have been particularly difficult to predict by using the acoustic features directly.

The ultimate goal of our work is to improve the conventional dimensional method in order to precisely predict values of the valence dimension as well as improve prediction of those of the activation and dominance. To achieve this goal, we construct an automatic speech emotion recognition system by adopting a three-layer model for human perception described by Scherer (Scherer, 1978) and developed by Huang and Akagi (Huang and Akagi, 2008). It was assumed that, a listener perceives the acoustic features and internally represented them as a smaller perception e.g. adjectives describing emotional voice such

as Bright, Dark, Fast, and Slow. These smaller percepts or adjectives are finally used to judge the emotional state of the speaker.

In this thesis, the proposed idea to improve automatic speech emotion recognition system can be done by imitating the process of human perception for emotional state from the speech signal. The conventional two-layer model has limited ability to find the most relevant acoustic features for each emotion dimension, especially valence, or to improve the prediction of emotion dimensions from acoustic features. To overcome these limitations, this study proposes a three-layer model to improve the estimating values of emotion dimensions from acoustic features. Our proposed model consists of three layers: emotion dimensions (valence, activation, and dominance) constitute the top layer, semantic primitives the middle layer, and acoustic features the bottom layer. A semantic primitive layer is added between the two conventional layers acoustic features and emotion dimensions.

We first, assume that the acoustic features that are highly correlated with semantic primitives will have a large impact for predicting values of emotion dimensions, especially for valence. This assumption can guide the selection of new acoustic features with better discrimination in the most difficult dimension. The second assumption is that human can judge the expressive content of a voice even without the understanding of one language, such as emotional state of the speaker from different language. Using the second assumption, we investigate the universality of the proposed speech emotion recognition system to detect the emotional state cross-lingually.

To sum up, the aims of this work is to investigate the following assumptions: (1) Selecting acoustic features based on the proposed three-layer model of human perception will help us to find the most related acoustic features for each emotion dimensions. (2) Using these selected acoustic features, as inputs to an automatic emotion recognition system will improve the accuracy of all emotion dimensions especially valence. (3) In addition, we investigate whether there are acoustic features that allow us to estimate the emotional state from the voice of a person no matter what language he/she speaks. We are interesting to build a global automatic emotion recognition system, which have the

ability to detect the emotional state regardless of language.

Therefore, the method we adopt to construct our speech emotion recognition system includes the following steps: first, we proposed a new acoustic feature selection algorithm to select the most relevant acoustic features for each emotion dimension by using a top-down method. Then, we build a perceptual three-layer model for each emotion dimension using a top-down method, one emotion dimension in the top layer, the highly correlated semantic primitive to this dimension in the middle layer, in the bottom layer the highly correlated acoustic feature to the highly correlated semantic primitives in the middle layer. Finally, a button-up method was used to estimate values of emotion dimensions from acoustic features by firstly, using fuzzy inference system (FIS) to estimate the degree of each semantic primitive from acoustic features, and then using another FIS to estimate values of emotion dimension from the estimated degrees of semantic primitives.

The proposed emotion recognition system was validated using two different languages (Japanese and German) in two different cases (speaker-dependent and multi-speaker). Firstly, the system was implemented for each language individually to investigate whether the system can be applied for any language. Secondly, the common acoustic features between the two languages are used to validate the second assumption.

The experimental results reveal that by using the proposed features selection algorithm for the two databases, we found many related acoustic features for each emotion dimension. The estimation accuracy for emotion dimensions is improved using the selected features comparing with all features. Moreover, the three-layer model can be applied for the two-different language databases with similar performance. The most important result is that the proposed three-layered model outperforms the conventional two-layered model. The speaker-dependent vs. multi-speaker emotion estimation was tested; it was found that the performance of speaker-dependent is better that multi-speaker. Finally, the estimated values of emotion dimensions are mapped into the given emotion categories using a Gaussian Mixture Model classifier for the Japanese and German databases. For the Japanese database, an overall recognition rate was up to 94% using emotion dimensions. For the German database, the recognition rate was up to 95.5% for speaker-dependent

tasks.

In order to investigate whether the automatic system can detect the emotion dimensions for one language by training the system using different language. The proposed speech emotion recognition system was trained using Japanese language and tested using German language and vice versa. It was found that the cross-language emotion recognition system could estimate emotion dimensions with small error comparing the estimation results from a system trained using the native language.

The results indicated that the three-layer system shows an internal structure of human perception clearly and has the recognition accuracy better than that of the two-layer system. In a sense of imitating the perception mechanism of humans, the constructed system provides a more effective emotion recognition system compared with the conventional methods.

Acknowledgments

First of all, I thank God for his countless bounties bestowed upon me and ask Him to guide and grant me mercy and forgiveness in the afterlife.

During my graduate studies at Japan Advanced Institute of Science and Technology (JAIST), I have received generous help from each and every one of JAIST, without which this thesis would have never been finished. I would like to express my most sincere gratitude to my advisor, Prof. Masato Akagi, for his constant support and guidance and encouragement throughout my stay. I consider myself fortunate to be a student of Prof. Akagi, who inspired me with his enthusiasm in exploring new scientific frontiers and unique insight in automatic emotion speech recognition. His patience, instructions and kind encouragement that sustained me through failures, which lead me to learn a lot about how to construct system for speech emotion recognition. Also, I want to thank Associate Prof. Masashi Unoki for his guidance, advice, and his helpful comments on lab meeting during my study. I am also want to thank Assist. Prof. Ryota Miyauchi for his guide and help especially during traveling for outside meetings.

I would like to pay sincere thanks to Prof. Jianwu Dang, and Associate Prof. Isao Tokuda for his guidance for my sub-theme and their precious advices and suggestions. I would like to extend my thanks to all the past and present group members of Prof. Akagi and Unoki lab, especial thanks for Hamada-san for his touter ship in the first days in Japan and until now, I will never forgot your help Hamada-san.

Last but not least I'd like to express my gratitude to my wife Dalia, my daughters Rawan and Ranim who were born in Egypt, My sons, Zeyad and Muaz who were born here in Japan during my study, and my extended family whose love, support, and belief in me has never seized. They waited for me through day and night and stood by me especially my mother. I once again thank you all for everything you did for me. Finally I

gratefully acknowledge financial support of my Ph.D research scholarship by Ministry of Higher Education of Arab Republic of Egypt.

Contents

List of Figures

List of Tables

Acronyms

HCI	Human-Computer Interaction
FIS	Fuzzy Inference System
AF	Acoustic Features
MAE	Mean Absolute Error
GMM	Gaussian Mixture Model
kNN	K-Nearest Neighbor
ANFIS	Adaptive Neuro Fuzzy Inference Systems
GUI	Graphical User Interface
RMSE	Root Mean Square Error
ED	Emotion Dimensions
SVR	Support Vector Regression
NN	Neural Network
ASR	Automatic Speech Recognition
AP	Accentual Phrase
ANOVA	ANalysis Of VAriance
LOOCV	Leave-One-Out-Cross-Validation

Chapter 1

Introduction

1.1 Introduction

Speech can be seen as a two-channel mechanism, involving not only actual meaning of the communication but also several prosodic distinctions. The linguistic channel deals with the actual information inferred by words ("What is said") whereas the paralinguistic channel gives additional information about the speaker ("How it is said"), namely his/her emotional state. The linguistic channel was the main focus for research in the past, but scientists have recently become more and more interested in this second implicit channel [74]. One of the goals of human-computer interaction (HCI) is the improvement of the user experience, trying to make this interaction closer to human-human communication. Inclusion of speech emotion recognition was one of the key points to include "perception" to multimedia devices. This improved their user interfaces. However, the analysis of affective states by the study of the implicit channel of communication (i.e. the recognition of not only what is said but also how it is said) may improve HCI making these applications more usable and friendly. This is because, in general, inclusion of skills of emotional intelligence to machine intelligence makes HCI more similar to human-human interaction [59, 65]. In other words, it is an attempt to make a computer capable of observing, interpreting and generating emotional states [57].

The research of automatic speech emotion recognition, not only can promote the further development of computer technology, but also greatly enhance the efficiency of peoples work and study, and help people to solve their problems more efficiently, as well as further enrich our lives and improve the quality of life. Automatic emotion recognition from speech has in the last decade shifted from a side issue to a major topic in human computer interaction and speech processing [89]. However, emotion detection from speech is a relatively new field of research, it has many potential applications. Therefore, accurate detection of emotion from speech has clear benefits for the design of more natural human-machine speech interfaces or for the extraction of useful information from large quantities of speech data. In human-computer or human-human interaction systems, emotion recognition systems could provide users with improved services by being adaptive

to their emotions. It is also becoming more and more important in computer application fields such as health care, children education, etc [100].

1.2 Problem statement

Most previous techniques for automatic speech emotion recognition focus only on the classification of emotional states as discrete categories such as happy, sad, angry, fearful, surprised, and disgusted [64, 48]. However, a single label or any small number of discrete categories may not accurately reflect the complexity of the emotional states conveyed in everyday interaction [2]. In the real-life, an emotional state have different degree of intensity, and may change over time depending on the situation from low to high degree. Therefore, an automatic speech emotion recognition system should be able to detect the degree or the level of the emotional state form the voice [2]. Hence, a number of researchers advocate the use of dimensional descriptions of human emotion, where emotional states are estimated as a point in a multi-dimensional space (e.g., [93, 76]).

In this study, a three-dimensional continuous model is adopted in order to represent the emotional states using emotion dimensions i.e. valence, activation and dominance. This approach is chosen because it exhibits great potential to model the occurrence of emotions in real world as in a realistic scenario, emotions are not generated in a prototypical or pure modality, but rather than in complex emotional states, which are a mixture of emotions with varying degrees of intensity or expressiveness. Therefore, this approach allows a more flexible interpretation of emotional states [101].

However, although the conventional dimensional model for estimating emotions from speech signals allows the representation of the degree of emotional state, this model has the following problems: (i) we do not know what acoustic features are related to each emotion dimension (ii) the acoustic features that correlate to the valence dimension are less numerous, less strong, and more inconsistent [76], and (iii) the values of emotion dimensions are difficult to estimate precisely only on the basis of acoustic information [25]. Due to these limitations, values of the valence dimension have been particularly

difficult to predict by using the acoustic features directly.

Conventional speech emotion recognition methods are mainly based on investigating the relationship between acoustic features and emotion dimensions as a two-layer model, i.e. acoustic feature layer and emotion dimension layer. For instance, Grimm et al. attempted to estimate the emotion dimensions (valence, activation, and dominance) from the acoustic features by using a fuzzy inference system (FIS) [31]. However, they found that activation and dominance were more accurately estimated than valence. Furthermore, many researchers also tried to investigate the most related acoustic features for each emotion dimension by using the correlation between a set of acoustic features and emotion dimensions [25, 93, 76, 80]. In all these studies, the valence dimension was found to be the most difficult dimension. Thus, some other studies focused only on exploring acoustic features related to valence dimension [77, 6]. Some emotions were found to share similar acoustic features such as happiness and anger, which were characterized by increased levels of fundamental frequency (F0) and intensity. This is one reason acoustic discrimination on valence dimension is still problematic: no strong discriminative acoustic features are available to discriminate between positive speech (e.g. happiness) and negative speech (e.g. anger), however, these emotions are usually not hard to distinguish for humans [80]. Therefore, a number of researchers tried to discriminate between the positive and negative emotions by combining acoustic and linguistic features to improve the valence estimation [42, 80]. However, valence was found to still be poorly estimated. All these studies suggest that finding relevant acoustic features to discriminate in the valence domain is one of the main challenges in speech emotion recognition.

On the other hand, an interesting question to ask is whether emotional states can be recognized universally or not. Culture and society have a considerable weight on the expression of emotions. This, together with the inherent subjectivity among individuals, can make us wonder about the existence of universal emotions. If we consider Darwins theory of evolution, emotions find their root in biology and therefore can be to some extent considered as universals [3]. Several studies have indeed shown evidence for certain universal attributes for both speech [7, 46] and music [83, 60], not only among individuals

of the same culture, but also across cultures. Dang et al. 2009, for instance, performed an experiment in which humans had to distinguish between 3 and 6 emotions respectively [15]. Their conclusion was that listeners are able to perceive emotion from speech sound without linguistic information with about 60% accuracy in a three-emotion evaluation and about 50% in a six-emotion evaluation.

Several studies have worked on the analysis of the most important acoustic features from the point of view of categorical model, working on mono-lingual [97, 8] and multi-lingual [66] data. However, they have not yet studied with the same depth the importance of acoustic features from the dimensional model point of view.

1.3 Objective of the present research work

In this study; our focus is on improving the dimensional method in order to precisely estimate values of emotion dimensions especially valence dimension. The first question that we try to answer is which acoustic features are mostly relevant for describing the valence dimension? In other words, we investigate the speech acoustic features that have a large impact for the prediction of emotion dimensions, and propose and construct an automatic speech emotion recognition system that has the ability to accurately predict the emotional state of the speaker based on the dimensional model. The second question is: whether there are acoustic features that allow us to estimate the emotional state from the voice of a person no matter what language he/she speaks? Even without the understanding of one language, human can still judge the expressive content of a voice, such as emotions. Therefore, we also investigate the universality of automatic speech emotion recognition, by investigate whether an automatic emotion recognition system trained using one language has the ability to detect the emotion dimension from different languages.

1.4 Proposed approach

Most of the previous studies used the two-layer model to investigate the relationship between acoustic features and emotion dimensions, however this model does not imitate human perception. Human perception, as described by Scherer [79] who adopted a version of Brunswik's lens model originally proposed in 1956 [10], is a three-layer model as shown in Figure 1.1.

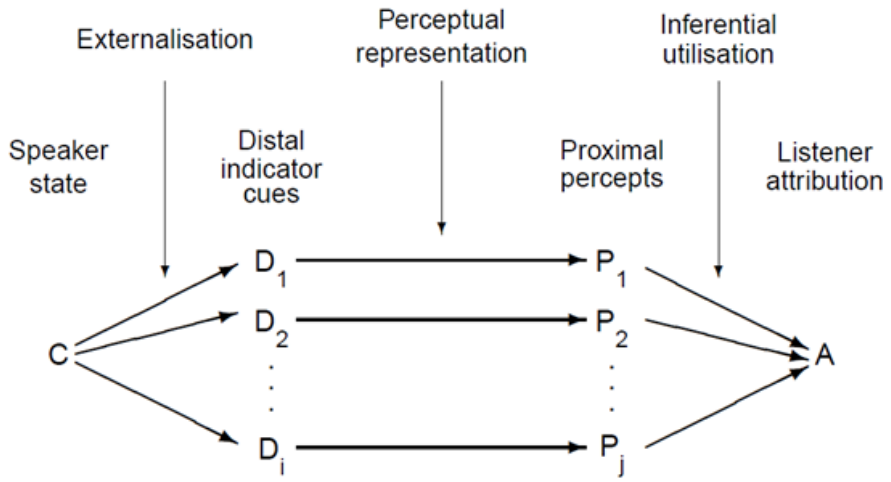


Figure 1.1: The Brunswikian lens model, adapted from Scherer (1978) [79].

The steps of human perception according to Scherer model are as follow:

- a speaker emotional state is expressed through a number of objectively measurable parameters, the so-called "distal indicator cues", in case of speech and emotion, these parameters are acoustic features.
- in the first step of the perceptual inference process, the acoustic features (distal cues) are perceived by a listener and internally represented as "proximal percepts".
- these percepts are used for "attribution" by the listener for inferring the speakers's sate. In speech and emotion examples of proximal percepts are subjectively perceived pitch or voice quality, while the attribution is the perceived speaker emotion.

Huang and Akagi adopted a three-layer model for human perception as shown in Figure 1.2. They tried to imitate human perception by using three-layer model instead of

two-layer model. Therefore, they assumed that human perception for emotional speech does not come directly from a change in acoustic features but rather a composite of different types of smaller perceptions that are expressed by semantic primitives or adjectives describing an emotional voice [35]. Akagi's model could be seen as a special case of Lens model, where the "distal indicator cues", 'proximal percepts", and "attribution" in Len's model correspond to the "acoustic features , "semantic primitives", and "emotional category", respectively in Akagi's model.

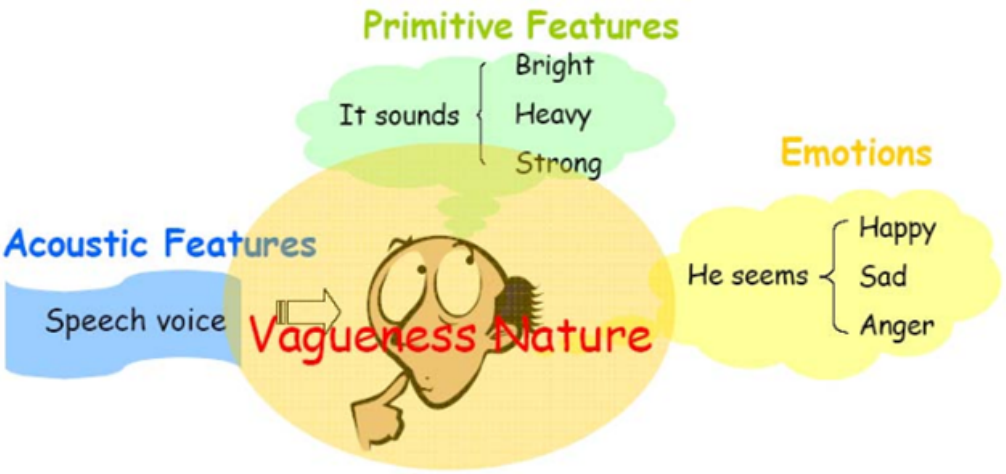


Figure 1.2: Schematic graph of human perception of emotional voices from [35].

In this thesis, the proposed idea to improve automatic speech emotion recognition system can be done by imitating the process of human perception for understanding the emotional state from the speech signal. The conventional two-layer model has limited ability to find the most relevant acoustic features for each emotion dimension, especially valence, or to improve the prediction of emotion dimensions from acoustic features. To overcome these limitations, this study proposes a three-layer model to improve the estimating values of emotion dimensions from acoustic features. Our proposed model consists of three layers: emotion dimensions (valence, activation, and dominance) constitute the top layer, semantic primitives the middle layer, and acoustic features the bottom layer. A semantic primitive layer is added between the two conventional layers acoustic features and emotion dimensions as shown in Figure 1.3.

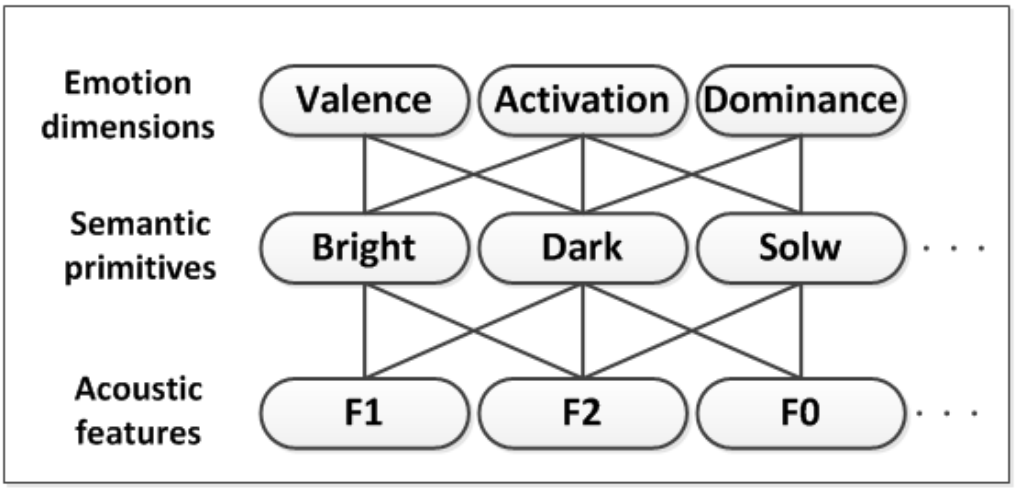


Figure 1.3: The proposed three-layer model.

We first, assume that the acoustic features that are highly correlated with semantic primitives will have a large impact for predicting values of emotion dimensions, especially for valence. This assumption can guide the selection of new acoustic features with better discrimination in the most difficult dimension.

The second assumption is that human can judge the expressive content of a voice even without the understanding of one language, such as emotional state of the speaker from different language. Using the second assumption, we investigate the universality of the proposed speech emotion recognition system to detect the emotional state cross-lingually. To accomplish this task, the most relevant acoustic features for each emotion dimension for the two different languages were investigated. Finally, the common acoustic features between the two languages can be used as the input of the cross-language speech emotion recognition system. The features found in one language were used to estimate emotion dimensions for the other language, and vice-versa.

1.5 Human perception for emotional state

In this study , we adopt the improved Brunswik's lens model for human perception by Huang and Akagi [35]. The human perception model is consists of three-layer: acoustic features, semantic primitives, and emotion dimensions layer, respectively. Figure 1.4

shows the process of human perception to judge the emotional state expressed by speakers. The human perception process is composed of two small process: the first process is semantic primitive perception in which the listener judge the degree of all adjectives describing the emotional voice, such as very Bright, very Slow, an so on, the final process is emotion perception process by judging the degree of emotional state from the adjectives describing this voice.

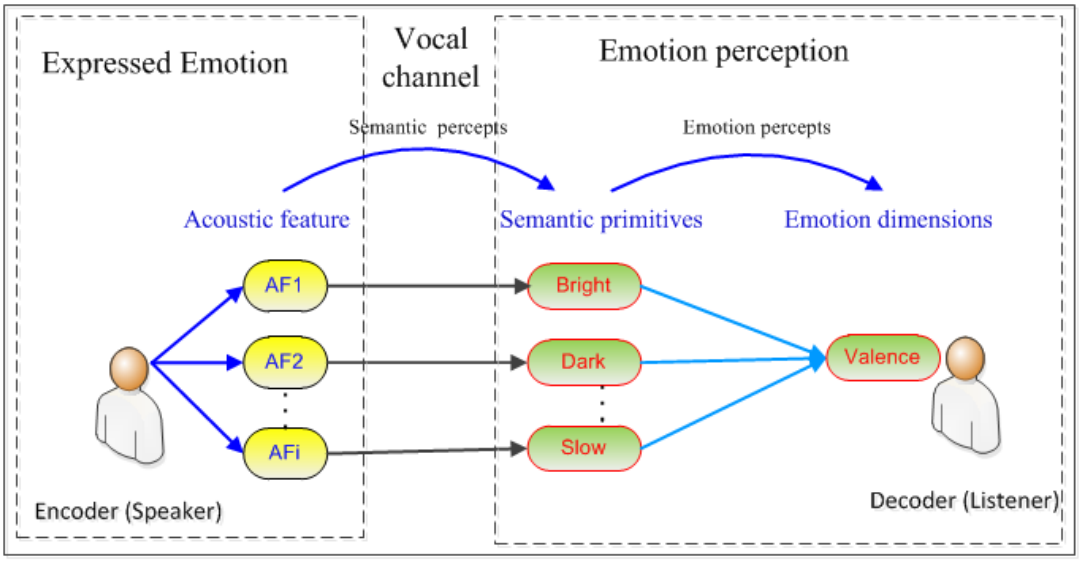


Figure 1.4: The improved Brunswik's lens model for human perception.

1.6 Research methodology

The feasibility of our three-layer model to improve emotion dimensions estimation; for valence, activation, and dominance was investigated. Our model consists of three layers: emotion dimensions (valence, activation, and dominance) constitute the top layer, semantic primitives the middle layer, and acoustic features the bottom layer. A semantic primitive layer is added between the two conventional layers acoustic features and emotion dimensions as shown in Figure 1.3.

Therefore, the approach we adopt includes the following steps:

- Feature selection: The most relevant acoustic features were selected by using a top-

down method. First, the most correlated semantic primitives were selected for each emotion dimension. Then, the most correlated acoustic features with the selected semantic primitives found in the first step were selected.

- Building a three-layer model for each emotion dimension: For example, in the case of valence dimension, the three layers are: valence dimension in the top layer, the highly correlated semantic primitives with valence dimension in the middle layer, all the highly correlated acoustic features with all semantic primitives in the bottom layer.

- Emotion dimensions estimation: By using the constructed three-layer model, a button-up method was used to estimate values of emotion dimensions from acoustic features as follows. First, fuzzy inference system (FIS) was used to estimate the degree of each semantic primitive from acoustic features, and then another FIS was used to estimate values of emotion dimension from the estimated degrees of semantic primitives in the first step.

Implementing an automatic emotion recognition system which estimate emotion dimensions based on a three-layer model of human perception should provide concrete support for our concept. To achieve our aims from this study: we construct mono-language emotion recognition system which can estimate emotion dimensions, across training and testing the system using the same language. Moreover, we construct a cross-language emotion recognition system which can estimate emotion dimension form the speech regardless of language, i.e. training the system using one language and testing using different language.

Therefore, the three-layer model was used to investigate whether there are acoustic features that allow us to estimate the emotional state from the voice of a person no matter what language he/she speaks.

To accomplish this, we work with two databases of emotional speech, one in Japanese and the other in German. We extract a variety of acoustic features and build the three-layer model for each dimension for the two languages individually. The top-down acous-

tic feature selection method was used to find the best acoustic feature subsets for each language. Finally, we construct two mono-language emotion recognition systems which predict the emotion for each language individually; Japanese-from-Japanese, and German-from-German and two cross-language emotion recognition systems which can estimate the emotion using cross-language mode Japanese-from-German, and German-from-Japanese. Using the following steps:

- We look for acoustic features that allow us to estimate emotional states from speech regard less the spoken language(Japanese/German)
- The constructed three layer model was used to predict the emotion dimensions for each language from the acoustic features of the other language

1.7 Outline of the thesis

The Thesis is organized as follows:

- **Chapter 1** describes the general aims and the specific issues of this study. Firstly, we introduce the objective of the present study and define the adopted problems and proposed solutions for these problems.
- **Chapter 2** introduces a general literature review on the state-of-the-art emotional research: concepts, theoretical frame work, and automatic emotion recognition system aspects. Thus, first the two emotion representation (categorical and dimensional) are presented. Then, merits of the dimensional representation are discussed. The relationship between the categorical approach and the dimensions approach is introduced. Moreover, this chapter gives an overview of the literature related to speech emotion recognition system. The literature will be reviewed under different aspects, among them emotion units, features, and classifiers. Finally, the process of emotion dimension estimation using fuzzy inference system estimator is introduced in details.

- **Chapter 3** introduces the elements of the proposed system; the used databases (German and Japanese) databases, acoustic features and experimental evaluation for semantic primitives and emotion dimensions using human evaluation. Firstly we extracted 21 acoustic features from the two databases. Two experiments were conducted for both Japanese and German database: the first experiment is to evaluate the 17 semantic primitives for each utterance, while the second experiment was conducted to evaluate emotion dimensions valence, activation, and dominance for each utterance. Inter-rater agreement was measured by means of pairwise correlations between subjects' mean ratings of each utterance, separately for each semantic primitives and emotion dimensions.

- **Chapter 4**, the first half of this chapter introduce the feature selection method, a top-down feature selection method was proposed to select the most related acoustic features based on the three-layer model. By firstly, selecting the highly correlated semantic primitives for emotion dimension, then selecting the set of all acoustic features which are highly correlated with the selected semantic. The set of selected acoustic features are considered the most related to the emotion dimension in the top layer. For each emotion dimension, a perceptual three-layer model was constructed as follows: the desired emotion dimension in the top layer, the most relevant semantic primitives in the middle layer, the most relevant acoustic features in the bottom layer.

  The second half this chapter, pretenses the implementation of the proposed system, the constructed perceptual three-layer model for each emotion dimension was used to estimate emotion dimensions using a bottom-up method. This method was used to construct our emotion recognition system as follows: the input of the proposed system are the acoustic features in the bottom layer, the output of are the emotion dimensions valence, activation, and dominance. Fuzzy inference system (FIS) was used to connect the elements of the proposed system. Firstly, one FIS was used to estimate each semantic primitive in the middle layer form the acoustic features

in the bottom layer. Then one FIS was used to estimate each emotion dimensions from the estimated semantic primitives.

- **Chapter 5** investigates the following questions: whether the selected acoustic features are effective for predicting emotion dimensions? second, whether the proposed emotion recognition system improve the estimation accuracy of emotion dimensions (valence, activation, and dominance) or not? The mean absolute error (MAE) is used to measure performance of the proposed system, by the distance between the estimated dimensions using the proposed system and the evaluated emotion dimensions using human listeners.

  Firstly, to investigate the first question, the most relevant acoustic features for each emotion dimension were used as inputs of the proposed emotion recognition system, to estimate values of emotion dimensions. Then, the estimation results of emotion dimensions are compared with those of estimation using the non-relevant acoustic features and all acoustic features.

  Furthermore, to investigate the second question which mean is how effectively our proposed system improve emotion dimensions estimation. Therefore, the performance of the proposed system was compared with that of the conventional two-layer system, using two different languages Japanese and German, with two different tasks (speaker-dependent task and multi-speaker task).

  Therefore, two emotion recognition system were constructed the first system was constructed based on the proposed approach and the other based on the conventional approach. The selected acoustic features group was used as input for both the proposed system and the conventional system.

  The most important results is that the proposed automatic speech emotion recognition system based on the three-layer model for human perception was superior to the conventional two-layer system.

- **Chapter 6** introduces a cross-lingual emotion recognition system that has the abil-

ity to estimate emotion dimensions for one language by training the system using another language. To accomplish this task, first, we investigate whether their are common acoustic features between the two languages. Second, we construct a cross-language emotion recognition system based on human perception three-layer model to accurately estimate emotion dimensions.

For both languages, our proposed feature selection method was used to select the most relevant acoustic features for each emotion dimension. Then, the common acoustic features between the two language were selected as inputs to the cross-language emotion recognition system, and the outputs of this system are the estimated emotion dimensions: valence, activation, and dominance.

For estimating emotion dimensions, the proposed cross-language emotion recognition system was trained using one language and testing using the second language. For instance, Japanese emotion dimensions were estimated form German database by training the system using acoustic features, semantic primitives, and emotion dimensions for each German speaker dataset individually, then the trained system was used to estimate Japanese emotion dimensions using Japanese acoustic features as inputs, in a similar way the German emotion dimensions were estimated from Japanese database.

The results of proposed cross-language emotion recognition system are presented and compared with the prediction from mono-language emotion recognition system.

- **Chapter 7**, the estimated emotion dimensions were mapped using Gaussian Mixture Model (GMM) classifier into emotion categories for both database. The results of the classification using the proposed method was compared with the classification of emotion categories from acoustic features directly using GMM.

  For the Japanese database, the overall recognition rate was 53.9% using direct classification using acoustic features and up to 94% using emotion dimensions. For the German database, the rate of classification directly from acoustic features was 60%, which was increased by up to 75% and 95.5% using emotion dimensions for

multi-speaker and speaker-dependent tasks, respectively. The result reveals that the recognition rate using the estimated emotion dimensions is higher than the direct classification using acoustic features directly.

- **Chapter 8**, finally concludes this thesis with respect to the research questions and give an outlook on future work.

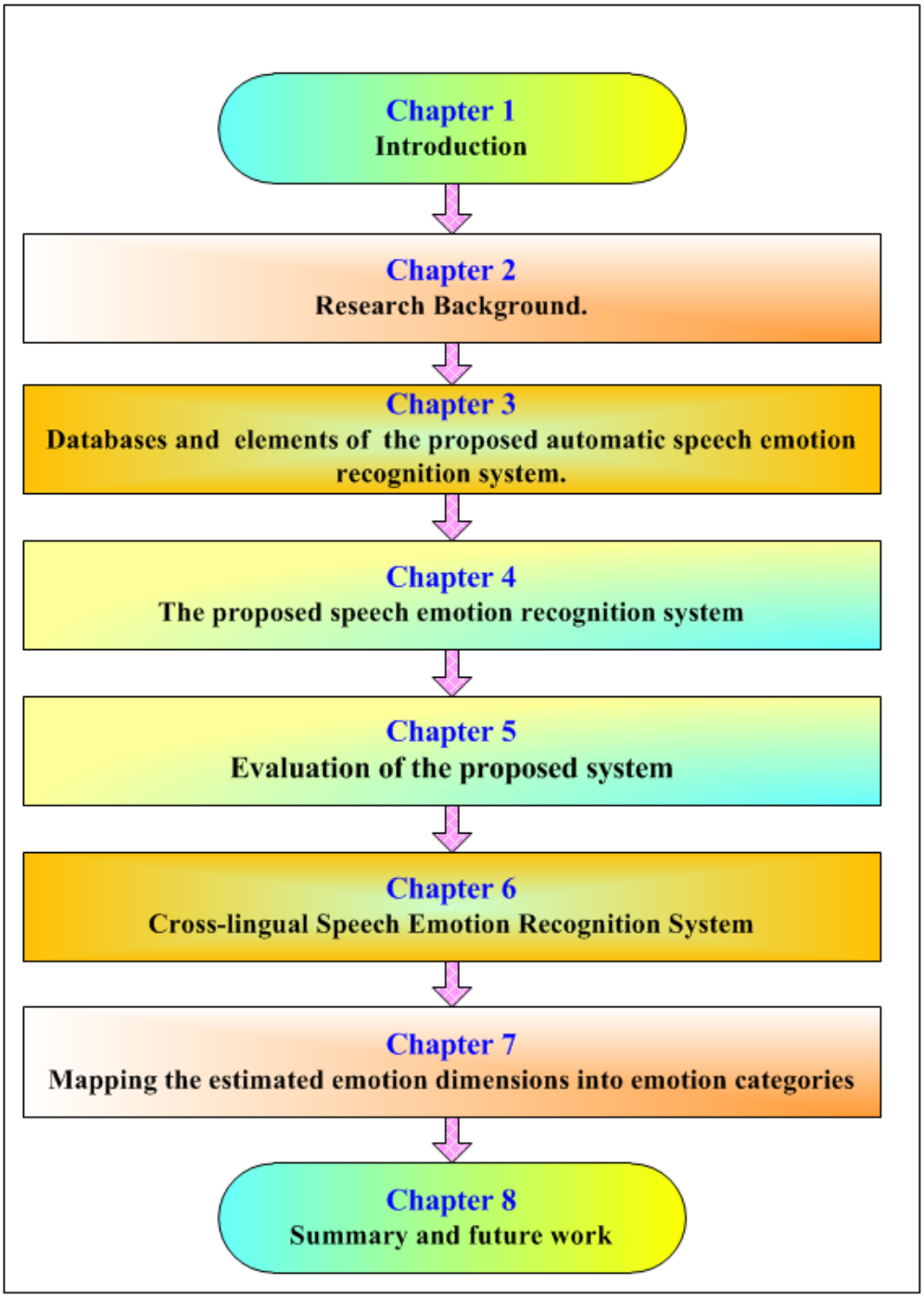


Figure 1.5: The Outline of the dissertation.

Chapter 2

Research Background

2.1 Introduction

This chapter introduces a general literature review on the emotional research from the speech signal: concepts, theoretical frame work and practical considerations necessary for constructing automatic emotion recognition system to detect the emotional state from speech. To recognize emotions, one first needs a precise idea of how to represent them. Emotion theories have a long tradition in psychology, having produced many models that can be used as basis for automatic speech emotion recognition. The most relevant ones in view of speech are presented in Section 2.2. They are also discussed regarding to what extent they are feasible to realize in practical applications.

The next question to deal with is that, where emotions can be observed? They are expressed in language, through acoustic, syntactic or semantic information, but also on other levels of human behavior as facial or body gestures. Machines, however, can also exploit information obtained by measuring body signals like heart rate or perspiration to predict the emotional state of a person. In this study, our focus on detecting the emotional state expressed in speech signal as introduced in Section 2.3.

Automatic emotion recognition is actually a pattern recognition problem depending strongly on: (1) the features extracted; (2) the classifier used; (3) the speech corpus used for training the classifier; (4) the emotion representation that the systems architecture is implemented for classifying. Having introduced these notions, a closer look on automatic emotion recognition from speech is presented in Section 2.4. After presenting a general system design, ranging from feature extraction over acoustic feature selection to the actual classification, possible features as acoustic correlates of emotions in speech are discussed and the traditional feature selection methods are described in details, since the finding of the most relevant acoustic features is a major part of this thesis.

Finally, the details of constructing and evaluating a speech emotion recognition system based on the dimensional approach using fuzzy inference system were introduced.

2.2 Types of emotion representation

In the area of automatic emotion recognition, mainly two classifying approaches have been used to capture and describe the emotional content in speech: categorical and dimensional approaches. Categorical approach is based on the concept of basic emotions such as anger, joy, and sadness, which are the most intense form of emotions from which all other emotions are generated by variations or combinations of them. They assume the existence of universal emotions that can be clearly distinguished from one another by most people. On the other hand, dimensional approach represents emotional states using a continuous multi-dimensional space. Both approaches, categorical and dimensional, provide complementary information about the emotional expressions observed in individuals. In the rest of this section the two representation will be introduced in more details. Finally, the advantages of the dimensional representation as will as the relation between the categorical and dimensional representation are also presented.

2.2.1 Categorical representation

The categorical theory proposes the presence of six basic, distinct, and universal emotions: happiness, anger, sadness, surprise, disgust, and fear [21, 19, 20, 18, 41, 84]. The simplest description of emotions is the use of emotion category labels. Most of the previous researchers treat the emotion recognition problem as a multiple classification task of several emotional categories such as angry, happy, and sad; or simply, negative and non-negative.

Discrete categorization allows a more particularized representation of emotions in applications where it is needed to recognize a predefined set of emotions. However, this approach ignores most of the spectrum of human emotional expressions. Some studies concentrate on only one or two selected categories.

One of the difficulties in comparing studies into emotions in research is that the choice of categories for a study varies and usually depend on an application that the researcher has in mind. There are a lot of problem facing the researcher who using the category approach such as: How many category they should use to describe the real-life emotion?

The short list of options shows that even if one decides to model emotions in terms of categories, it is not immediately clear what categories to use. The most frequently used categories may not be the most suitable ones for a given research question or application. In contrast, it is also important to detect the variability within a certain emotion (e.g., "a little happy" or "very happy") in addition to the emotion categories. This is supported by the fact that human soften or emphasize their emotional expressions flexibly depending on the situation in actual human speech communication.

Therefore, a single label or any small number of discrete categories may not accurately reflect the complexity of the emotional states conveyed in everyday interaction [2].

2.2.2 Dimensional representation

Many different approaches reported in the psychological literature have led to the proposal of dimensions underlying emotional concepts, through representing the emotional state as a point in a multi-dimensional space [70, 71, 75]. The used dimensions in this representation are gradual in nature and represent the essential aspects of emotion concepts (how negative or positive, how aroused or relaxed, how powerful or weak) rather than the fine specifications of individual emotion categories. It is important to know that the names used for these dimensions were actually selected by the individual researchers interpreting their data, and did not arise from the data itself. In this study, the following names for emotion dimensions are used: the terms valence (synonymous to evaluation or pleasure), activation (used as synonymous to arousal and activity) and dominance (potency or power).

In general, there are several ways to represent emotions in a multi-dimensional emotion space. Two-dimensional representations include one dimension that describes the valence taking values (from positive to negative). The other emotion dimension describes the activation or arousal from high to low) as shown in Figure 2.1, basic emotions are marked as areas within the two-dimensional space in this figure.

Three-dimensional representations additionally include a third dimension defining the

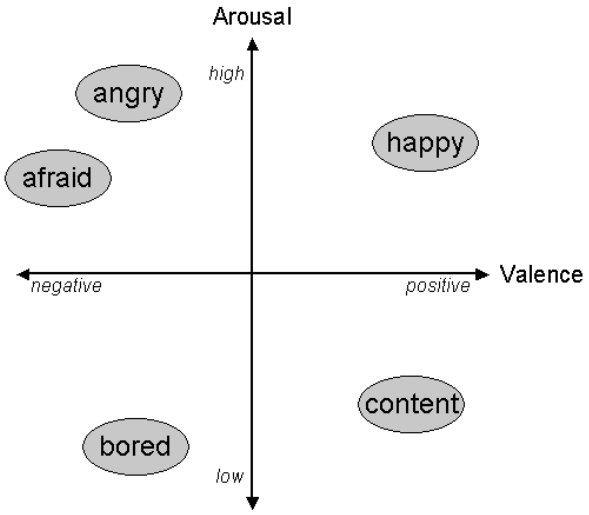


Figure 2.1: A two-dimensional emotion space with a valence and an arousal axis. Basic Emotions are marked as areas within the space.

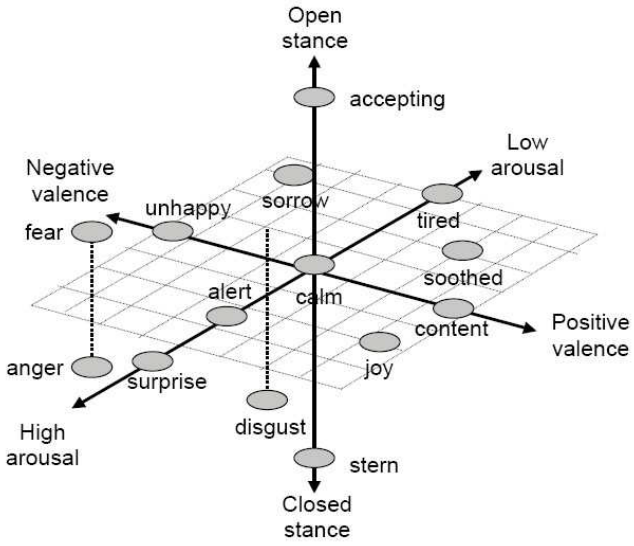


Figure 2.2: Emotional categories mapped into Arousal-Valence-Stance space, Fourteen emotions located in Arousal-Valence-Stance space [4].

apparent strength of the person, which is referred to as dominance (or power). According to both the work of Schlosberg (1954) and Scherer et al. (2006) this dimension is even more important than the activation dimension [75, 67]. Especially in the case of high activation, Gehm & Scherer (1988) found that taking the level of control and social power of an individual into account is useful in distinguishing certain emotion [28]. This finding

is supported by Russell & Mehrabian (1977) [67], who could show that anger and fear both consist of similarly very negative and high activation values and can only be distinguished due to their different values on the dominance scale as shown in Figure 2.2. This third dimension is necessary to distinguish anger from fear, since the dominance (or the ability to handle a situation) is the only discriminating element in this case.

One powerful representation is in terms of the three emotional attributes introduced by Grimm et al., they proposed a generalized framework using a continuous-valued, three-dimensional emotion space method [32, 31]. This method defines emotions as points in a three-dimensional emotion space spanned by the three basic dimensions valence (negative-positive), activation (calm-excited), and dominance (weak-strong). Figure 2.3 shows a schematic sketch of this emotional space.

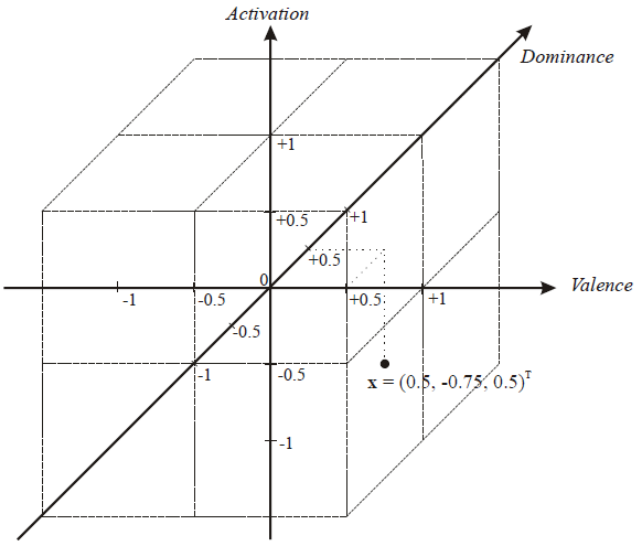


Figure 2.3: Three-dimensional emotion space, spanned by the primitives valence, activation, and dominance, with a sample emotion vector added for illustration of the component concept.

2.2.3 Merits of the dimensional representation

It is important to think carefully about the type of representation most suitable for a given task. Emotions have different degree of intensity, and may change over time depending on the situation from low to high degree, for example, human listener may detect or

describe the emotional state as little happy or very happy. Consequently, an automatic speech emotion recognition system should be able to detect the level or the intensity of the emotional state form the voice.

In addition, it seems reasonable to assume that most human-machine interaction will require the machine to recognize only mild, non-extreme emotional states. Therefore, the need to express full-blown emotions is a marginal rather than a central requirement, while the main focus should be on the systems capability to express a large variety of emotional states of low to medium intensity. Emotion dimensions are a representation of emotional states which fulfills these requirements: They are naturally gradual, and are capable of representing low-intensity as well as high-intensity states as shown in Figure 2.4, each emotion category have different level or degree for example, happy in the first quarter is represented by three faces which represent little happy, happy, and very happy, respectively.

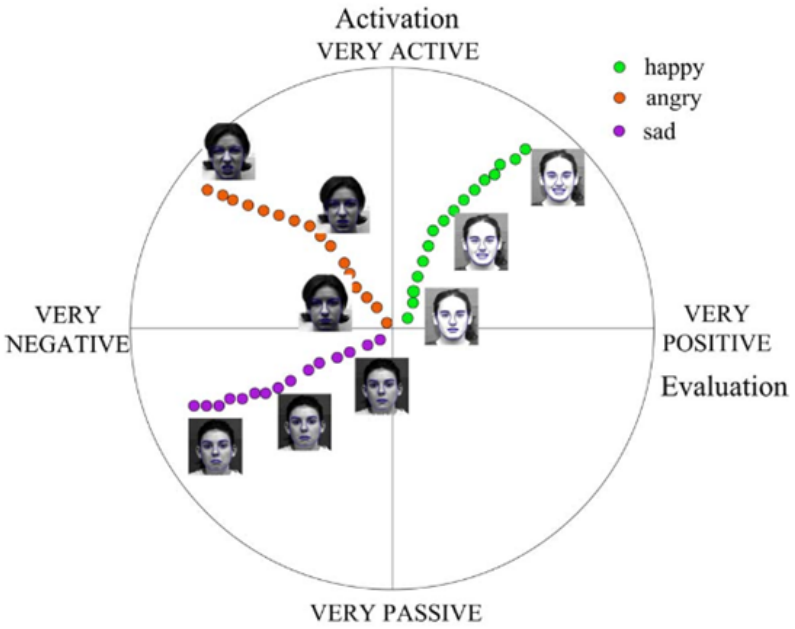


Figure 2.4: Labeling of facial image sequences in the emotional space [96].

In the categorical approach, where each emotional state is classified into a single category, a complex mental or affective state or blended emotions perhaps too difficult to

handle [99]. Contrarily, in the dimensional approach, emotional transitions can be easily captured, the numerical representations are more appropriate to reflect the gradient nature of emotion expressions, in which observers can indicate their impression of moderate (less intense) and authentic emotional expressions on several continuous scales [56, 94] .

In this work, the three-dimensional continuous model is adopted in order to represent the emotional states using emotion dimensions i.e. valence, activation and dominance. This approach is chosen because it exhibits great potential to model the occurrence of emotions in real-world as in a realistic scenario, emotions are not generated in a prototypical or pure modality, but rather than in complex emotional states, which are a mixture of emotions with varying degrees of intensity or expressiveness [24]. Therefore, this approach allows a more flexible interpretation of emotional states [101].

2.2.4 Mappings between emotion representations

The categorical and the dimensional approach are closely related, i.e. by detecting the emotional content using one of these two schemes, it will be essay to infer its equivalents in the other scheme. For example, if an utterance is estimated with positive valence and high activation, then, it could inferred that the emotional category for this utterance is Happy, and vice versa. Therefore, any improvement in dimensional approach will leading to an improvement in the categorical approach.

The estimated values of emotion dimensions (valence, activation, and dominance) are found to be transferable to emotion categories, if desired [30], for example in [30] the estimated emotion dimensions were mapped to the emotion categories using k-nearest neighbor (kNN) classifier. The results reveal that, the achieving recognition rate significantly higher than the traditional categorical classification from acoustic features directly.

In [82] the experimental results indicate that an alternative way of classifying emotions can be seen as finding a place in the emotional space, and infer from such location and from additional information, i.e. context, application, if available, the intended emotion.

The advent of describing emotion as a point in multi-dimension space has led to

Table 2.1: Emotion and Speech Parameter (From Murray and Arnott, 1993)

	Anger	Happiness	Sadness	Fear	Disgust
Speech rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflects
Articulation	Tense	Normal	Slurring	Precise	Normal

identify the exact emotion for the speaker. Also, the three-dimensional emotion attribute estimates could be classified into the emotion categories. This procedure allows for a comparison of the calculated estimation errors to classical recognition rates. Therefore, mapping emotion dimensions into emotion categories will strengthen our findings in this study by demonstrating that, the dimensional approach can actually help us to improve the automatic emotion classification.

2.3 The expression of emotions in human speech

After having reviewed how emotions can be described, the next question is where emotions can be observed. In this study, our focus on detecting the emotional state expressed in speech signal. Information on emotion is encoded in all aspects of language, in what we say and in how we say or pronounce it, and the "how" is even more important than the "what". Looking at all levels of language, the following thing can be considered: a speakers intention is highly correlated with his emotional state.

To improve the speech emotion recognition accuracy we can achieve this goal only if there are some reliable acoustic correlates of emotion in the acoustic characteristics of the signal. A number of researchers have already investigated this question. Murray and Arnott have conducted a literature review on human vocal emotion (Table 2.1) and concluded that in general, the correlation of the acoustic characteristics, both prosody and voice quality, and the speakers emotional state are consistent among different studies, with only minor differences being apparent [53].

They concluded that the pitch envelope (i.e., the level, range, shape, and timing of the pitch contour) is the most important parameter in differentiating the basic emotions, and the voice quality is important in differentiating the secondary emotions. It has also been noted that the acoustic correlates of basic emotions are cross-cultural, but those of the secondary emotions are culturally dependent.

In the previous studies, researchers have often found it useful to define emotions in some multi-dimensional space. Past research has been relatively successful at discovering acoustic correlates distinguishing emotions on the grounds of activation, but less successful for valence. These studies have found that positive-activation emotions have high mean F0 and energy as well as a faster speaking rate than negative-activation emotions. In this study we are interested in looking at the correlations between acoustic features and the three emotion dimensions: valence, activation, and dominance.

2.4 Automatic speech emotion recognition system

Automated recognition of emotions conveyed in the speech is an important research topic that has emerged in recent years with a number of possible applications. The most important one is probably the improvement of the man-machine interface, knowing that human communication contains a large amount of emotional messages which should be recognized by machines such as robot assistants in the household, computer tutors, or Automatic Speech Recognition (ASR) units in call-centers [32].

One of the first practical works on automatic speech emotion recognition based on the dimensional approach was conducted by by Grimm et al. [31] in 2007. In this contribution they proposed a generalized framework using a real-valued, three-dimensional representation. This representation defines emotions as points in a three-dimensional emotion space spanned by the three basic dimensions: valence (positive-negative axis), activation (calm-excited axis), and dominance (weak-strong axis). However, they found that activation and dominance were more accurately estimated than valence. For instance, Grimm et al. attempted to estimate the emotion dimensions (valence, activation,

and dominance) from the acoustic features by using a fuzzy inference system (FIS) [31]. However, they found that activation and dominance were more accurately estimated than valence. Therefore, our aim in this study is to improve the dimensional approach in order to estimate accurately all emotion dimensions.

In this section, first a general overview of a system for emotion recognition from speech is given. Subsequently, an overview of the most important acoustic feature for classifying emotions, and in the end of this section the traditional features selection is introduced in more details.

2.4.1 Overview of speech emotion recognition system

A speech emotion recognition system consists of three principal parts as shown in Figure 2.5, feature extraction, feature selection and emotion recognition. The purpose of the feature extraction is to find those properties of the acoustic signal that are characteristic of emotions and to represent them in an n-dimensional feature vector.

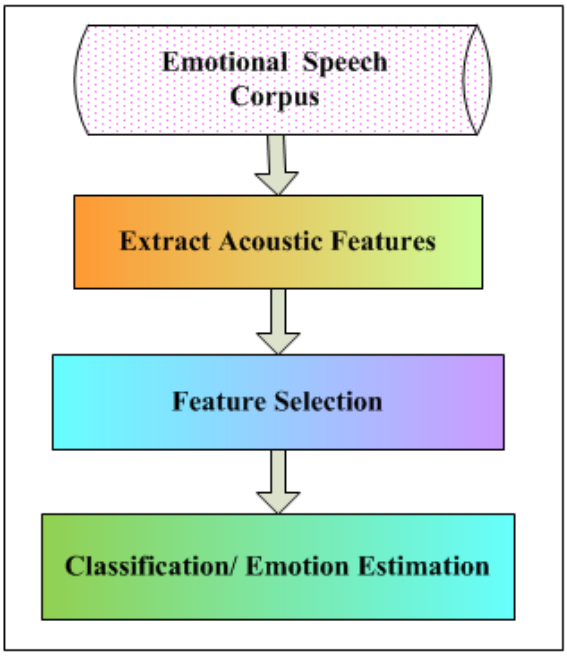


Figure 2.5: The process of speech emotion recognition.

So far, there is not yet a general agreement on which features are the most important

ones and good features seem to be highly data dependent, As a consequence, most approaches compute a high number of features and apply then a feature selection algorithm, in order to reduce the dimensionality of the input data. The feature selection algorithm chooses the most significant features with respect to the data for the given task. In this study, correlation analysis method have been applied as feature selection method, to find a subset of features for improving prediction accuracy.

After the feature selection method, each emotion unit is represented by one or more feature vectors, and the problem of emotion recognition can now be considered a general pattern classification problem as shown in Figure 2.6. The output of emotion classification techniques is a prediction value for the three dimensions valence, activation, and dominance.

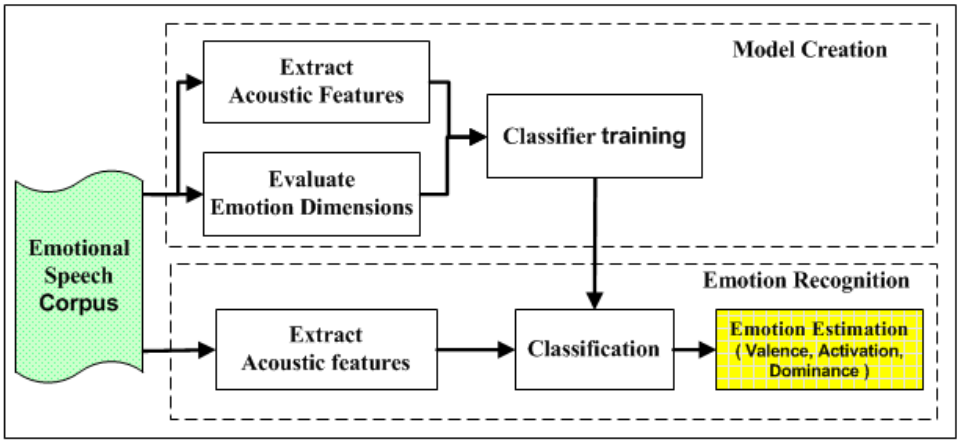


Figure 2.6: Block diagram of emotion recognition analysis using the two-layer model.

Static as well as dynamic classification approaches are considered. In static modeling one feature vector represents one emotion unit, while in dynamic modeling, one emotion unit is represented by a sequence of feature vectors. The latter therefore consider also the temporal behavior of the features. As timing is very important for emotions, temporal information has to be encoded in the feature vector for static modeling, which usually leads to high-dimensional vectors, while the feature vectors used in dynamic modeling are generally smaller. So, in principle, any classifier can be used, though Support Vector Machines, Neural Networks and Bayesian classifiers for static modeling and HMMs for

dynamic modeling are most commonly found in the literature on emotional speech recognition. Currently, static modeling approaches prevail. In this study we adopt the static modeling, the parameters of the classifier are learnt from training data.

As a first step towards exploiting the dimensional approach is extraction of the most important acoustic features for automatically estimate emotion dimensions. Then, this estimation can be used to locate the individual's emotional state in the multi-dimensional space, and if necessary, to map it to a basic emotion as described in Section 2.2.4. In order to improve the emotion speech recognition, we have to search for the golden set of vector features. Chapter 4 in this study describe a novel feature selection method based on a three-layer model of human perception to select the best correlated acoustic features for each dimension.

2.4.2 Acoustic features related to emotion speech

For automatic speech emotion recognition system the acoustic features are the most important cues that can be used to detect the emotional content from the speech signal. The acoustic features can be divided into prosodic features, spectral features, and voice quality features. The first question that we try to investigate is which acoustic features are mostly relevant for describing the emotion dimensions? So far, as numerous studies have investigated this type of effect, there are various acoustic features that can be extracted from the speech signal, some are found to be significant to the perception of expressive but some are not. For example, the acoustic effects of anger, joy and fear are very similar, although these emotions are usually not hard to distinguish for humans. Therefore, there is no easy mapping from acoustics to emotions.

In the following those acoustic measures that are the basis of most features used for automatic emotion recognition will be described. The importance of these acoustic features For human perception is quite well studied and explored, however, the extent of their respective significance for automatic emotion recognition is not yet ultimately resolved.

In this chapter, before extracting of acoustic features in the next chapter, it is necessary to consider what acoustic features should be extracted. According to much previous studies, the most important features for emotion recognition are conveyed by the parameters of fundamental frequency (vocal pitch), duration, intensity, formant, and voice quality. These features are considered to be very important factors correlated to the perception of expressive speech.

These features are described below.

- **Fundamental frequency (F0)** Fundamental frequency is one of the three features that is most consistently used in the literature because it represents the intonation and the pitch contour of a speech utterance. Therefore, when analyzing acoustic correlates to emotional speech, fundamental frequency is crucially important.

- **Duration** The speed of the speech utterance or the length varies greatly when a speaker utter the same word in different emotional states. For example, when some one in anger state he/she speak very fast, while when someone in sad emotional state the speed of speaking is very slow. In a speech utterance, the durations of phonemes, words, phrases, and pauses compose the prosodic rhythm. Consequently, duration is also important to investigate.

- **Intensity** Intensity of a speech utterance is perceived primarily as loudness. It is determined by the volume of the air flow of breath sent out by the lungs and is represented by the power contour of the speech signal. A voice with different levels of loudness can be perceived differently. When people are in a good mood, such as happy or cheerful, the voice is usually at a higher level of loudness. Conversely, when people are in a sad mood such as depressed or upset, the voice is usually at a lower level of loudness. Therefore, for studying the perception of expressive speech, intensity is another feature that needs to be investigated.

- **Formants** Formants are those regions in the spectrogram where the amplitude of the acoustic energy is high. They reflect the natural resonance frequencies of the

vocal tract. These natural resonances are not fixed, but can be changed by altering the shape of the vocal tract. The resonances of the vocal tract can be changed, for example, by changing the position of the tongue or the jaw [4]. Formants are numbered from the lowest frequency upward (F1, F2, F3, etc.). A spectrogram can contain as many as ten identifiable formants, but most analyses do not go beyond the first five. Here we report on the most important formants for vowel identification, the first and second formant (F1 and F2). These two formants can be used to uniquely determine vowel identity. In [27] they investigated the effect of emotion dimension on the placement of F1 and F2. The results show that emotion has a sizable influence on formant positioning. For higher activation, the mean of F1 of all three vowels /a/. /i/, and /u/ is higher, while for lower activation is lower. In addition, for higher activation emotions have a significantly lower F2 for /a/. These effects confirm the important role of arousal in the vocal expression of emotion. Therefore, formants are one of the meaningful acoustic cues that should be investigated.

However, one problem that arises when measuring formants is that vocal tracts are different among different people. The analysis is more reasonable when the voices analyze are produced by the same people. Therefore, for justification all acoustic feature measured including formant were normalized to the neutral category for all speakers as described in Section 2.4.3.1. This normalization process was adopted to avoid the speaker-dependency of acoustic features such as formant acoustic feature.

2.4.3 Acoustic feature selection

Feature selection presents several advantages. To begin with, small feature subsets require less memory and computations, whereas they also allow for a more accurate statistical modeling, thus improving performance. On the contrary, large feature sets may yield a prohibitive computational time for classifier training. For example, neural networks face difficulties, when they are fed with extraordinary many features. Employing large

feature sets increases the risk of including features with reduced discriminating power. Additionally, feature selection eliminates irrelevant features, leading to a reduction in the cost of acquisition of the data. Furthermore, if all the features are employed, there is the risk for over-fitting. In addition, feature selection can increase performance when the number of training utterances is not sufficient or when a real-time problem needs to be handled.

However, before applying feature selection it is important to pre-process our features carefully in order to guarantee the quality of data. For that reason feature normalization are carried out.

2.4.3.1 Feature normalization

As a prerequisite to feature selection, normalization takes place. Feature normalization improves the features generalization ability and guarantees that all the features obtain the same scale in order to ensure an equal contribution of each feature to the feature selection algorithm.

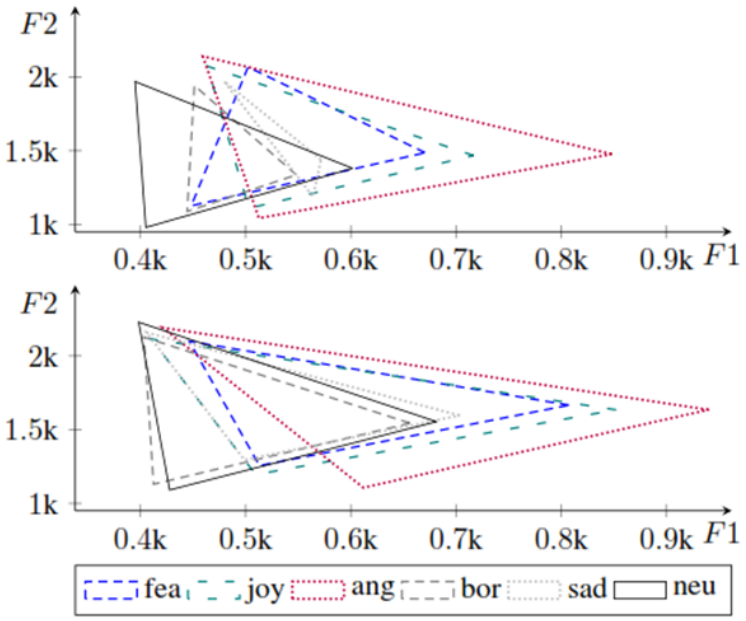


Figure 2.7: Classical vowel triangle form for different speakers emotional states. Speakers: male (top), female (bottom).

The values of acoustic features greatly changed for emotionally colored and neutral speech samples, for example, it was found a significant difference for the vowel triangles form and their position in F1-/F2-dimensional space for emotionally colored and neutral speech samples [88] as shown in Figure 2.7. This difference illustrates why automatic speech recognition models trained on neutral speech are not able to provide a reliable performance for emotion speech recognition. Another problem that arises when measuring formants is that vocal tracts are different among different people, i.e. formant frequency is speaker-dependent and emotion-dependent acoustic feature.

To justify the acoustic features to avoid both speaker-dependency and emotion-dependency, all acoustic feature measured in this study were normalized to the neutral one; we adopt an acoustic feature normalization method. Our normalization performed by dividing the values of acoustic features by the mean value of neutral utterances.

2.4.3.2 Feature selection

Feature selection. In most cases of affective speech analysis, conventional acoustic features are used without paying much attention to their selection. Whereas in our approach we take it of great importance because finding the acoustic correlates of valence emotion dimension in speech itself is a challenging task, and is crucial for the later stages as well.

The traditional methods for selecting the acoustic features for each emotion dimension were based on the two-layer model. These methods were based on correlations $R_i^{(j)}$ between acoustic features (i) and emotion dimensions (j) (valence, activation, dominance) as a two-layer model as follows:

Let $f_i = \{f_{i,n}\}(n = 1, 2, \ldots, N)$ be the sequence of values of the i^{th} acoustic feature. Moreover, let $x^{(j)} = \{x_n^{(j)}\}(n = 1, 2, \ldots, N)$ be the sequence of values of the j^{th} emotion dimension, $j \in \{Valence, Activation, Dominance\}$. Where M, N is the number of used acoustic features and the number of utterances in our databases, respectively.

The correlation coefficient $R_i^{(j)}$ between the acoustic features f_i and the emotion dimension $x^{(i)}$ can be determined by the following equation:

$$R_i^{(j)} = \frac{\sum_{n=1}^{N}(f_{i,n} - \overline{f_i})(x_n^{(j)} - \overline{x}^{(j)})}{\sqrt{\sum_{n=1}^{N}(f_{i,n} - \overline{f_i})^2}\sqrt{\sum_{n=1}^{N}(x_n^{(j)} - \overline{x}^{(j)})^2}} \tag{2.1}$$

where $\overline{f_i}$, and $\overline{x}^{(j)}$ are the arithmetic mean for the acoustic feature and emotion dimension, respectively.

Many researchers tried to investigate the most related acoustic features for each emotion dimension by using the correlation between a set of acoustic features and emotion dimensions [25, 93, 76, 80]. In all these studies, the valence dimension was found to be the most difficult dimension. Chapter 4 introduces in details the proposed acoustic feature selection method in this study.

2.5 Emotion dimension estimation

The aim of speech emotion recognition system based on the dimensional approach can be viewed as using an estimator to map the acoustic features to real-valued emotion dimensions (valence, activation, and dominance). The selected acoustic features form the previous section can be used as an input to the automatic speech emotion recognition system to predict emotion dimensions.

Emotion dimension values can be estimated using many estimator such as K-nearest neighborhood (KNN), Support Vector Regression (SVR), or Fuzzy Inference System FIS. In this study, for selecting the best estimator among KNN, SVR, and FIS, pre-experiments in our previous work [16] indicated that our best results were achieved using FIS estimator. The reason for using fuzzy logic is explained, before describing how to use it for emotion dimension estimation from the selected acoustic features.

2.5.1 The advantage of using fuzzy logic

Most of the statistical methodology mainly based on a linear and precise relationships between the input and the output, while the relationship between acoustic features and

emotion dimensions are non-linear. Therefore, fuzzy logic is a more appropriate mathematical tool for describing this non-linear relationship. The reasons are as follows:

- Fuzzy logic is a tool for embedding existing structured of human knowledge into mathematical models [45] using If-Then rules, and this is exactly what the model proposes to do in dealing with the perception of expressive speech.
- Fuzzy logic models non-linear functions of arbitrary complexity [92], and the relationship between emotion dimensions and acoustic features are certainly complex and non-linear. Therefore, fuzzy logic is appropriate to model these relationships.
- Fuzzy logic is based on natural language [40], and the natural language used in our model is in the form of semantic primitives (the middle layer of our model).

2.5.2 Fuzzy inference system

A FIS implements a nonlinear mapping from an input space to an output space by a number of fuzzy if-then rules constructed from human knowledge. The success of a FIS depends on the identification of the fuzzy rules and membership functions tuned to a particular application. It is usually difficult in terms of time and cost, and sometimes impossible, however, to transform human knowledge into a rule base [54]. Even if a rule base is provided, there remains a need to tune the membership functions to enhance the performance of the mapping. Neuro-fuzzy systems overcome these limitations by using artificial neural networks to identify fuzzy rules and tune the parameters of membership functions in FIS automatically. In this way, the need for the expert knowledge usually required to design a standard FIS is eliminated. A specific approach in neuro-fuzzy systems is ANFIS, which is a Sugeno type FIS implemented in the framework of adaptive neural networks [38].

To understand the fuzzy relationship between linguistic description of acoustic perception and expressive speech, a fuzzy inference system (FIS) will be built.

2.5.3 Adaptive Neuro Fuzzy Inference Systems ANFIS

Fuzzy inference system (FIS) is usually used as mathematical tool for approximating non-linear functions. This model can import qualitative aspects of human knowledge and reasoning process by data sets without employing precise quantitative analysis. The structure of the fuzzy inference system is shown in Figure 2.8 it is composed of the following five functional components:

- A rule base containing a number of fuzzy if-then rules.
- A database defining the membership functions of the fuzzy sets.
- A decision-making unit as the inference engine.
- A fuzzification interface which transforms crisp inputs to linguistic variables.
- A defuzzification interface converting fuzzy outputs to crisp output.

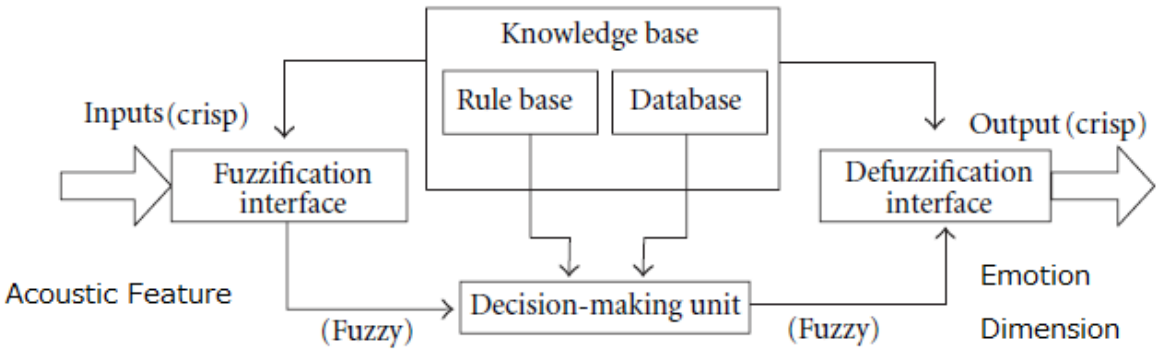


Figure 2.8: The structure of the fuzzy inference system.

ANFIS is a very common artificial intelligence technique in literature, which was proposed by Jang (1993) [38]. Although Fuzzy structure has strong inference system, it has no learning ability. In the contrary, Neural Network (NN) has powerful learning ability. ANFIS merges these two desired features in its own structure. As stated in the literature, the ANFIS can effectively predict highly non-linear models with much smaller root mean

square error values at the same number of iterations as compared to the conventional neural network-based models. Therefor, ANFIS is a neural-fuzzy system which contains both neural networks and fuzzy systems.

A fuzzy-logic system can be described as a non-linear mapping from the input space (acoustic features) to the output space (emotion dimensions). This mapping is done by converting the inputs from numerical domain to fuzzy domain. To convert the inputs, firstly, fuzzy sets and fuzzifiers are used. After that process, fuzzy rules and fuzzy inference engine is applied to fuzzy domain [37, 38]. The obtained result is then transformed back to arithmetical domain by using defuzzifiers. Gaussian functions are used for fuzzy sets and linear functions are used for rule outputs on ANFIS method. The standard deviation, mean of the membership functions and the coefficients of the output linear functions are used as network parameters of the system. The summation of outputs is calculated at the last node of the system. The last node is the rightmost node of a network. In Sugeno fuzzy model, fuzzy if-then rules are used (Sugeno and Kang 1988) (Takagi and Sugeno 1985) [78, 81]. The following is a typical fuzzy rule for a Sugeno type fuzzy system:

$$if\ x\ is\ A\ and\ y\ is\ B\ then\ z = f(x, y) \tag{2.2}$$

In this rule, A and B are fuzzy sets in anterior. The crisp function in the resulting is $z = f(x, y)$. This function mostly represents a polynomial. But exceptionally, it can be another kind of function which can properly fit the output of the system inside of the fuzzy region that is characterized by the anterior of the fuzzy rule. In this study first-order Sugeno fuzzy model is used for cases which are having $f(x, y)$ as a first-order polynomial. This model was originally proposed in (Sugeno and Kang 1988) (Takagi and Sugeno 1985) [78, 81]. Zero-order Sugeno fuzzy model is used for cases where f is constant. This can be called as a special case for Mamdani fuzzy inference system [51]. In this case, a fuzzy singleton is defined for each rules resultant. Or, this can be also called as a special case for Tsukamotos fuzzy model [85]. In this case, a membership function of a step function

is defined where it is centered at the constant for each rules consequent. Additionally, a radial basis function network under certain minor constraints is functionally correlative to a zero order Sugeno fuzzy model (Jang 1993). Lets investigate a first-order Sugeno fuzzy inference system having two rules:

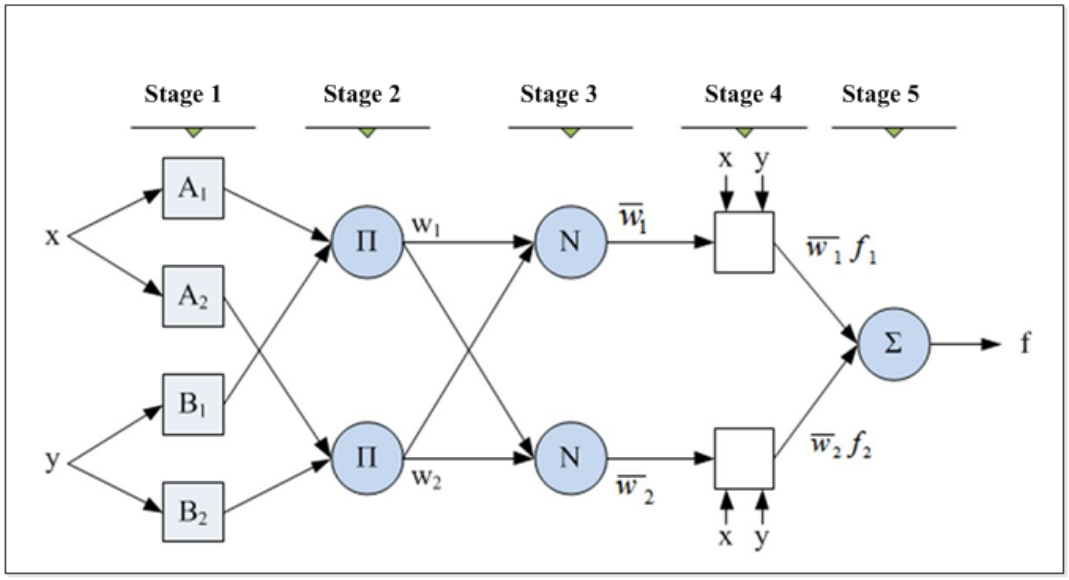


Figure 2.9: The Basic Architecture of ANFIS.

$$Rule\ 1:\ if x\ is\ A_1\ and\ y\ is B_1,\ then\ f_1 = p_1x + q_1y + r_1 \tag{2.3}$$

$$Rule\ 2:\ if x\ is\ A_2\ and\ y\ is B_2,\ then\ f_2 = p_2x + q_2y + r_2 \tag{2.4}$$

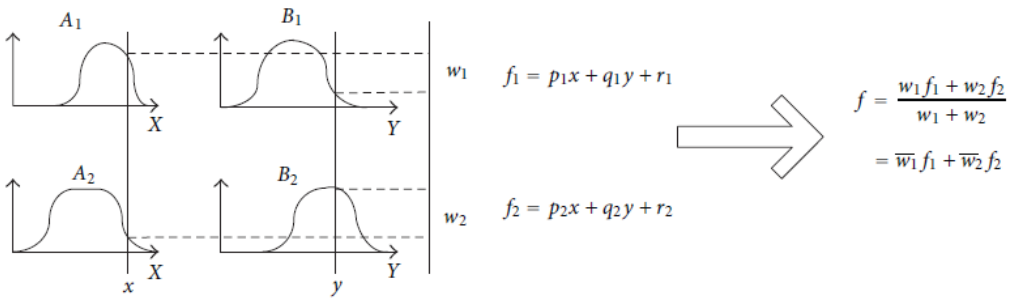


Figure 2.10: A two-input first-order Sugeno fuzzy model with two rules.

In the Figure 2.9, and 2.10, the fuzzy reasoning system is illustrated in a shortened form (Jang 1996) [39]. In order to avoid excessive computational complexity in the process of defuzzification, only weighted averages are used. On the previous figure, we see a fuzzy reasoning system. This system generates an output which is shown as f. To generate this output, system accepts an input vector $[x, y]$. The output is calculated by computing each rules weighted average. Those weights are achieved from the product of the membership grades in the assumption part. Using adaptive networks which are bound with the fuzzy model can compute gradient vectors. This computation is very helpful for learning of the Sugeno fuzzy model.

The learning algorithm that ANFIS uses contains both gradient descent and the least-squares estimate. This algorithm runs over and over till an acceptable error is reached. Running process of each iteration has two phases: forward step and backward step. In forward step, linear least-squares estimate method is used for obtaining consequent parameters and precedent parameters are corrected. In backward step, fixing of consequent parameters is done. Gradient descent method is used for updating precedent parameters. And also, the output error is back-propagated through network.

It is very important that the number of training epochs, the number of membership functions and the number of fuzzy rules hold a critical position in the designing of ANFIS. Adjusting of those parameters is very crucial for the system because it may lead system to over-fit the data or will not be able to fit the data. This adjusting is made by a hybrid algorithm combining the least squares method and the gradient descent method with a mean square error method. The lesser difference between ANFIS output and the actual objective means a better (more accurate) ANFIS system. So we tend to reduce the training error in training process.

A brief summary of 5 stages illustrated in Figure 2.9 of the ANFIS algorithm will be explained, each stage is described and necessary formulas are stated as follows:

- **Stage 1:** fuzzification stage: the parameters used in this stage is called premise parameters and rearranged according to output error in every loop. For this stage;

every node is an adaptive node with a node function and output calculated by Equation (2.5).

$$O_{1,i} = \mu_{A_i}(x), i = 1, 2, and \\ = \mu_{B_i}(y), i = 3, 4, \tag{2.5}$$

These parameters are membership grades of a fuzzy set and input parameters.

- **Stage 2:** A fixed node labeled Π, whose output is the product of all the incoming signals can be computed via Equation (2.6).

$$O_{2,i} = \mu_{A_i}(x)\mu_{B_i}(y), i = 1, 2, \tag{2.6}$$

Every output of the stage 2 affects the triggering level of the rule in the next stage. Trigger level is called firing strength and N norm operator is called AND operator in fuzzy system.

- **Stage 3:** This layer can be called as normalization layer. In this stage, all firing strengths are re-arranged again by their own weights as shown in Equation (2.7).

$$O_{3,i} = \overline{\omega_i} = \frac{\omega_i}{\omega_1 + \omega_2}, i = 1, 2, \tag{2.7}$$

- **Stage 4:** (Defuzzication): This stage is a preliminary calculation of the output for real world. This stage has adaptive nodes and it is expressed as functions and if ANFIS model is Sugeno type then Equation (2.8) is valid to calculate output of this layer. This type is called first order Sugeno type (Takagi and Sugeno, 1985).

$$f_i = p_i x_i + q_i y_i + r_i \tag{2.8}$$

Here, p and q are consequent parameters and the consequent parameters are adjusted while the antecedent parameters remain fixed. Output of this stage four can be calculated using Equation (2.9).

$$O_{4,i} = \overline{\omega_i} f_i \tag{2.9}$$

- **Stage 5:** Summation neuron; this stage is a fixed node, which computes the overall output as the summation of all incoming signals by using Equation (2.10).

$$O = \sum_{n=1}^{n} \overline{\omega_i} f_i \tag{2.10}$$

Figure 2.11 presents the resultant network in case of eight inputs, every input have four membership functions, and one output also have four membership functions, in this system there are four rules. This network architecture is called as ANFIS (Adaptive Neuro-Fuzzy Inference System).

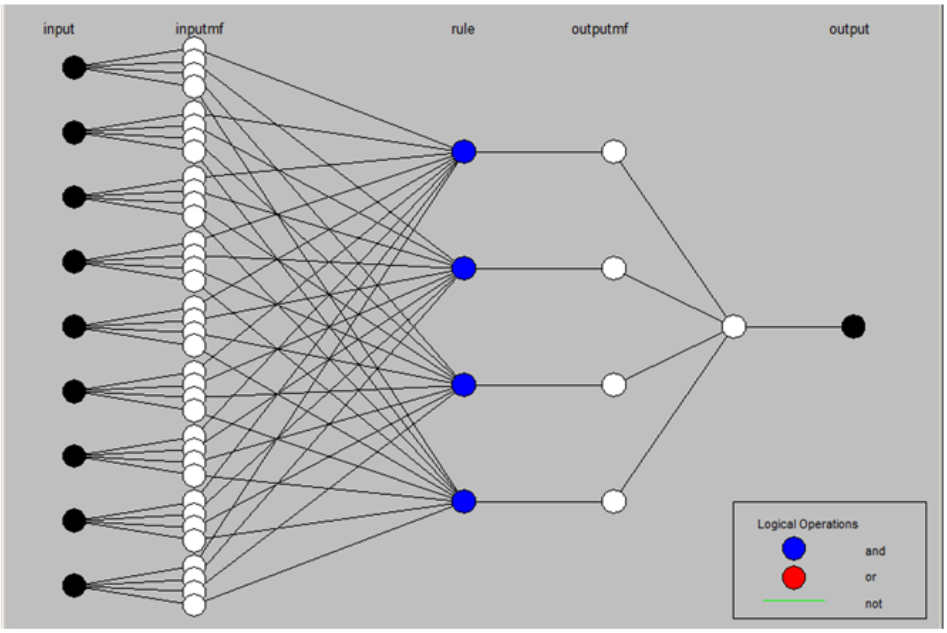


Figure 2.11: ANFIS model of fuzzy inference eight inputs every one have four membership functions, the number of rules are four.

2.5.4 Development of ANFIS Model For Emotion Dimensions Estimation

Having identified the best acoustic features set we constructed individual classifiers to estimate each emotion dimension. The speech emotion recognition system consists of two main stages (training and testing) as shown in Figure 2.12.

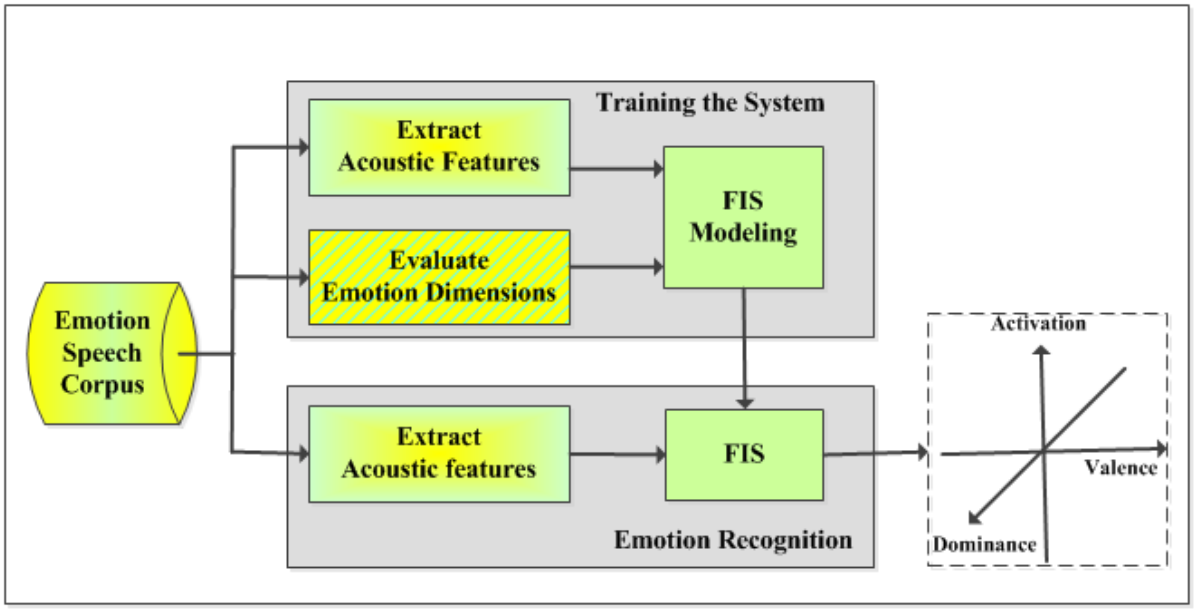


Figure 2.12: Emotion recognition system based on a two-layer model.

In order to train FIS, first, the values of emotion dimensions must be evaluated by human subjects using a listening test. Then, the parameter of the estimator FIS are learn from training it using the input acoustic features and the output emotion dimensions to build If-Then rules between them.

To estimate emotion dimensions: valence, activation ,and dominance three FISs were used. Since FIS is multi-input and one output as described in the previous section, it is required three FIS, one for each emotion dimension individually. Figure 2.13 shows the used FIS to estimate valence dimension from the most correlated acoustic features to valence. ANFIS was used to construct the used FIS.

The input for each FIS are the selected acoustic features and the output is the estimated emotion dimension.

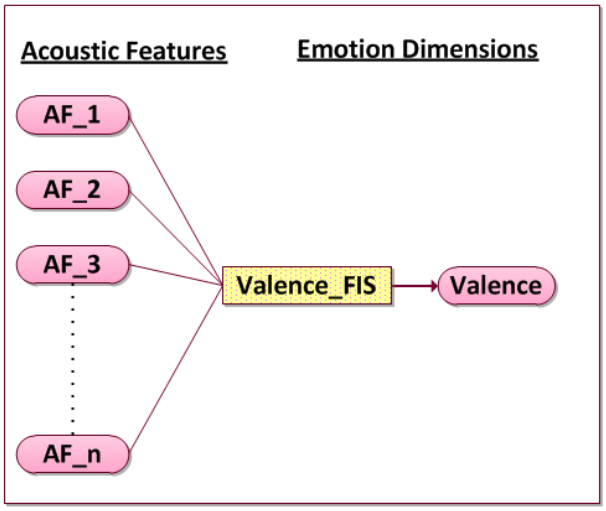


Figure 2.13: Valence dimension estimation using a two-layer model.

2.6 System Evaluation

In order to assess the performance of emotion recognition system, the mean absolute error and correlation between (the estimated values of emotion dimensions and the evaluated values by listeners). The smaller mean absolute error the closer estimated value to the human evaluation. While, the correlation between the output of the systems and the human evaluation is used to investigate the similarity between the estimated output and experimental data obtained by listening to each utterance using human subjects.

For each emotion dimension, we calculated the mean absolute error between the estimated output using the system, and the evaluated values by listeners as follows:

Let $y = \{y_i^{(j)}\}(i = 1, 2, \ldots, N)$ is the output sequence for the automatic system, where $j \in \{Valence, Activation, Dominance\}$, and let $x = \{x_i^{(j)}\}(i = 1, 2, \ldots, N)$ is the evaluated values by human subjects in the experimental evaluation as described in chapter 2. The mean absolute error $E^{(j)}$ ($j \in \{Valence, Activation, Dominance\}$) is calculated according to the following equation.

$$E^{(j)} = \frac{\sum_{i=1}^{N} |x_i^{(j)} - y_i^{(j)}|}{N} \tag{2.11}$$

The correlation coefficient $R^{(j)}$ between system output y and the human evaluation x, will be determined by the following equation.

$$R^{(j)} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{2.12}$$

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

In order to evaluate the performance of a trained classifier on new unseen data, it has to be tested with data that was not used in training stage. So far, the cross-validation method is used for evaluating the performance of emotion recognition systems, where the database is split into n parts and by iterating over all parts, always $n-1$ parts are used for training and the remaining part is used for testing. The overall recognition accuracy then results from the combination of the accuracies on all test splits. A most frequently found value for n is 10; in this thesis mostly $n = 1$ or $n = 5$ is assumed. Usually, all splits are made of equal size and it is also considered that the class distribution remains equal.

In this study two type of cross-validation are used: $n = 1$ (Leave-one-out cross validation) and $n = 5$ (5-fold cross validation). The Leave-one-out cross validation is used when the number of the used data is very small while, the 5-fold cross validation is used when we have large number of used data for training and testing.

2.6.1 Leave-one-out cross validation

Leave-one-out cross validation is used in the field of machine learning to determine how accurately a learning algorithm will be able to predict data that it was not trained on. When using the leave-one-out method, the learning algorithm is trained multiple times, using all but one of the training set data points. The form of the algorithm is as follows:

- For k = 1 to R (where R is the number of training set points). Temporarily remove the kth data point from the training set.
- Train the learning algorithm on the remaining R - 1 points.

Table 2.2: 5-Folds Cross Validation of Data

Group NO.	Training Data	Testing Data
1	f2+f3+f4+f5	f1
2	f1+f3+f4+f5	f2
3	f1+f2+f4+f5	f3
4	f1+f2+f3+f5	f4
5	f1+f2+f3+f4	f5

- Test the removed data point and note your error.
- Calculate the mean error over all R data points.

Leave-one-out cross validation is useful because it does not waste data. When training, all but one of the points are used, so the resulting classification rules are essentially the same as if they had been trained on all the data points.

2.6.2 5-fold cross validation

The 5-fold cross-validation is applied for training and testing the ANFIS classifier. The training data set is divided into 5 disjoint sets namely fold1 (f1), fold2 (f2), fold3 (f3), fold4 (f4), fold5 (f5). Furthermore, these folds are formed with the following groups for training and testing data preparation for emotion dimension estimation. Table 3.5 shows the 5-fold cross-validation of data.

2.7 Summary

This chapter introduced to basic concepts related to emotions and speech. First, different representations to describe emotions were presented. The merits of the dimensional representation are discussed therefore this representation is adopted in this study. The relationship between the two representations are discussed. Furthermore, potential sources for information on emotions were identified, which included in the speech signal. Afterwards, a general speech emotion recognition system with the three major steps of acoustic feature extraction, acoustic feature selection and classification was described. In particular, possible acoustic features that are or could be important for emotions were

discussed. The process of the traditional feature selection was presented and drawback for this approach is listed. The next chapter will introduce the database and elements of the proposed automatic emotion recognition system. Finally, the details of estimating emotion dimensions using FIS were introduced.

Chapter 3

Databases and elements of the proposed speech emotion recognition system

3.1 Introduction

This study attempt to build an automatic speech emotion recognition system that has the ability to accurately estimate emotion dimensions (valence, activation, and dominance) from a speech signal. As mentioned in the previous chapters most of the previous study estimate emotion dimensions by mapping the acoustic features into emotion dimensions directly as considered two-layer model. The conventional two-layer model has limited ability to find the most relevant acoustic features for each emotion dimension, especially valence, or to improve the prediction of emotion dimensions from acoustic features. Several studies assumed that human perception for emotional speech is a three-layer model [79, 10, 35]. Therefore, the key point to improve automatic speech emotion recognition for estimating emotion dimensions can be done by imitating humans perception process. Thus, we adopt the three-layer model for human perception as described in Figure 3.1

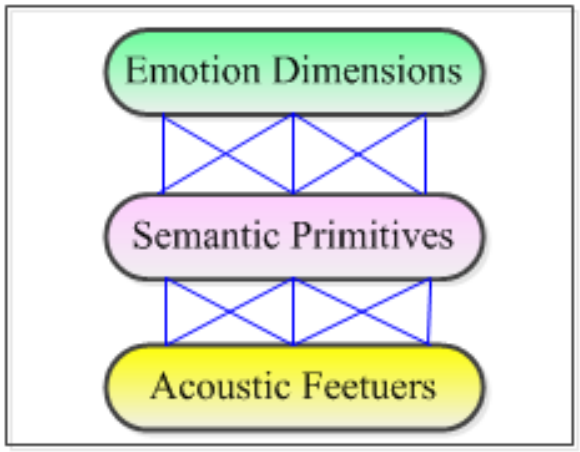


Figure 3.1: Block diagram of the three-layered model for emotion perception.

The proposed model consists of three-layer: emotion dimensions valence, activation, and dominance in the top layer, semantic primitives (adjectives describing speech) in the middle layer, while the acoustic features is in the bottom layer.

This chapter introduces the elements of the proposed emotion recognition system, ranging from the used databases, over acoustic feature extraction, to the experimental evaluation for emotion dimensions and semantic primitives using two listening tests by human subjects. The elements of this system were collected in this chapter, however, the

proposed system will be implemented in chapter 4 in details. To validate the proposed system in chapter 5, two databases were selected Japanese and German database as described in the next section. The input of the automatic system are the acoustic features, therefore the used acoustic features will be extracted in Section 3.3. The three emotion dimensions (valence, activation, and dominance) are the final output for the proposed system, a subjective evaluation was used to evaluate these dimensions using listening test experiment by human subjects for the two databases as described in section 3.5.2. Semantic primitives are adjectives describing emotional voice, this is the new layer we added between the two traditional layers: acoustic features and emotion dimensions. In this study, 17 semantic primitives are used to represent the new layer as follow: (Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow). The 17 semantic primitives are evaluated in 5-point scale by using another listening experiment as presented in Section 3.5.3. Inter-rater agreement must be done in order to exclude the subjects who have very low correlation coefficient among all subjects.

3.2 Databases

In this study, our aim is to prove a new concept proposed to improve the conventional method for automatic speech emotion recognition , not to construct a real-life application. Consequently, acted emotions are quite adequate as a testing data [68]. Therefore, in order to validate, evaluate the performance, and investigate the effectiveness of the proposed system, we used two acted databases of emotional speech: one in Japanese (single-speaker) and the other in German (multi-speaker). The first one is Fujitsu database, a multi-emotion single speaker produced and recorded by Fujitsu Laboratory. The second is Berlin database of emotional speech. The Berlin database is widely used in emotional speech recognition [11]. It is easily accessible and well annotated. Thus, by evaluating these two databases, that are described in detail in the following, a wide variety of emotions is covered and results can be expected to be general.

Table 3.1: The English translation for all 20 Japanese sentences used in Fujitsu database. The first column shows the id numbers of the sentences, the second column shows the pronunciation of the Japanese sentences in English, the third column shows the English translation for all sentences in the database.

	Japanese Sentences	English Translation
1	Atarashi meru ga todoite imasu	You've got a new mail.
2	Atama ni kuru koto nante arimasen	Theres 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	Son'na no furui meishindesu yo	That's an old superstition.
7	Min'na 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	I'd 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 dont come on time, please.
20	Sabisueria de goryu shimashou	Meet me at the service area.

3.2.1 Japanese Database

The Japanese database is the multi-emotion single-speaker Fujitsu database produced and recorded by Fujitsu Laboratory. A professional actress was asked to produce utterances using five emotional speech categories, i.e., neutral, joy, cold anger, sadness, and hot anger. In the database, there are 20 different Japanese sentences as shown in Table 3.1.

The actress speaker was asked to spoke each sentence nine times: one utterance in neutral and two utterances in each of the other four categories (joy, cold anger, sadness, and hot anger) as shown in Table 3.2. Thus, there are nine utterances for each sentence and 180 utterances for all 20 sentences. However, one cold anger utterance is missing so, the total number of utterance for Japanese database is 179.

Table 3.2: The used categories in Japanese database. The first column shows the utterances id (UID). Their are two patterns for each emotion category: Joy, Cold Anger, Hot Anger, and Sadness. And only one pattern for Neutral.

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

Table 3.3: Specification of speech data for Japanese database.

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

The detailed information of the Japanese database is shown in Table 3.3.

3.2.2 Berlin Database of Emotional Speech

The Berlin database of emotional speech was recorded at the Technical University of Berlin [11]. This database comprises of seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral speech. Ten professional German actors (five female and five male) spoke ten sentences with emotionally neutral content in the seven different emotions. Five of the ten sentences consisted of one phrase, the other five consisted of two phrases. As the recordings were intended for phonetic analysis of emotions and emotional speech synthesis they were conducted under very controlled conditions and so are marked by a very high audio quality. After the recordings a listening test was performed with 20 human subjects who should recognize the emotion of every utterance and rate it for its naturalness. The utterances in German and their translation to English

Table 3.4: The 10 utterances recorded in the Berlin database of emotional speech

	German Sentences	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 Karl.
b10	Die wird auf dem Platz sein, wo wir sie immer hinlegen.	It will be in the place where we always store it.

Table 3.5: The number of utterances for each category in the German database

Category	Anger	Boredom	Disgust	Fear	Happiness	Neutral	Sadness	All
Number	127	81	46	68	71	79	62	534

can be found in Table 3.4.

The number of utterances for each category in the German database are shown in Table 3.5. Table 3.6 shows the details information about speakers who spoke all utterances in Berlin database.

3.2.3 The selected dataset from Berlin Database of Emotional Speech

The Japanese database is inadequate for validating our emotion recognition system, because it is a single speaker database which is only suitable for speaker-specific task. To investigate the effectiveness of the proposed system for multi-speaker and different languages, a Berlin database [11] was selected. This database was selected because: (1) it is an acted-speech database the same as the Fujitsu database, (2) it contains four categories similar to those in the Fujitsu database (happy, angry, sad, and neutral), and (3) it is

Table 3.6: Information about the speakers who spoke utterances of Berlin database.

SID	Gender and age
M03	male, 31 years old
F08	female, 34 years
F09	female, 21 years
M10	male, 32 years
M11	male, 26 years
M12	male, 30 years
F13	female, 32 years
F14	female, 35 years
M15	male, 25 years
F16	female, 31 years

Table 3.7: The number of utterances for the selected categories (Anger Happiness Neutral Sadness) from the Berlin database

Category	Anger	Happiness	Neutral	Sadness	Total
Male	60	27	39	25	151
Female	67	44	40	37	188
Total	127	71	79	62	339

a multi-speaker and multi-gender database which enable us to investigate the effect of speaker and gender variation in speech emotion recognition. To compare the results of the two databases, we used only the four similar categories from Berlin database as shown in Table 3.7.

The used utterances in Berlin database were not equally distributed between the various emotional states: 69 frightened; 46 disgusted; 71 happy; 81 bored; 79 neutral; 62 sad; 127 angry, as shown in Table 3.5. Therefore, for training propose we used equally distributed between the four emotional states, and gender distribution i.e. equal number of emotional state for each gender male and female. It is found that the male sadness utterances are only 25 utterances, therefore in order to have same number of utterances for each category for male and female, we selected 25 utterances from each emotional state for male, as well as 25 utterances from each emotional state for female as shown in Table 3.8.

Finally, 50 happy, 50 angry, 50 sad, and 50 neutral; in total 200 utterances were selected form Berlin database: 100 utterances were uttered by 5 males and the other 100 by 5 females divided equally between the four emotional states. Tables 3.9 and

Table 3.8: Selected utterances from the Berlin database

Category	Anger	Happiness	Neutral	Sadness	Total
Male	25	25	25	25	100
Female	25	25	25	25	100
Total	50	50	50	50	200

Table 3.9: Selected utterances for male from Berlin database

Speaker ID	Neutral	Happy	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
	25	25	25	25	100

3.10 show the number of utterances which selected form the male and female utterances, respectively.

Table 3.11 show the distribution of the 200 utterances for the 10 sentences for both male and female utterances.

3.3 Acoustic feature analysis

To construct a speech emotion recognition system, acoustic features are needed to be investigated. In this research, the most relevant acoustic features that have been successful in related works and features used for other similar tasks were selected. Therefore, 16 acoustic features that originate from F0, power envelope, power spectrum, and duration were selected from the work by Huang and Akagi [35]. In addition to these 16 acoustic features, five new parameters related to voice quality are added, because voice quality is one of the most important cues for the perception of expressive speech. Acoustic features related to duration are extracted by segmentation, and the rest are extracted by the high quality speech analysis-synthesis system STRAIGHT [44], leading to extraction of a set of 21 acoustic features, which can be grouped into five subgroups as shown in Table 3.12.

Table 3.10: Selected utterances for female from Berlin database

Speaker ID	Neutral	Happy	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
	25	25	25	25	100

Table 3.11: Selected utterances for each sentence from Berlin database

Utterance	Male	Female	Total
a01	13	11	24
a02	15	7	22
a04	13	9	22
a05	8	13	21
a07	12	15	27
b01	5	10	15
b02	14	7	21
b03	5	9	14
b09	6	8	14
b10	9	11	20
	100	100	200

3.3.1 Segmentation and vowels information

Segmentation is needed for most of the selected acoustic features therefore, firstly we explain the segmentation process then rest of acoustic features will be explained in details in next subsections. In this study, we use two level of segmentation, the first level is phoneme level, the second is the accentual phrasal level. The smaller phrasal units is considered as an Accentual Phrases (AP). The term "accentual phrase" refer to a smaller unit from the utterance which usually terminated with a drop in F0, and sometimes with a pause. An accentual phrase is often composed of two words or more. One utterance can be uttered by different number of accentual phrases depend on the speaker and his/her emotional state.

On the first stage of our evaluation we estimated the phoneme boundaries, which were manually determined. In order to execute a vowel level analysis a phoneme level transcription is needed, which requires a corresponding lexicon containing phonetic transcription of words presented in a corpus. Unfortunately, the Fujitsu and Berlin databases does

Table 3.12: The used acoustic features.

Group	Acoustic 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.

not provide such a lexicon, so we created it by ourselves. Therefore, All utterances were manually segmented at the phoneme level. The Japanese and German language sharing 5 common vowels; /a/, /e/ /i/, /o/, and /u/. The total amounts of vowel instances presented in selected speech databases are presented in Table 3.13 and 3.14 for Japanese and German language, respectively.

For each utterance in the used databases, firstly the number of accentual phrases is counted by a listening test, then the bounders of each accentual phrases is determined.

3.3.2 F0 related features

F0 contour and power envelope varied greatly with different expressive speech categories, both for the accentual phrases as well as for the overall utterance. To measure these acoustic features, it is necessary first to separate one utterance into several accentual phrases depended on the content as explained in Section 3.3.1. Then, F0 contours were

Table 3.13: Number of vowels for each category for Fujitsu Database.

	a	e	i	o	u
Neutral	92	30	61	43	26
Joy	184	60	122	86	52
Cold Anger	181	59	118	83	49
Sad	184	60	122	86	52
Hot Anger	184	60	122	86	52
	825	269	545	384	231

Table 3.14: Number of vowels for each category for Berlin Database.

	a	e	i	o	u
Neutral	161	79	219	44	30
Joy	168	82	223	46	33
Anger	164	71	218	38	29
Sad	189	122	248	51	45
	682	354	908	179	137

measured for accentual phrases within the utterance as well as for the entire utterance using STRAIGHT [44]. For each utterance we extract four acoustic features related to F0. The definition of these acoustic features are as follows:

- **F0_HP:** Highest F0 is the maximum of F0 for each utterance.
- **F0_AP:** Average F0 is the mean value of F0 for each utterance.
- **F0_RS1:** Rising slope of the first accentual phrase is the slope of a regression line that fits the raising part of first accentual phrase which is from the start point of first accentual phrase to the highest point of the first high accent part.
- **F0_RS:** Rising slope of the entire utterance is mean value of all rising slopes for the accentual phrases within the utterance.

3.3.3 Power envelope related features

Power envelope was measured in a similar way to that for the F0 contour. All these acoustic features were measured using STRAIGHT analysis. For each utterance, the measurements were as follows:

- **PW_R:** The power range which is the difference between the maximum and the minimum power in each utterance.

- **PW_RHT:** The ratio between the average power in the high frequency portion (over 3 kHz) and the average power.

- **PW_RS1:** Rising slope of the first accentual phrase is the slope were measured based on the same concept of that for measuring F0 acoustic features.

- **PW_RAP:** The mean value of the power range for the accentual phrases within each utterance.

3.3.4 Power spectrum related features

Formants measures were the mean value of (first formant frequency (SP_F1), second formant frequency (SP_F2), third formant frequency (SP_F3) taken approximately at the midpoint of the vowels /a/,/e/,/i/,/o/, and /u/. All utterances were manually segmented at the phoneme level. After segmentation, formants were extracted for the vowels /a/, /i/, /e/,/o/ and /u/ using PRAAT speech analysis software with standard settings [9]. These include the maximum number of formants tracked (five), the maximum frequency of the highest formant (set to 5000 for male and 5500 for female speakers), the time step between two consecutive analysis frames (0.01 seconds), the effective duration of the analysis window (0.025 seconds) and the amount of pre-emphasis (50 Hz). These settings generally resulted in acceptable results. Formant measures were taken approximately at the vowel midpoint of the vowels /a/, /e/, /i/, /o/, and /u/. Finally, for spectrum, we used formants, spectral tilt, and spectral balance.

- **SP_F1:** The mean value of first formant frequency taken approximately at the midpoint of all vowels for each utterance.

- **SP_F2 and SP_F3:** Are the mean value of second and the third formant frequency, respectively. They were calculated in a similar way as the first formant frequency

- **SP_TL:** Spectral tilt is used to measure voice quality and it was calculated from $(A1 - A3)$, where A1 is the level in dB of the first formant and A3 is the level of the harmonic whose frequency is closest to the third formant.

- **SP_SB:** Spectral balance, this parameter serves for the description of acoustic consonant reduction, and was calculated according to the following equation:

$$SP\_SB = \frac{\Sigma_i |S(f_i)|.f_i}{\Sigma_i |S(f_i)|} \tag{3.1}$$

  where $|S(f_i)|$ is the amplitude of the spectrum and f_i is the frequency in Hz.

3.3.5 Duration related features

The speed of the speech utterance or the length varies greatly when a speaker utter the same word in different emotional states, especially for vowels. For example, when some one in anger or happy state he/she speak very fast, while when someone in sad emotional state the speed of speaking is very slow. For each sentence, the duration of all phonemes, both consonants and vowels, were measured as described in 3.3.1. The duration measurements were as follows:

- **DU_TL:** The total length, is the duration of the entire utterance, calculated from the segmentation data.

- **DU_CL:** Consonant length, is the summation of the durations of all consonant within the utterance,

- **DU_RCV:** the ration of consonant duration to vowel duration is calculated as follow we first calculate Consonant length DU_CL as done in previous step, then we calculate the summation of the durations of all vowel within the utterance DU_VL. Finally, the DU_RCV is DU_CL divided by DU_VL.

3.3.6 Voice quality related features

Voice quality is one of the most important cues for the perception of emotional speech. Voice quality conveys both linguistic and paralinguistic information, and can be distinguished by acoustic source characteristics. The emotional state of the speaker is reflected in his or her vocal utterances. Listeners know this and are able to recognize emotional states based on vocal cues alone [27].

Voice quality is primarily associated with the spectral properties of the speech signal. Spectral shape can provide cues to relevant aspects of voice quality, such as $H1-H2$ and $H1-A3$ where H1, H2, A3 are the amplitudes in (dB) of the first harmonic, the second harmonic, and the level of the harmonic whose frequency is closest to the third formant, respectively as shown in Figure 3.2.

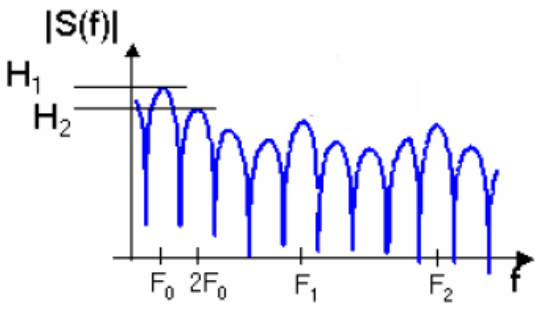


Figure 3.2: Speech spectrum in dB, showing harmonics H1, H2.

$H1-A3$ reflects glottal cycle characteristics, i.e., speed of closing of vocal folds while $H1-H2$ is concerned with glottal opening [50, 34]. H1 and H2 are related to the open quotient (OQ) using Equation (3.2) as reported by Fant in [26].

$$H1-H2 = -6 + 0.27exp(5.5 \times OQ) \tag{3.2}$$

In this study, we focus on harmonics amplitude H1 and H2 as a spectral property of the speech signal that reflects voice quality. Finally, $H1-H2$ has been used as an indication for voice quality. In order to investigate the effect of voice quality for each

vowel individually, the mean value of $H1 - H2$ for vowel /a/,/e/,/i/,/o/, and /u/ per utterance are used as an indication for voice quality and calculated as follows:

For evaluation we use an average $H1 - H2$ value extracted from vowel segments. Harmonic amplitudes were computed pitch-synchronously, using standard optimization techniques to find the maximum of the spectrum around the peak locations as estimated by F0. Therefore, F0 was extracted by using STRAIGHT [44], Harmonic structure is determined through spectral analysis using FFT then H1 and H2 are calculated through integer multiplication of the F0 value obtained from the STRAIGHT.

Figure 3.3 shows the trajectories of H1, H2 for vowels segment, for four emotional states: neutral, joy, hot anger and sad utterances taken from the Fujitsu database. The four Figures contains words (lexical content) bout with different four emotional speech categories, however, the shape varied greatly in the trajectories of H1, H2, therefore, $H1 - H2$ is very important factor for classification of the emotional state.

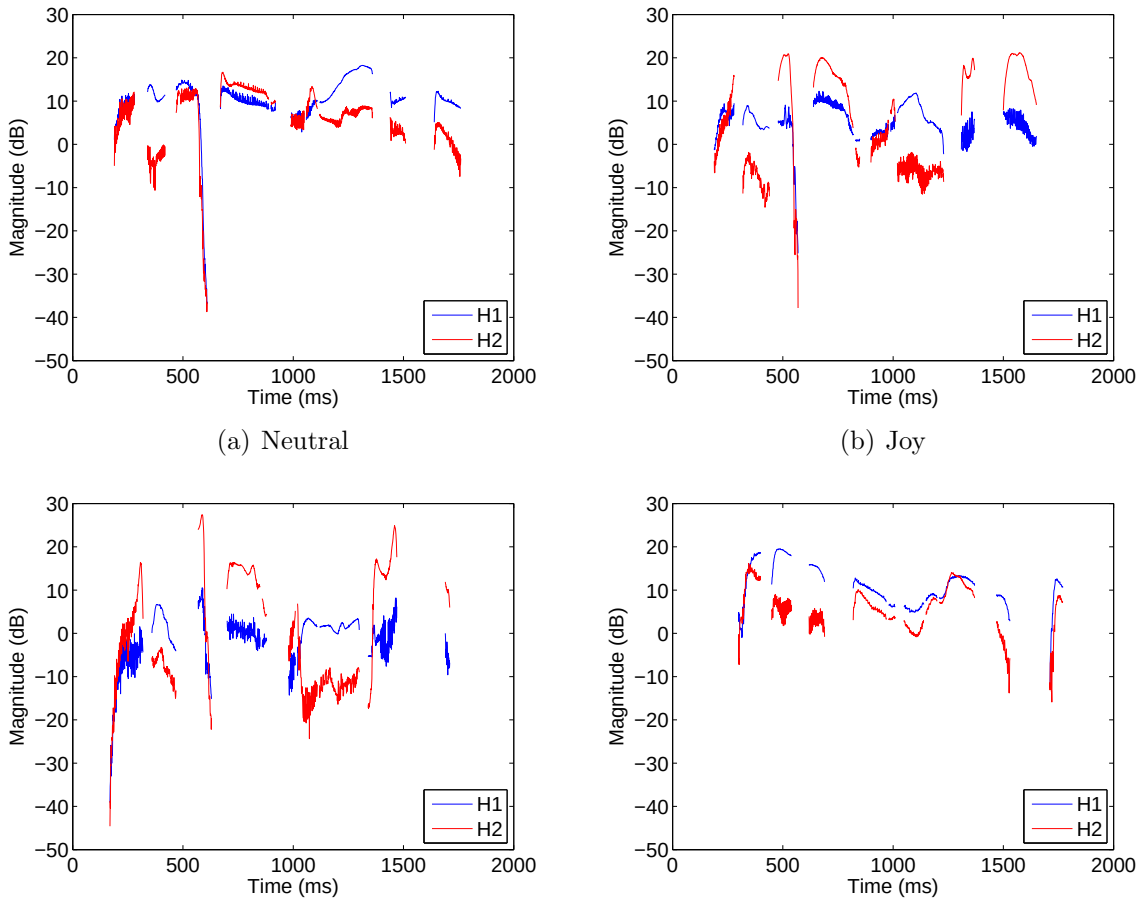
Finally, the mean value of $H1 - H2$ for vowel /a/,/e/,/i/,/o/, and /u/ per utterance MH_A, MH_E, MH_I, MH_O, and MH_U, respectively are used as an indication for voice quality, where MH_j can be calculated using:

$$MH\_j = \frac{1}{t_k^{(j)}} \sum_{i=1}^{t_k^{(j)}} (H1_i^{(j)} - H2_i^{(j)}) \tag{3.3}$$

where $j \in \{A, E, I, O, U\}$, and t_k is a number of discrete estimations of first harmonics values within a vowel segment, $H1_i$, $H2_i$ are an estimation of first and second harmonic value respectively at discrete time i. we estimated the average of $H1 - H2$ values for each vowel individually.

3.4 Normalization

The values of acoustic features greatly changed for emotionally colored and neutral speech samples, for example, it was found a significant difference for the vowel triangles form and



(a) Neutral (b) Joy

(c) Hot Anger (d) Sad

Figure 3.3: The trajectories of H1, H2 for vowels segment. Emotion relevant to H1, H2 acoustic features shown for a neutral, a joy, hot anger and a sad utterance taken from the Fujitsu database of emotional speech. The text spoken in each of the utterances was "Arigato wa iimasen." ("I wont say thank you.").

their position in F1-/F2-dimensional space for emotionally colored and neutral speech samples [88]. Another problem that arises when measuring formants is that vocal tracts are different among different people, i.e. formant frequency is speaker-dependent and emotion-dependent acoustic feature.

All the 21 acoustic features were extracted for both Fujitsu and Berlin databases. In order to justify the acoustic features to avoid speaker-dependency and emotion-dependency on the acoustic features. We adopt a novel a acoustic feature normalization method, in which all acoustic feature values are normalized by those of the neutral speech. This was performed by dividing the values of acoustic features by the mean value of neutral utterances for all acoustic features for 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:

$$\hat{f}_i = \frac{f_i}{(\sum_{i=1}^{K} f_i/K)} \tag{3.4}$$

3.5 Experimental evaluation for emotion dimensions and semantic primitives

Having extracted the acoustic features for each utterance which are the inputs to the automatic emotion recognition system, it is also required to subjectively evaluate semantic primitives and emotion dimensions for each utterance using human subjects. Therefore, all utterances must be labeled in term of emotion dimensions and semantic primitives.

Emotion dimensions are the elements of the top layer of the proposed model and they are the outputs of the proposed automatic emotion recognition system. Most exciting emotional speech databases have been annotated using the categorical approach, while, few databases have been annotated using the dimensional approach [89]. The Fujitsu and Berlin databases are categorical databases. Therefore, listening tests are required to annotate each utterance in the used databases using the dimensional approach. Thus, the two databases were evaluated by the listening tests along three dimensions: valence, activation, and dominance, as described in Section 3.5.2

The semantic primitives were used as the the middle layer between acoustic features and emotion dimension therefore, values of each semantic primitive for all utterances must be evaluated. Section 3.5.3 shows the listening experiment which was conducted for each database using human subjects to label each utterance along the 17 semantic primitives.

Table 3.15: The number of subjects who labeled the two databases.

Used database	Subjects	Gender
Japanese database	11	9M 2F
German Database	9	8M 1F

Table 3.16: The Stimuli used for experimental evaluation.

Used database	Stimuli	Emotion Categories
Japanese database	179	5 Categories:Neutral, Joy, Cold Anger, Sad, Hot Anger
German Database	200	4 Categories:Neutral, Joy, Sad, Anger

3.5.1 Human subject evaluation

Two listening tests were used to evaluate semantic primitives and emotion dimensions. From the previous studies, no agreement about how many subjects must be used for conducting a listening test for evaluating database in term of emotion dimensions. For example, some researcher conducted a listening test by only 2 subjects such as Vidrascu and Devillers [14] or, at most, 4 to 5 subjects such preformed by Lee and Narayanan in [47]. Moreover, Grimm in [31] preformed a three listening tests using 18, 17, and 6 subjects for three different databases in his research corpora, respectively.

The number of subjects used to evaluate the two databases are listed in Table 3.15, moreover, stimuli are presented in Table 3.16. The Fujitsu database was evaluated by 11 graduate students, all native Japanese speakers (nine male and two female). While Berlin database was evaluated using nine graduate students, all native Japanese speakers (eight male and one female). No subjects have hearing impairments. In comparison with other studies on emotion recognition which included 2 to 5 independent subjects, we used a much higher number of evaluators to gain statistical confidence.

Subjects for listening tests were 11 graduate students, native Japanese speakers without any hearing troubles evaluate the Japanese database. Due to of graduation of 5 subjects, therefore, we asked a anther three subjects to participate in evaluation for German. In total nine subjects evaluate the German database six who participate in the evaluation of Japanese database and three new subjects who did not participate in Japanese database. Japanese subjects were selected for both experiment because these excrements

measure how speaker express emotion by voice not what the speaker said i.e. without understating the content of the speech utterance.

It is very important to conduct statistical analysis for the subjects ratings to obtain a much more reliable and rich annotation. In this study, an inter-rater agreement was conducted between subjects ratings, then, ratings evaluations from subjects' who agree with high degree were selected. Finally, the average of all agreed subjects is used as final label for each utterance.

3.5.2 Emotion Dimensions Evaluation

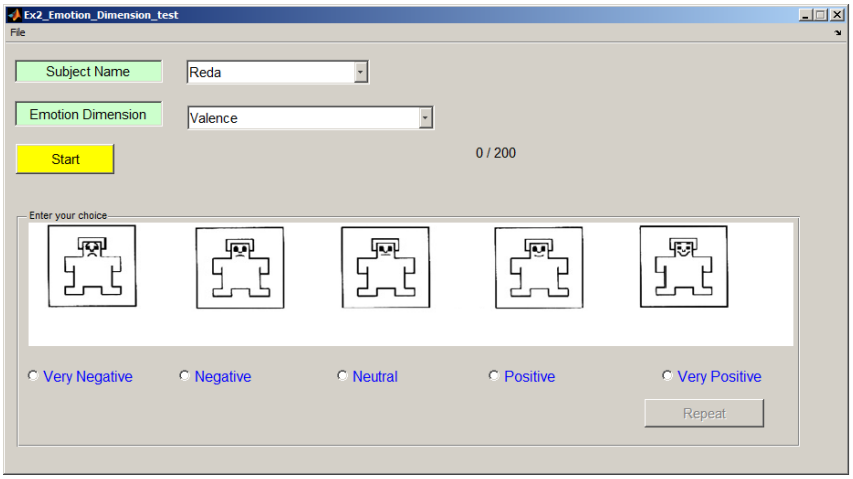
The purpose of this experiment is to evaluate the used databases in term of the degree of each emotion dimensions valence, activation, and dominance. For each utterance, the three values of emotion dimensions: (valence, activation, dominance) represent the emotional state as one point in emotional space. These values not only represent the emotional state but also the degree of emotional state. Emotion dimensions evaluation by human subjects will be used as a reference to evaluate the performance of the automatic emotion recognition systems later.

For emotion dimensions evaluation, a 5-point scale {-2, -1, 0, 1, 2} was used: valence (from -2 very negative to +2 very positive), activation (from -2 very calm to +2 very exited), and dominance (from -2 very weak to +2 very strong). The subjects used a MATLAB Graphical User Interface (GUI) in this experiment to evaluate the stimuli, as shown in Figure 3.4.

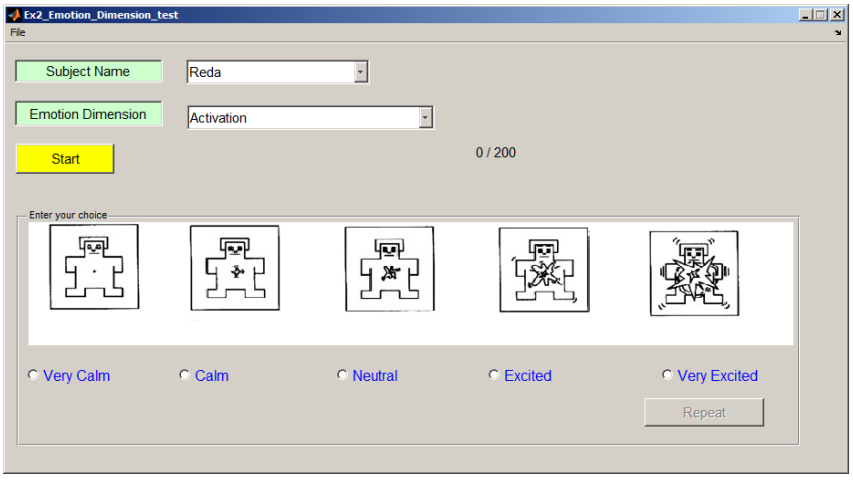
They were asked to evaluate one emotion dimension for the whole database in one session. There were three sessions, one for each emotion dimension. As done in the work of Mori et al. [49] for emotion dimension evaluation, the basic theory of emotion dimension was explained to the subjects before the experiment started. Then they took a training session to listen to an example set composed of 15 utterances, which covered the used five-point scale, three utterances for each point in the used scale. In this test, the stimuli were presented randomly, for each utterance. Subjects were asked to evaluate their perceived impression from the way of speaking, not from the content itself, and then choose score on the five-point scale for each dimension individually, repetition was allowed. Finally, the average of the subjects rating for each emotion dimension was calculated per utterance.

3.5.2.1 Agreement Between Subjects

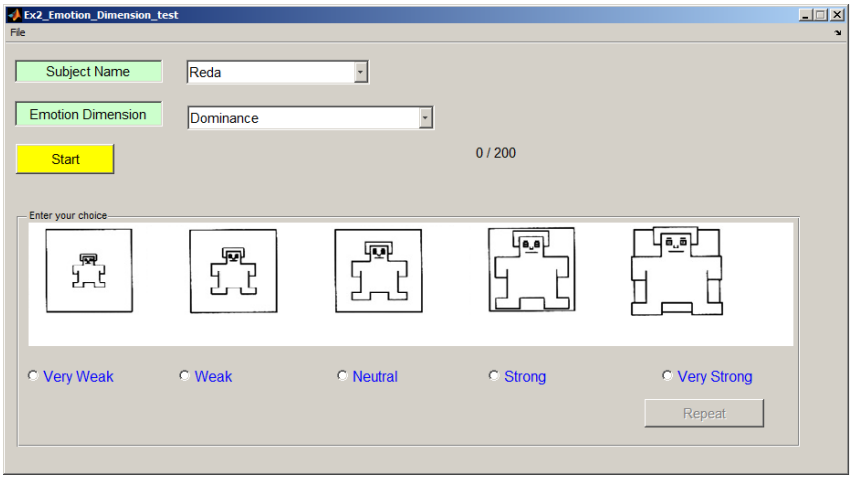
Inter-rater agreement must be done in order to exclude the subjects who have very low correlation coefficient among all subjects. The inter-rater agreement was measured by



(a) MATLAB GUI for valence evaluation.



(b) MATLAB GUI for activation evaluation.



(c) MATLAB GUI for dominance evaluation.

Figure 3.4: MATLAB GUI for evaluating emotion dimensions.

Table 3.17: Pairwise correlations of rated valence dimension of each utterance, demonstrating the degree of inter-rater agreement between subjects for the listening test.

(a) Japanese Database using 11 subjects

Valence	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11
S01	1	0.91	0.91	0.88	0.92	0.92	0.86	0.92	0.91	0.93	0.90
S02		1	0.92	0.89	0.94	0.92	0.87	0.90	0.92	0.94	0.89
S03			1	0.87	0.94	0.92	0.87	0.91	0.91	0.94	0.90
S04				1	0.89	0.88	0.84	0.86	0.86	0.88	0.84
S05					1	0.92	0.88	0.92	0.94	0.95	0.92
S06						1	0.86	0.90	0.92	0.95	0.92
S07							1	0.86	0.87	0.87	0.85
S08								1	0.89	0.91	0.88
S09									1	0.94	0.90
S10										1	0.91
S11											1

(b) German Database using 9 subjects

Valence	S01	S02	S03	S04	S05	S06	S07	S08	S09
S01	1	0.83	0.83	0.82	0.82	0.82	0.78	0.74	0.85
S02		1	0.86	0.86	0.86	0.86	0.78	0.82	0.85
S03			1	0.85	0.86	0.84	0.82	0.82	0.86
S04				1	0.87	0.85	0.82	0.82	0.89
S05					1	0.87	0.84	0.84	0.86
S06						1	0.80	0.82	0.84
S07							1	0.79	0.84
S08								1	0.81
S09									1

means of pairwise Pearson's correlations between two subjects' ratings, for each emotion dimension, separately. For example Table 3.17 listed the correlations between every pair of subjects for the ratings evaluations of valence dimension for both Japanese and German database.

Tables 3.18(a) and 3.18(b) summarize the minimum, maximum, and average of Pearson's correlations between every two subjects' ratings for all emotion dimensions, for the two database. It was found that all subjects agreed to high degree for all emotion dimension evaluation, for the two databases, as shown in Tables 3.18(a) and 3.18(b) for Japanese and German database, respectively. For Japanese database, the average of Pearson's correlation coefficient among every pairs of two subjects were as follows: 0.90, 0.85, and 0.89 for valence, activation, and dominance, respectively, moreover, for German

Table 3.18: Minimum (Min), Maximum (Max) and Average (Ave) for the correlation coefficients between subjects ratings for evaluating emotion dimensions.

(a) Japanese database

	Min	Max	Ave
Valence	0.84	0.95	0.90
Activation	0.75	0.94	0.85
Dominance	0.79	0.98	0.89

(b) German database

	Min	Max	Ave
Valence	0.74	0.89	0.83
Activation	0.73	0.93	0.87
Dominance	0.79	0.92	0.86

database 0.83, 0.87, and 0.86 for valence, activation, and dominance, respectively. This indicate that all subjects agreed to a high degree for all emotion dimension evaluation, for the two databases.

3.5.3 Evaluations of Semantic Primitives

The propose of this experiment is to evaluate the used databases in term of semantic primitives (adjectives) which is the middle layer of the proposed model. Semantic primitives are required as the bridge between the acoustic features and the emotion dimensions in the proposed model. These evaluation values for semantic primitives will be used to find which acoustic features are related to each emotion dimension.

In this study, the human perception model as described by Scherer [79] is adopted. This model assumes that human perception is a three-layer process. It was assumed that the acoustic features are perceived by a listener and internally represented by a smaller perception e.g. adjectives describing emotional voice as reported by Huang and Akagi [35]. In this study 'smaller perception' means an earlier process of perception. These smaller percepts or adjectives are finally used to detect the emotional state of the speaker. These adjectives can be subjectively evaluated by human subjects. Therefore, the following set of adjectives describing the emotional speech were selected as candidates for semantic primitives: Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow. These adjectives were selected because they reflect a balanced selection of widely used adjectives that describe emotional speech. They are originally from the work of Huang and Akagi [35].

For the evaluation, we used listening tests. In these tests, the stimuli were presented

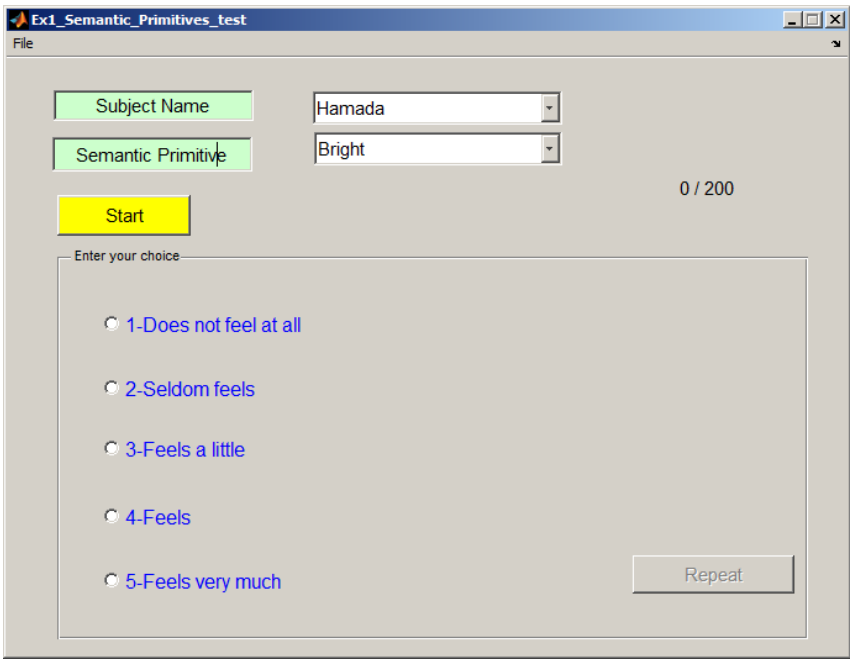


Figure 3.5: MATLAB GUI used for Semantic Primitives evaluation experiment.

randomly to each subject through binaural headphones at a comfortable sound pressure level in a soundproof room. Subjects were asked to rate 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" as shown in the MATLAB GUI used for this experiment in Figure 3.5. The 17 semantic primitives were evaluated for the two databases, and then ratings of the individual subject were averaged for each semantic primitive per utterance.

3.5.3.1 Inter-rater agreement

The inter-rater agreement was measured by means of pairwise Pearson's correlations between two subjects' ratings, separately for each semantic primitive. Tables 3.19(a) and 3.19(b) summarize the minimum, maximum, and average of Pearson's correlations between every two subjects' ratings for all semantic primitives, for the two database. The results for the two database is as follows: for Japanese database, the average of Pearson's correlation among every pairs of two subjects for all semantic primitives evaluation were ranged between 0.68 and 0.85, moreover, for German database, the average of correlations were ranged between 0.66 and 0.86. This result suggests that all subjects agreed from a

Table 3.19: Minimum (Min), Maximum (Max) and Average (Ave) for the correlation coefficients between subjects ratings for evaluating semantic primitives.

(a) Japanese database

	Min	Max	Ave
Bright	0.64	0.99	0.83
Dark	0.58	0.95	0.79
High	0.51	0.95	0.77
Low	0.62	0.94	0.81
Strong	0.63	0.96	0.84
Weak	0.66	0.91	0.80
Calm	0.61	0.97	0.77
Unstable	0.57	0.94	0.80
Well-modulated	0.67	0.93	0.81
Monotonous	0.60	0.95	0.76
Heavy	0.58	0.92	0.76
Clear	0.57	0.93	0.76
Noisy	0.64	0.93	0.79
Quiet	0.69	0.99	0.85
Sharp	0.54	0.92	0.76
Fast	0.54	0.84	0.70
Slow	0.53	0.83	0.68

(b) German database

	Min	Max	Ave
Bright	0.57	0.83	0.71
Dark	0.52	0.89	0.73
High	0.52	0.93	0.80
Low	0.55	0.88	0.77
Strong	0.73	0.92	0.85
Weak	0.72	0.93	0.86
Calm	0.57	0.92	0.77
Unstable	0.25	0.89	0.69
Well-modulated	0.63	0.89	0.78
Monotonous	0.36	0.95	0.74
Heavy	0.45	0.95	0.71
Clear	0.51	0.94	0.78
Noisy	0.17	0.89	0.72
Quiet	0.59	0.88	0.80
Sharp	0.49	0.84	0.72
Fast	0.53	0.78	0.66
Slow	0.53	0.83	0.71

moderate to a very high degree.

3.6 Summary

In this chapter, the elements of the proposed automatic emotion recognition system were collected. The two acted emotional speech databases were selected to validate the proposed system and prove our concept in this study, these database are databases Fujitsu and Berlin database as described in details in Section 3.2.1 and Section 3.2.2, respectively. The proposed three-layer model consists of emotion dimensions, semantic primitives, acoustic features. In order to construct the three-layer model, all of these elements must be evaluated objectively or subjectively. Firstly, 21 acoustic features were selected from the literature of emotional speech as an initial set of acoustic features. These acoustic features were extracted as described in Section 3.3. Emotion dimension in this study are valence, activation, and dominance which subjectively evaluated using a listening experiment as described in Section 3.5.2. The used semantic primitives are also subjectively

evaluated using a listening experiment as described in Section 3.5.3. Chapter 4 investigates which acoustic features of the selected set will be mostly related to each emotion dimension.

Chapter 4

The proposed speech emotion recognition system

4.1 Introduction

The ultimate goals of this chapter, are (1) attempt to answer the most challenging question for speech emotion recognition, what are the most related acoustic features for each emotion dimensions?, (2) to improve the predict of emotion dimensions values by constructing a speech emotion recognition system based on the process of human perception. Chapter 3 introduces the first step towards building emotion recognition system by extracting the initial set of acoustic features. In this chapter, we introduce the second and the third step, which are: selecting the most related acoustic features for each emotion dimensions, and estimating emotion dimensions based on the proposed method.

As motioned in Chapter 1 that most of the previous studies for emotion dimensions estimation were based on a two-layer model i.e. acoustic feature layer and emotion dimension layer. Using the two-layer model, the acoustic feature selection was based on the correlation between acoustic features and emotion dimension. The acoustic features correlated to valence dimension were very few, very weak, and inconsistent. Due to these limitation, the prediction of valence dimension very difficult to be estimated from acoustic features only. Thus, the conventional two-layer model has limited ability to find the most relevant acoustic features for each emotion dimension, especially valence, or to improve the prediction of emotion dimensions from acoustic features. This model does not imitate human perception, this is reason behind the poor estimation of valence dimension.

In this thesis, the proposed idea to improve the prediction of emotion dimensions can be done by imitating the process of human perception for recognizing the emotional state from a speech signal. Therefore, to overcome these limitations, this study proposes a three-layer model to improve the estimating values of emotion dimensions from acoustic features based on human perception described by [10, 79, 35]. Our proposed model consists of three layers: emotion dimensions (valence, activation, and dominance) constitute the top layer, semantic primitives the middle layer, and acoustic features the bottom layer as described in Figure 4.1. A semantic primitive layer is added between the two conventional layers acoustic features and emotion dimensions.

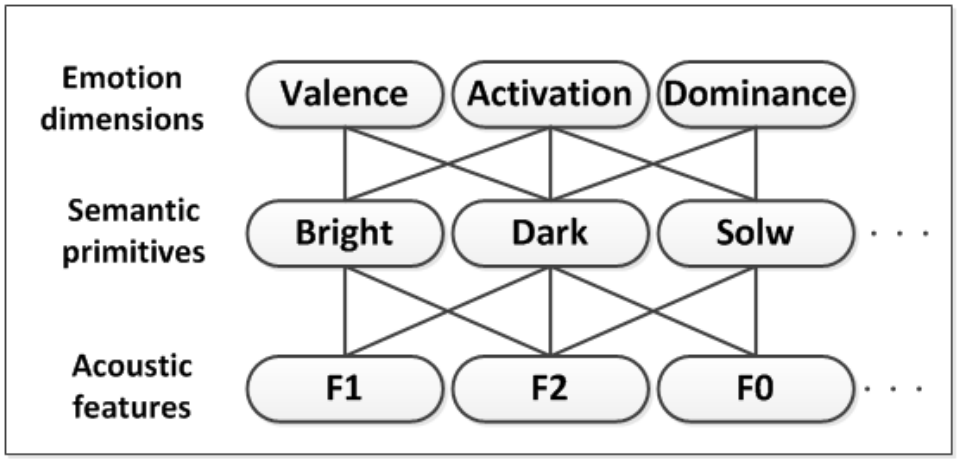


Figure 4.1: The three layer model.

To answer the first question, first, we assume that the acoustic features that are highly correlated with semantic primitives will have a large impact for predicting values of emotion dimensions, especially for valence. This assumption can guide the selection of new acoustic features with better discrimination in the most difficult dimension. Thus, a novel top-down feature selection method is proposed to select the most related acoustic features based on the three-layer model.

The selection procedures of the traditional two-layer method was compared with that of the proposed method. To accomplish this task, the most related acoustic features are investigated using; the traditional method as described in Section 4.2, and using the proposed method as introduced in Section 4.3.

The second issue to be solved in this chapter, is to improve the exciting emotion recognition system in order to accurately estimate emotion dimensions from acoustic features. Having select the most related acoustic features for each emotion dimension, we build an emotion recognition system using a bottom-up method to estimate emotion dimensions form acoustic features by imitating the human perception process. Firstly, estimating semantic primitives from acoustic features, then estimating emotion dimensions from the estimated semantic primitives as presented in Section 4.4.

4.2 The traditional method for acoustic features selection

This section, introduce the traditional method for selecting the acoustic features for each emotion dimension based on the two-layer model. The traditional method for acoustic features selection were based on correlations between acoustic features and emotion dimension as a two-layer model. For example in [23] they used the subset evaluation method to find the subsets of acoustic features that have high correlation with emotion dimensions.

In order to investigate the relationship between acoustic features and emotion dimensions by using the traditional two-layered model, the correlations coefficients between extracted parameter values for each acoustic feature and evaluated scores of each dimension are calculated as follows: Let $f_m = \{f_{m,n}\}(n = 1, 2, \ldots, N)$ be the sequence of values of the m^{th} acoustic feature, $m = 1, 2, \ldots, M$, where M is the number of extracted acoustic features in this study ($M = 21$ acoustic features as described in Chapter 3). Moreover, let $x^{(i)} = \{x_n^{(i)}\}(n = 1, 2, \ldots, N)$ be the sequence of values of the i^{th} emotion dimension, $i \in \{Valence, Activation, Dominince\}$, where N is the number of utterances in our databases ($N = 179$ for Japanese, $N = 200$ for German). Then the correlation coefficient $R_m^{(i)}$ between the acoustic parameter f_m and the emotion dimension $x^{(i)}$ can be determined by the following equation:

$$R_m^{(i)} = \frac{\sum_{n=1}^{N}(f_{m,n} - \overline{f_m})(x_n^{(i)} - \overline{x}^{(i)})}{\sqrt{\sum_{n=1}^{N}(f_{m,n} - \overline{f_m})^2}\sqrt{\sum_{n=1}^{N}(x_n^{(i)} - \overline{x}^{(i)})^2}} \tag{4.1}$$

where $\overline{f_m}$, and $\overline{x}^{(i)}$ are the arithmetic mean for the acoustic feature and emotion dimension respectively.

Tables 4.1 and Table 4.2 show the correlation coefficients for all acoustic features and all emotion dimensions for Japanese and German database, respectively. For Japanese language from Table 4.1, it is evident that eight acoustic features have high correlation with the activation and dominance dimensions as demonstrated by the absolute value of

the correlation, which was greater than 0.45 as shown in bold in this table. Furthermore, the emotion dimension valence shows a smaller absolute values of correlations than the activation and dominance. For German language from Table 4.2, it is evident that eight acoustic features have high correlation with the activation and dominance dimensions. For both Japanese and German language, For both Japanese and German language, valence shows a smaller absolute values of correlations than the activation and dominance. These results are consistent with many previous studies [29, 76]. The poor correlation between the acoustic features and valence is the reason behind the very low performance for valence estimation using the traditional approach.

The correlations between acoustic features and emotion dimensions for Japanese and German database have similar trend with the valence dimension, only one acoustic feature was found to be correlated with valence. The reason for weak correlations between acoustic features and valence dimensions is that the acoustic effects of anger and joy are very similar as reported in many studies, though these emotions are usually not hard to distinguish for human. Therefore, the acoustic feature selection using the two-layer model was failed to select a discriminate acoustic feature for valence dimension. Thus, the traditional method can not answer this question yet: which acoustic features can be used to estimate the valence dimension?. Therefore, this study suggest to adopt the three-layer model which imitate human perception for acoustic feature selection as described in the next section.

Table 4.1: Japanese Database: Correlation coefficients between acoustic features (AF) and emotion dimensions (ED).

m	AF \ ED	Valence	Activation	Dominance	#
1	MH_A	-0.23	**-0.85**	**-0.83**	2
2	MH_E	-0.10	**-0.56**	**-0.57**	2
3	MH_I	0.27	-0.03	-0.17	0
4	MH_O	0.13	**0.76**	**0.75**	2
5	MH_U	0.08	-0.17	-0.23	0
6	F0_RS	0.34	**0.78**	**0.65**	2
7	F0_HP	0.29	**0.77**	**0.64**	2
8	F0_AP	-0.08	-0.11	-0.12	0
9	F0_RS1	-0.09	-0.16	-0.17	0
10	PW_R	0.24	**0.53**	**0.50**	2
11	PW_RHT	**-0.47**	0.33	0.37	1
12	PW_RS1	-0.02	-0.22	-0.23	0
13	PW_RAP	0.17	0.39	0.36	0
14	SP_F1	-0.10	0.30	0.28	0
15	SP_F2	-0.10	0.09	0.11	0
16	SP_F3	0.01	0.33	0.36	0
17	SP_TL	0.40	0.29	0.27	0
18	SP_SB	0.05	0.40	0.36	0
19	DU_TL	-0.12	-0.30	-0.31	0
20	DU_CL	-0.27	**-0.61**	**-0.58**	2
21	DU_RCV	-0.30	**-0.61**	**-0.57**	2
	#	1	8	8	17

Table 4.2: The correlation coefficients between the acoustic features and the emotion dimensions for German Database.

m	ED / AF	Valence	Activation	Dominance	#
1	MH_A	-0.33	**-0.82**	**-0.81**	2
2	MH_E	-0.18	**-0.70**	**-0.71**	2
3	MH_I	-0.03	-0.19	-0.24	0
4	MH_O	-0.28	**-0.67**	**-0.68**	2
5	MH_U	-0.25	**-0.47**	**-0.47**	2
6	F0_RS	0.21	**0.69**	**0.65**	2
7	F0_HP	0.19	**0.59**	**0.54**	2
8	F0_AP	-0.05	-0.14	-0.13	0
9	F0_RS1	-0.05	-0.10	-0.09	0
10	PW_R	0.23	**0.75**	**0.74**	2
11	PW_RHT	-0.25	0.44	**0.49**	1
12	PW_RS1	0.08	0.14	0.14	0
13	PW_RAP	0.08	0.36	0.35	0
14	SP_F1	**-0.55**	**-0.49**	-0.43	2
15	SP_F2	-0.03	-0.29	-0.29	0
16	SP_F3	-0.04	-0.04	0.01	0
17	SP_TL	0.28	0.26	0.26	0
18	SP_SB	-0.02	-0.05	-0.02	0
19	DU_TL	-0.28	-0.38	-0.39	0
20	DU_CL	-0.24	-0.36	-0.36	0
21	DU_RCV	-0.14	-0.39	-0.37	0
	#	1	8	8	17

4.3 Selection of Acoustic Features and Semantic Primitives

This section describes the proposed acoustic features selection method to identify the most relevant acoustic features for emotion dimensions valence, activation, and dominance. For this purpose, we proposed a thee-layer model that imitates the human perception to understand the relationship between acoustic features and emotion dimensions.

4.3.1 Selection Procedures

Our selection method is based on the following assumptions: 1) semantic primitives which are highly correlated with the emotion dimension are given large impact in the estimation of that dimension, and 2) acoustic features which are highly correlated with the semantic primitive are given large impact in the estimation of that semantic primitive. In this study, we consider the correlation highly correlated if its absolute value is grater than or equal to 0.45. To accomplish this task, the top-down method shown in Fig 4.2 was used as follows:

- **Step (1):** Calculating the correlation coefficients between each emotion dimension (top-layer) and each semantic primitives (middle layer).
- **Step (2):** Selecting the highly correlated semantic primitives for each emotion dimension.
- **Step (3):** Calculating the correlation coefficients between each selected semantic primitive (middle layer) in step 2 and each acoustic feature (bottom layer).
- **Step (4):** Selecting the highly correlated acoustic features for each semantic primitive.

For each emotion dimension, the selected acoustic features in the final step are considered as the most relevant features to the dimension in the top-layer.

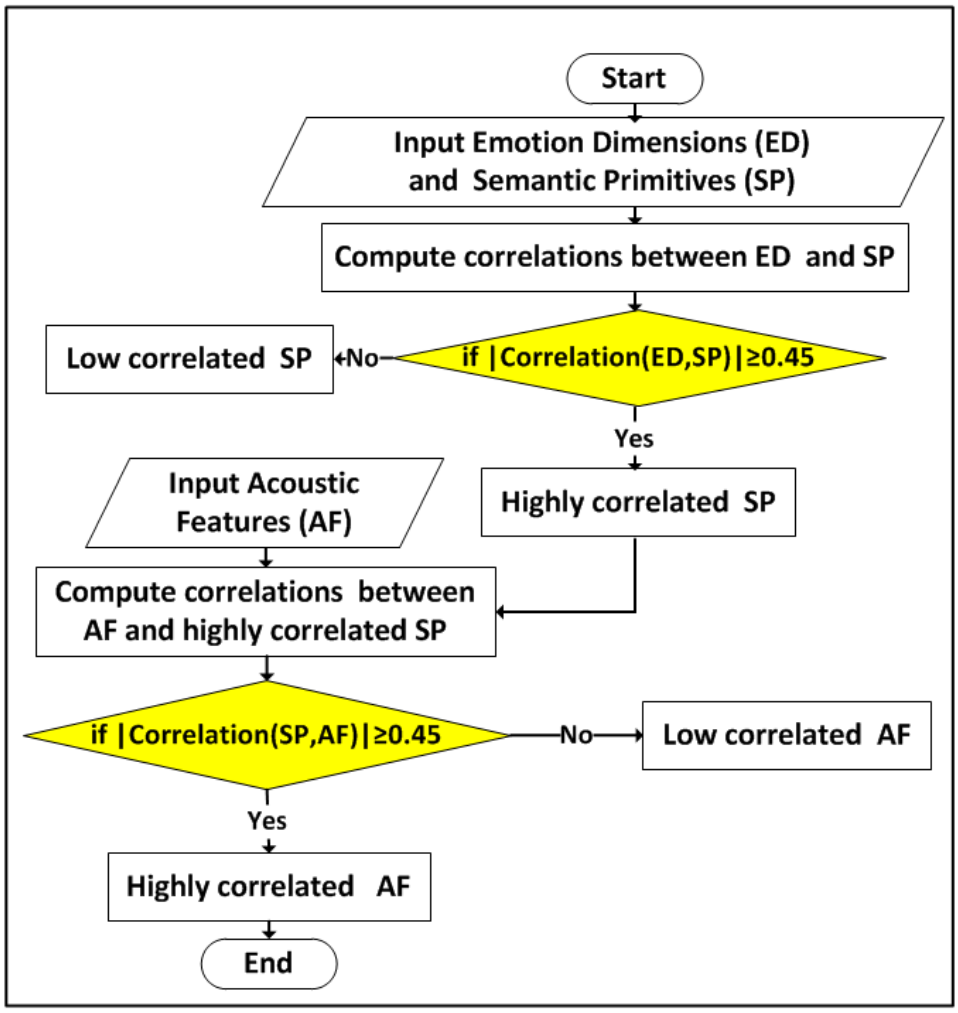


Figure 4.2: Process for acoustic feature selection.

4.3.2 Correlation between elements of the three-layer model

To select the most related acoustic features for each emotion dimensions, first the correlation between elements of the proposed model were calculated as preformed in the next two subsections.

4.3.2.1 The correlation between emotion dimensions and semantic primitives

First, the correlations between the elements of the top layer and the middle layer were calculated as follows: let $x^{(i)} = \{x_n^{(i)}\}(n = 1, 2, \ldots, N)$ be the sequence of the rated values of the i^{th} emotion dimension by the listening test, $i \in \{Valence, Activation, Dominance\}$. Moreover, let $s^{(j)} = \{s_n^{(j)}\}(n = 1, 2, \ldots, N)$ be the sequence of the rated values of the j^{th}

semantic primitive from another listening test, $j \in \{Bright, Dark, \ldots, Slow\}$. Where N is the number of utterances in used database ($N = 179$ for Japanese) and ($N = 200$ for German). Then the correlation coefficient $R_j^{(i)}$ between the semantic primitive $s^{(j)}$ and the emotion dimension $x^{(i)}$ can be determined by the following equation:

$$R_j^{(i)} = \frac{\sum_{n=1}^{N}(s_{j,n} - \overline{s_j})(x_n^{(i)} - \overline{x}^{(i)})}{\sqrt{\sum_{n=1}^{N}(s_{j,n} - \overline{s_j})^2}\sqrt{\sum_{n=1}^{N}(x_n^{(i)} - \overline{x}^{(i)})^2}} \tag{4.2}$$

where $\overline{s_j}$ and $\overline{x}^{(i)}$ are the arithmetic mean for the semantic primitive and emotion dimension, respectively.

Tables 4.3 and 4.4 show the correlation coefficients between all semantic primitives and all emotion dimensions for Japanese and German database, respectively. Where, the numbers in bold represent the higher correlations demonstrated by the absolute value of the correlation, which is $\geq$0.45. In addition, '#' in the last row and last column represents the number of higher correlations, for emotion dimensions and semantic primitives, respectively. The correlations between emotion dimension and semantic primitives revalues that there are many semantic primitives found to be highly correlated with valence dimension, six and seven semantic primitives in case of Japanese and German database, respectively. These semantic primitives can be used to to accurately estimate the valence dimension in the recognition process as will be described in Section 4.4. From Tables 4.3 and 4.4 we can easily note that the activation and dominance dimensions have similar correlations with all semantic primitives for both database, the reasons behind this is that, the acoustic characteristics of the dominance dimension may overlap with the activation, rendering this dimension redundant [76], and usually the third dimension dominance is added to emotional representation for distinguish between fear and anger as explained in chapter 2, however, the used emotion categories in both databases does not include the fear emotion category, therefore, the results of dominance and activation are very similar.

4.3.2.2 The correlation between semantic primitives and acoustic features

Second, the correlations coefficients between elements of the middle layer (semantic primitive), and the bottom layer (acoustic feature) are calculated as follows: Let $a_m = \{a_{m,n}\}(n = 1, 2, \ldots, N)$ be the sequence of values of the m^{th} acoustic feature, $m = 1, 2, \ldots, M$, where M be the number of extracted acoustic features in this study $M = 21$. Then the correlation coefficient $R_m^{(j)}$ between the acoustic feature a_m and the semantic primitive $s^{(j)}$ can be determined by the following equation:

$$R_m^{(j)} = \frac{\sum_{n=1}^{N}(a_{m,n} - \overline{a_m})(s_n^{(j)} - \overline{s}^{(j)})}{\sqrt{\sum_{n=1}^{N}(a_{m,n} - \overline{a_m})^2}\sqrt{\sum_{n=1}^{N}(s_n^{(j)} - \overline{s}^{(j)})^2}} \tag{4.3}$$

where $\overline{a_m}$, and $\overline{s}^{(j)}$ are the arithmetic mean for the acoustic feature and semantic primitive respectively.

Tables 4.5 and 4.6 lists the correlations coefficients between all semantic primitives and all acoustic features, for the Japanese and German database, respectively. The correlations analysis between semantic primitives and acoustic features show a stronger correlations, many acoustic features highly correlated with each semantic primitives. This strong correlations indicate that the prediction of semantic primitives from acoustic features will be more accurate and precise.

Table 4.3: Japanese Database: The correlation coefficients between the semantic primitives and the emotion dimensions.

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	Valence	**0.76**	**-0.59**	**0.43**	**-0.45**	-0.15	-0.04	0.01	-0.11	0.01	0.05	**-0.72**	**0.96**	-0.17	-0.11	-0.23	0.04	-0.11	6
2	Activation	**0.69**	**-0.86**	**0.79**	**-0.79**	**0.91**	**-0.95**	**-0.94**	**0.90**	**0.84**	**-0.73**	**-0.64**	0.35	**0.89**	**-0.93**	**0.89**	**0.76**	**-0.73**	16
3	Dominance	**0.54**	**-0.76**	**0.64**	**-0.65**	**0.96**	**-0.98**	**-0.91**	**0.88**	**0.78**	**-0.67**	**-0.48**	0.21	**0.88**	**-0.89**	**0.92**	**0.75**	**-0.71**	16
	#	3	3	2	2	2	2	2	2	2	2	3	1	2	2	2	2	2	36

Table 4.4: German Database: The correlation coefficients between the semantic primitives and the emotion dimensions.

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	Valence	**0.85**	**-0.67**	**0.65**	**-0.60**	0.08	-0.37	-0.20	0.13	0.26	-0.24	**-0.87**	**0.78**	0.12	-0.36	0.26	0.41	**-0.48**	7
2	Activation	**0.74**	**-0.92**	**0.88**	**-0.93**	**0.92**	**-0.97**	**-0.91**	**0.89**	**0.92**	**-0.85**	**-0.58**	**0.66**	**0.92**	**-0.98**	**0.95**	**0.78**	**-0.81**	17
3	Dominance	**0.64**	**-0.87**	**0.80**	**-0.88**	**0.95**	**-0.97**	**-0.89**	**0.89**	**0.89**	**-0.81**	**-0.47**	**0.60**	**0.93**	**-0.97**	**0.95**	**0.78**	**-0.80**	17
	#	3	3	3	3	2	2	2	2	2	2	3	3	2	2	2	2	3	41

Table 4.5: Japanese Database: The correlation coefficients between the acoustic features and semantic primitives.

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	**MH_A**	**-0.67**	**0.79**	**-0.73**	**0.72**	**-0.77**	**0.81**	**0.82**	**-0.77**	**-0.71**	**0.62**	**0.61**	-0.35	**-0.79**	**0.84**	**-0.74**	**-0.52**	**0.48**	**16**
2	**MH_E**	-0.38	**0.50**	-0.41	0.41	**-0.53**	**0.57**	**0.52**	**-0.47**	-0.41	0.34	0.35	-0.19	**-0.51**	**0.55**	**-0.50**	-0.34	0.31	**8**
3	MH_I	0.34	-0.23	0.40	-0.40	-0.20	0.11	0.00	-0.03	0.03	-0.09	**-0.45**	0.36	-0.02	-0.10	-0.15	-0.10	0.06	1
4	**MH_O**	**0.57**	**-0.71**	**0.65**	**-0.65**	**0.70**	**-0.76**	**-0.75**	**0.69**	**0.62**	**-0.53**	**-0.53**	0.27	**0.74**	**-0.79**	**0.68**	0.43	-0.40	**14**
5	MH_U	0.02	0.08	0.04	-0.02	-0.23	0.23	0.13	-0.11	-0.06	0.01	-0.04	0.05	-0.15	0.14	-0.18	-0.09	0.08	0
6	**F0_RS**	**0.83**	**-0.88**	**0.98**	**-0.98**	**0.60**	**-0.69**	**-0.80**	**0.74**	**0.74**	**-0.69**	**-0.87**	**0.53**	**0.76**	**-0.87**	**0.63**	**0.51**	**-0.51**	**17**
7	**F0_HP**	**0.79**	**-0.84**	**0.94**	**-0.94**	**0.60**	**-0.67**	**-0.80**	**0.75**	**0.76**	**-0.72**	**-0.82**	**0.47**	**0.76**	**-0.85**	**0.64**	**0.53**	**-0.53**	**17**
8	F0_AP	-0.12	0.13	-0.11	0.12	-0.08	0.11	0.09	-0.08	-0.07	0.03	0.12	-0.11	-0.10	0.13	-0.08	-0.03	0.03	0
9	F0_RS1	-0.13	0.17	-0.12	0.12	-0.14	0.20	0.12	-0.07	-0.09	0.06	0.13	-0.11	-0.12	0.15	-0.11	0.02	-0.01	0
10	**PW_R**	**0.50**	**-0.57**	**0.50**	**-0.53**	0.45	**-0.53**	**-0.47**	0.43	0.45	-0.37	**-0.48**	0.34	**0.47**	**-0.55**	0.42	0.26	-0.26	**9**
11	**PW_RHT**	-0.09	-0.04	0.15	-0.12	**0.50**	-0.37	**-0.48**	**0.51**	0.43	**-0.46**	0.11	-0.41	**0.52**	-0.36	**0.55**	0.35	-0.29	**6**
12	PW_RS1	-0.11	0.20	-0.16	0.17	-0.22	0.25	0.18	-0.16	-0.11	0.09	0.16	-0.07	-0.18	0.21	-0.20	-0.18	0.18	0
13	PW_RAP	0.39	-0.43	0.40	-0.41	0.32	-0.38	-0.36	0.32	0.36	-0.31	-0.37	0.25	0.36	-0.42	0.31	0.16	-0.13	0
14	SP_F1	0.24	-0.28	0.33	-0.31	0.30	-0.29	-0.38	0.37	0.34	-0.30	-0.21	-0.02	0.43	-0.42	0.31	0.11	-0.06	0
15	SP_F2	0.00	-0.02	0.05	-0.05	0.15	-0.10	-0.11	0.13	0.18	-0.18	0.03	-0.07	0.12	-0.07	0.15	0.06	-0.02	0
16	SP_F3	0.14	-0.26	0.21	-0.23	0.36	-0.36	-0.33	0.30	0.23	-0.17	-0.17	0.07	0.31	-0.31	0.35	0.30	-0.32	0
17	SP_TL	0.35	-0.39	0.21	-0.25	0.16	-0.28	-0.16	0.09	0.05	0.02	-0.35	0.40	0.10	-0.23	0.09	0.15	-0.19	0
18	SP_SB	0.32	-0.37	0.40	-0.39	0.36	-0.39	-0.42	0.41	0.39	-0.36	-0.32	0.14	0.43	**-0.45**	0.36	0.24	-0.23	1
19	**DU_TL**	-0.13	0.21	-0.13	0.15	-0.27	0.27	0.24	-0.22	-0.15	0.10	0.13	-0.10	-0.17	0.18	-0.24	**-0.45**	**0.47**	**2**
20	**DU_CL**	**-0.46**	**0.58**	**-0.47**	**0.50**	**-0.51**	**0.57**	**0.53**	**-0.47**	-0.38	0.28	**0.47**	-0.33	**-0.47**	**0.54**	**-0.47**	**-0.55**	**0.57**	**14**
21	**DU_RCV**	**-0.60**	**0.68**	**-0.61**	**0.63**	**-0.50**	**0.58**	**0.57**	**-0.50**	-0.44	0.33	**0.60**	-0.41	**-0.56**	**0.67**	**-0.46**	-0.33	0.33	**12**
	#	7	8	7	7	8	8	9	8	4	5	8	2	9	9	8	5	5	117

Table 4.6: German Database: The correlation coefficients between the acoustic features and semantic primitives.

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	**MH_A**	**-0.61**	**0.78**	**-0.74**	**0.80**	**-0.78**	**0.84**	**0.74**	**-0.72**	**-0.73**	**0.68**	**0.50**	**-0.62**	**-0.76**	**0.84**	**-0.82**	**-0.71**	**0.74**	**17**
2	**MH_E**	**-0.46**	**0.62**	**-0.60**	**0.65**	**-0.69**	**0.69**	**0.68**	**-0.68**	**-0.65**	**0.62**	0.38	-0.40	**-0.67**	**0.69**	**-0.67**	**-0.56**	**0.58**	**15**
3	MH_I	-0.06	0.17	-0.01	0.12	-0.15	0.23	0.07	-0.06	-0.05	-0.02	0.03	-0.11	-0.12	0.21	-0.13	-0.26	0.31	0
4	**MH_O**	**-0.51**	**0.64**	**-0.60**	**0.64**	**-0.64**	**0.68**	**0.62**	**-0.60**	**-0.60**	**0.58**	**0.43**	**-0.50**	**-0.63**	**0.68**	**-0.64**	**-0.53**	**0.56**	**16**
5	**MH_U**	-0.38	**0.49**	-0.42	**0.46**	-0.41	**0.51**	0.38	-0.38	-0.38	0.33	0.31	-0.43	-0.40	**0.50**	**-0.46**	**-0.46**	**0.49**	**7**
6	**F0_RS**	**0.52**	**-0.64**	**0.73**	**-0.73**	**0.70**	**-0.66**	**-0.75**	**0.74**	**0.75**	**-0.77**	**-0.51**	0.44	**0.71**	**-0.70**	**0.73**	0.44	-0.43	**14**
7	**F0_HP**	**0.49**	**-0.55**	**0.66**	**-0.63**	**0.60**	**-0.55**	**-0.70**	**0.68**	**0.70**	**-0.71**	-0.44	0.35	**0.63**	**-0.59**	**0.62**	0.35	-0.34	**13**
8	F0_AP	-0.11	0.14	-0.14	0.15	-0.12	0.15	0.13	-0.11	-0.13	0.12	0.14	-0.10	-0.14	0.14	-0.13	-0.12	0.15	0
9	F0_RS1	-0.09	0.11	-0.10	0.11	-0.09	0.12	0.08	-0.07	-0.08	0.07	0.13	-0.10	-0.09	0.11	-0.10	-0.10	0.14	0
10	**PW_R**	**0.52**	**-0.69**	**0.69**	**-0.72**	**0.75**	**-0.75**	**-0.78**	**0.77**	**0.78**	**-0.77**	-0.40	0.42	**0.75**	**-0.76**	**0.74**	**0.49**	**-0.51**	**15**
11	**PW_RHT**	0.07	-0.27	0.31	-0.34	**0.56**	-0.43	**-0.54**	**0.57**	**0.53**	**-0.51**	-0.02	0.01	**0.56**	**-0.46**	**0.52**	0.23	-0.19	**8**
12	PW_RS1	0.15	-0.17	0.14	-0.16	0.14	-0.17	-0.15	0.14	0.15	-0.14	-0.13	0.17	0.16	-0.15	0.16	0.12	-0.11	0
13	**PW_RAP**	0.25	-0.31	0.36	-0.36	0.38	-0.35	-0.43	0.43	**0.46**	**-0.48**	-0.17	0.18	0.41	-0.38	0.37	0.03	-0.07	**2**
14	**SP_F1**	**-0.59**	**0.58**	**-0.53**	**0.55**	-0.32	**0.48**	0.34	-0.31	-0.37	0.33	**0.55**	**-0.60**	-0.33	**0.47**	-0.40	-0.44	**0.50**	**9**
15	SP_F2	-0.18	0.23	-0.27	0.29	-0.32	0.28	0.36	-0.36	-0.32	0.35	0.12	-0.16	-0.34	0.31	-0.31	-0.21	0.19	0
16	SP_F3	-0.09	0.04	-0.09	0.05	-0.04	0.01	0.15	-0.13	-0.12	0.21	0.04	0.06	-0.07	0.04	0.02	0.15	-0.16	0
17	SP_TL	0.26	-0.31	0.20	-0.25	0.17	-0.27	-0.09	0.08	0.12	-0.05	-0.26	0.34	0.15	-0.27	0.22	0.31	-0.35	0
18	SP_SB	-0.08	0.04	-0.11	0.07	-0.05	0.02	0.10	-0.12	-0.15	0.16	0.05	-0.03	-0.09	0.04	-0.03	0.17	-0.15	0
19	**DU_TL**	-0.34	0.40	-0.29	0.37	-0.30	0.41	0.20	-0.19	-0.17	0.12	0.28	**-0.48**	-0.23	0.38	-0.33	-0.42	**0.47**	**2**
20	DU_CL	-0.30	0.37	-0.28	0.33	-0.29	0.37	0.22	-0.21	-0.20	0.14	0.24	-0.38	-0.25	0.36	-0.33	-0.40	0.45	0
21	DU_RCV	-0.30	0.36	-0.37	0.37	-0.37	0.36	0.37	-0.36	-0.37	0.34	0.22	-0.23	-0.38	0.37	-0.37	-0.29	0.30	0
	#	7	8	7	8	7	8	7	7	8	8	3	4	7	9	8	5	7	118

4.3.3 Selection Results

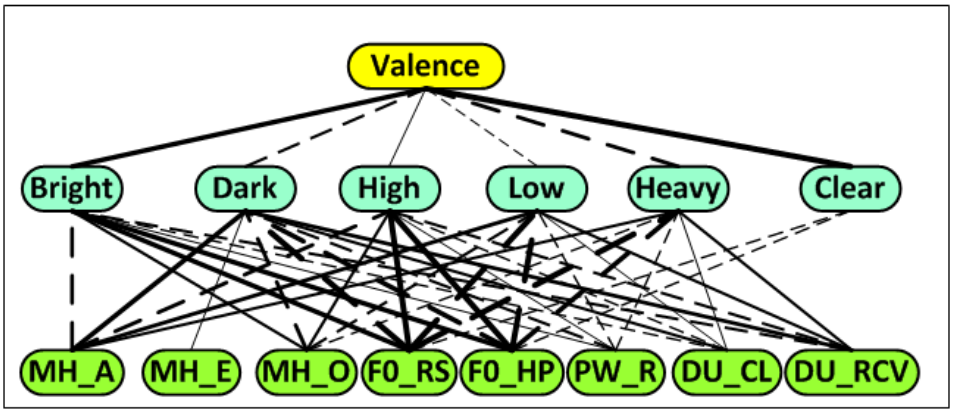
For each emotion dimension, a perceptual three-layer model was constructed using the top-down procedure described in the proposed feature selection method, for example, to construct the perceptual three-layer model for valence dimension for the Japanese, the following steps is preformed:

- valence dimension will be in the top layer then;
- using step 2 in the proposed method, the highly correlated semantic primitives for valence dimension are selected, from Table 4.3 it was found 6 semantic primitive highly correlated with valence dimension the are (Bright, Dark, High, Low, Heavy, Clear). These semantic primitives composes the middle layer;
- using step 4 in the proposed method, for each selected semantic, we select the highly correlated acoustic features and the combination of all selected acoustic features will be the most related acoustic feature to the valence dimension and these acoustic features composes the the bottom layer.

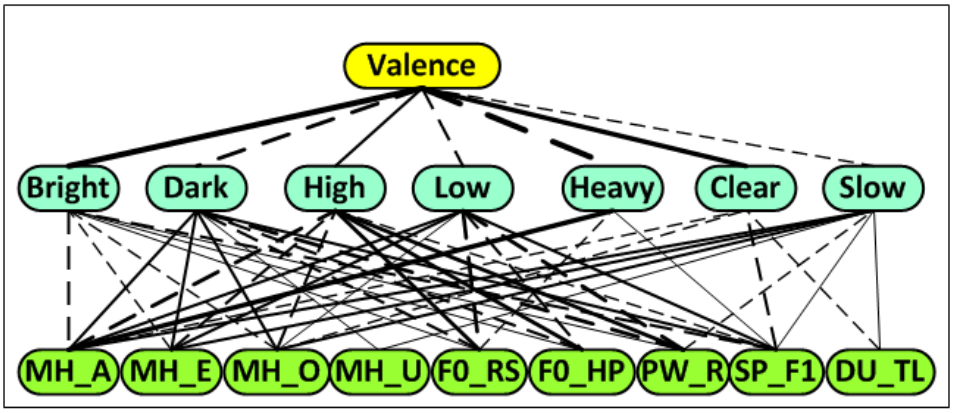
For example, Figs. 4.3(a) and 4.3(b) illustrate the valence perceptual model for Japanese and German database, respectively. Where the solid and dashed lines in these figures represent positive and negative correlations, respectively. Also, the thickness of each line indicates the strength of the correlation: the thicker the line, the higher the correlation.

In case of valence dimension for the Japanese database as shown in Fig. 4.3(a), it is evident that six semantic primitives were found that highly correlated with valence as shown in the middle layer. These six semantic primitives are highly correlated with eight acoustic features as shown in the bottom layer of this figure.

The valence perceptual model for Japanese and German language are compared as follows: For both languages, the valence dimension is found to be positively correlated with Bright, High and Clear semantic primitives, while it is negatively correlated with Dark, Low, and Heavy semantic primitives. Therefore, the two languages not only share



(a) Japanese database.

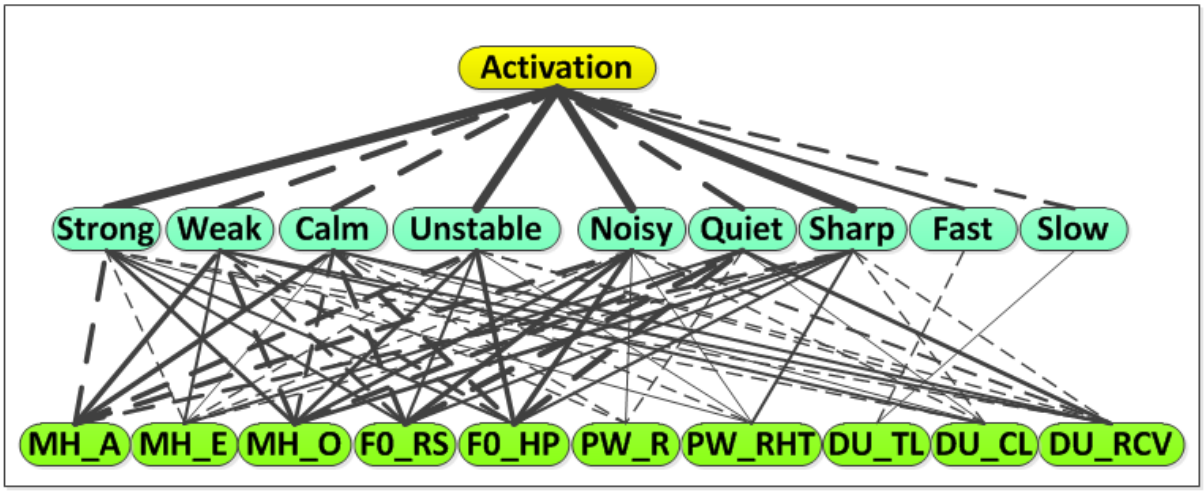


(b) German database.

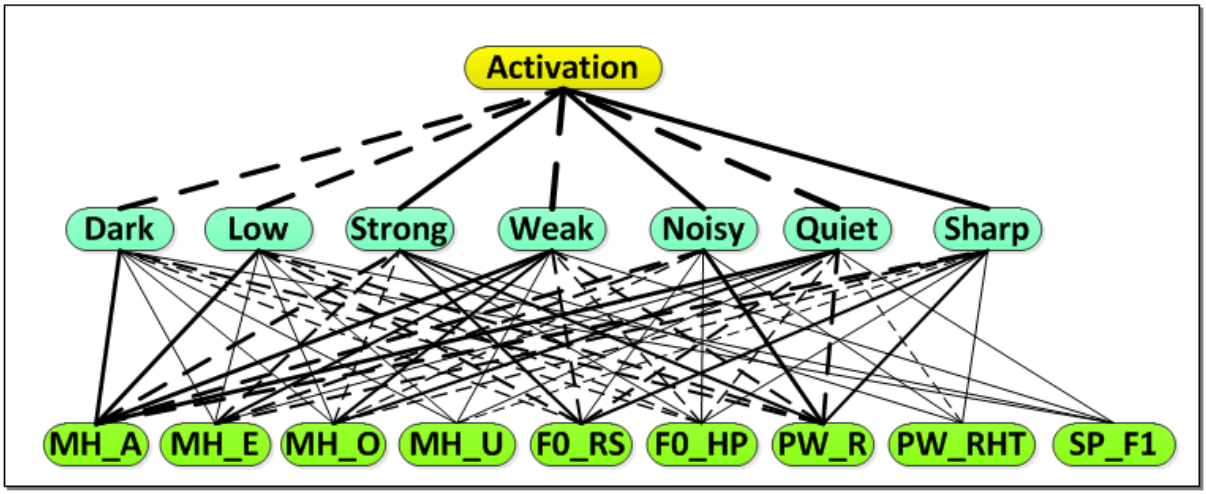
Figure 4.3: Valence perceptual model.

six semantic primitives but also similar correlations between the emotion dimensions and the corresponding semantic primitives.

In addition, comparing the relationship between semantic primitives and acoustic features, it is found that the six semantic primitives that were shared by both German and Japanese have a similar correlations with six common acoustic features (MH_A, MH_E, MH_O, F0_RS, F0_HP, and PW_R). This finding suggests the possibility of some type of universality of acoustic cues associated with semantic primitives.

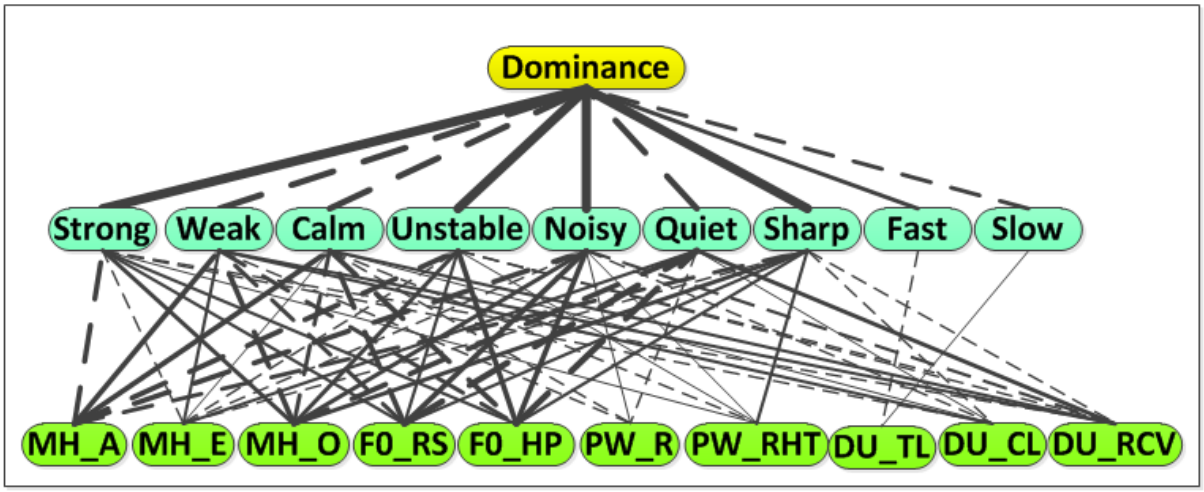


(a) Japanese database.

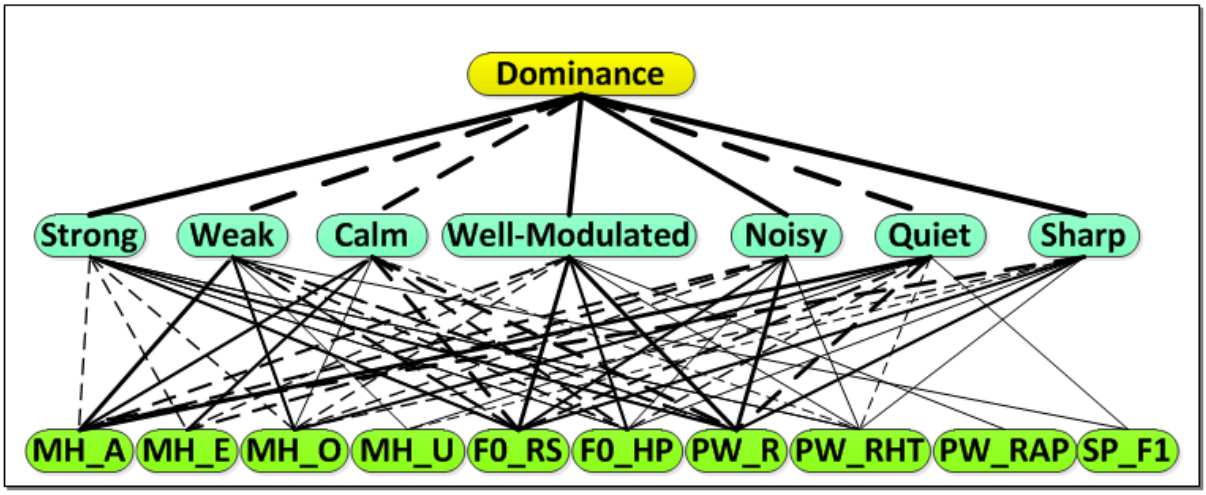


(b) German database.

Figure 4.4: Activation perceptual model.



(a) Japanese database.



(b) German database.

Figure 4.5: Dominance perceptual model.

Table 4.7: Selected acoustic features for each emotion dimension for Japanese database.

	Valence	Activation	Dominance
Voice quality	MH_A MH_E MH_O	MH_A MH_E MH_O	MH_A MH_E MH_O
Pitch	RS HP	RS HP	RS HP
Power envelope	PW_R -	PW_R PW_RHT	PW_R PW_RHT
Duration	- CL RCV	DU_TL CL RCV	DU_TL CL RCV
Number	8	10	10

Table 4.8: Selected acoustic features for each emotion dimension for German database.

	Valence	Activation	Dominance
Voice quality	MH_A MH_E MH_O MH_U	MH_A MH_E MH_O MH_U	MH_A MH_E MH_O MH_U
Pitch	RS HP	RS HP	RS HP
Power envelope	PW_R -	PW_R PW_RHT -	PW_R PW_RHT PW_RAP
Power spectrum	F1	F1	F1
Duration	TL	-	-
Number	9	9	10

Therefore, the proposed method can be used effectively to select the most relevant acoustic features for each emotion dimension regardless the used language. In a similar way the perceptual three-layer model were constructed for activation and dominance dimensions for both Japanese and German as shown in Figures 4.4 and 4.5.

4.3.4 The selected acoustic features

Finlay after applying these steps for Japanese and German database, the results are summarized in Table 4.7 and Table 4.8, respectively.

From the above Tables, it is clearly that, there are eight and nine acoustic features related to valence dimension in case of Japanese and German language, respectively.

Therefor, the new feature selection method outperform the traditional feature selection method which was not consider the human perception into account.

4.3.5 Discussion

Our model mimics the human perception process for understanding emotions on the basis of Brunswick's lens model [10], where the speaker expresses his/her emotional state through some acoustic features. These acoustic features are interpreted by the listener into some adjectives describing the speech signal, and from these adjectives, the listener can judge the emotional state. For example, if the adjectives describing the voice are Dark, Slow, Low, and Heavy, these make the human listener feel that the emotional state is negative valence and very weak activation, resulting in it being detected as a very Sad emotional state in the categorical approach.

On the other hand, the conventional acoustic features selection method was based on the correlations between acoustic features and emotion dimension as a two-layer model. To investigate the effectiveness of the proposed feature selection method, the results were compared with the conventional method. For example, Table 4.2, shows the correlations coefficients between acoustic features and emotion dimensions directly in case of German language. From this table, evidently only one acoustic feature highly correlated with the valence dimension ($|correlation(SP\_F1, Valence)| = 0.55 > 0.45$), while eight acoustic features highly correlated with the activation and dominance dimensions. Therefore, valence shows a smaller number of highly correlated acoustic features than the activation and dominance. These results are similar to those of many previous studies [29, 76]. Due to this drawback, most previous studies achieved a very low performance for valence estimation using the conventional approach [31, 95].

The most important result is that, using the proposed three-layer model for feature selection, the number of relevant acoustic features to emotion dimensions increases. For example, the number of relevant features for the most difficult dimension valence increases from one to nine using the proposed method. Moreover, the number of features increased

from eight to nine for activation and from eight to ten for dominance. The selected acoustic features can be used to improve emotion dimensions estimation as described in detail in the next section.

4.4 The proposed speech emotion Recognition System

This section introduces the implementation of the proposed model into a speech emotion recognition system. The task of speech emotion recognition system based on the dimensional approach can be viewed as using an estimator to map the acoustic features to real-valued emotion dimensions (valence, activation, and dominance). The perceptual three-layer models were built for all emotion dimensions as described in Section 4.3.3. These models are used to construct our proposed automatic speech emotion recognition system in order to estimate emotion dimensions from acoustic features.

The bottom-up method was used to construct our system, the selected acoustic features (bottom layer) form the previous section are used as an input for the proposed system to predict emotion dimensions (top layer).

4.4.1 System Implementation

Emotion dimension values can be estimated using many estimator such as K-nearest neighborhood (KNN), Support Vector Regression (SVR), or Fuzzy Inference System FIS. In this study, for selecting the best estimator among KNN, SVR, and FIS, pre-experiments not included here indicated that our best results were achieved using an FIS estimator. Therefore, FIS was used to connect the elements of the three-layer model. Most statistical methodology is mainly based on a linear and precise relationship between the input and the output, while the relationships among acoustic features, semantic primitives, and emotion dimensions are non-linear. Therefore, fuzzy logic is a more appropriate mathematical tool for describing this non-linear relationship [31, 35, 38]. In Chapter 2 the FIS estimators

are described in more details.

A FIS implements a nonlinear mapping from an input space to an output space by a number of fuzzy if-then rules constructed from human knowledge [38]. Using artificial neural networks to identify fuzzy rules and tune the parameters of membership functions in FIS automatically is called Adaptive-Network-based Fuzzy Inference System (ANFIS). Designing a standard FIS need the expert knowledge, However using ANFIS this need is eliminated. Therefore, in this study, ANFIS was used to construct FISs models that connect the elements of our recognition system.

Figure 4.6 shows a block diagram of the proposed automatic emotion recognition system based on the three-layered model, this system consists of two main stages: The first stage is model creation which is employed for training the model, and the second stage is applying emotion recognition to test the model. Our system was constructed by using FIS to build the mathematical relationship between the elements of the three-layer model as follows: (1) FIS1 is used to map the acoustic features onto semantic primitives, (2) FIS2 is used to map the semantic primitives onto emotion dimensions. The desired output is an estimation of a real-valued for emotion dimensions: valence, activation, and dominance.

The conventional emotion dimensions estimators were usually used to map the acoustic features to a real-valued emotion dimensions as a two-layer only. However, in this study, these estimators were used to estimate emotion dimensions in the bases of three-layer model; simply by mapping the acoustic features into a real-valued semantic primitives (adjectives) followed by mapping the semantic primitives to a real-valued emotion dimensions.

4.4.2 Emotion Dimensions Estimation using the three-layer model

In Section 4.3.3 a perceptual three-layer model were constructed for each emotion dimension using the top-down method and the feature selection algorithm, for Japanese

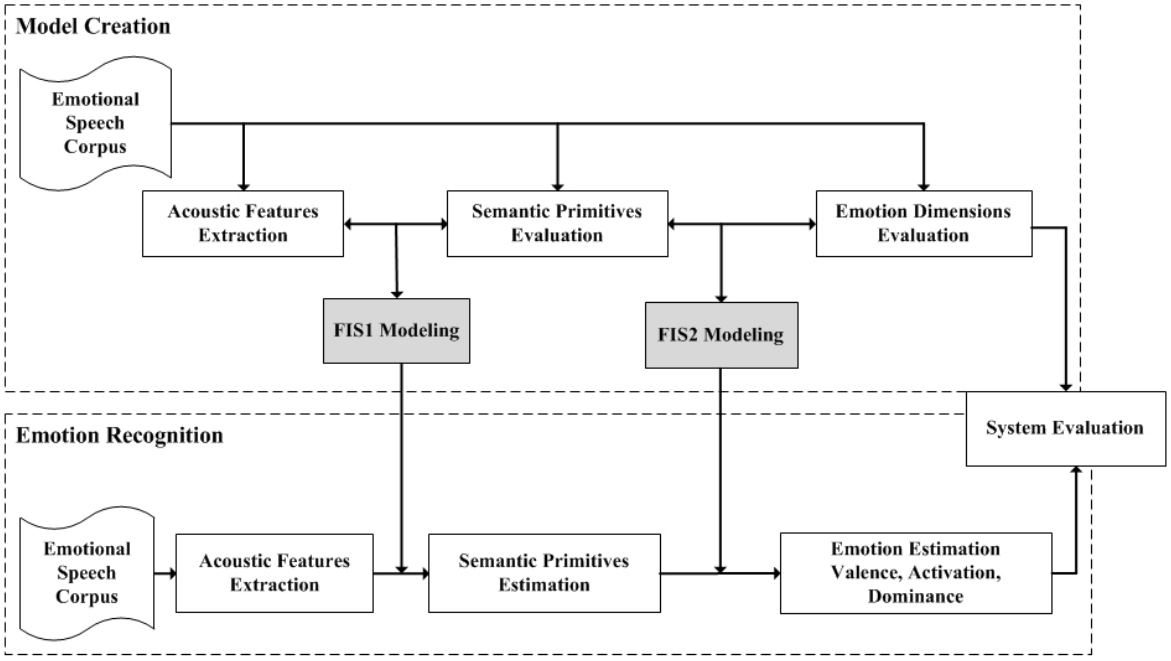


Figure 4.6: Block diagram of the proposed emotion recognition system based on the three-layered model.

and German database. The perceptual three-layer model for valence, and activation and dominance for both Japanese database, are shown in Figures 4.3, 4.4, and 4.5, respectively. In the recognition stage, these model were used by a bottom-up method to estimate emotion dimension.

The bottom-up method was used to estimate emotion dimensions from the acoustic features in two main steps:

- **Step 1:** Semantic primitive estimation: in this step each semantic primitive is estimated from acoustic features. For each semantic primitives one FIS is needed to estimate this semantic primitives from all acoustic features; the input here is the acoustic features and the output are the estimated semantic primitives, as described in Section 4.5.

- **Step 2:** Emotion dimension estimation: the estimated semantic primitives is used to estimate emotion dimension. One FIS system is required for each emotion dimension. The input for each FIS are the estimated semantic primitives and the output

Table 4.9: The elements in the perceptual model for Japanese-valence

Layer	Elements	Number
Top layer	Valence	1
Middle Layer	Bright, Dark, High, Low, Heavy, Clear	6
Bottom Layer	MH_A, MH_E, MH_O, F0_RS, F0_HP, PW_R, DU_CL, DU_RCV	8

is the estimated emotion dimension, as described in Section 4.5.1.

In the rest of this chapter we will explain these steps for valence dimension in Japanese database, Similarly, the activation and dominance can be estimated using these steps. In order to estimate valence dimension for Japanese language, the perceptual model for valence dimension was used from bottom to the top as shown in Figure 4.7 which describe the perceptual model for Japanese-valence.

The elements in each layer for this model are listed in Table 4.9 as follow: valence dimension in the top layer, six semantic primitives in the middle layer, and eight acoustic features in the bottom layer.

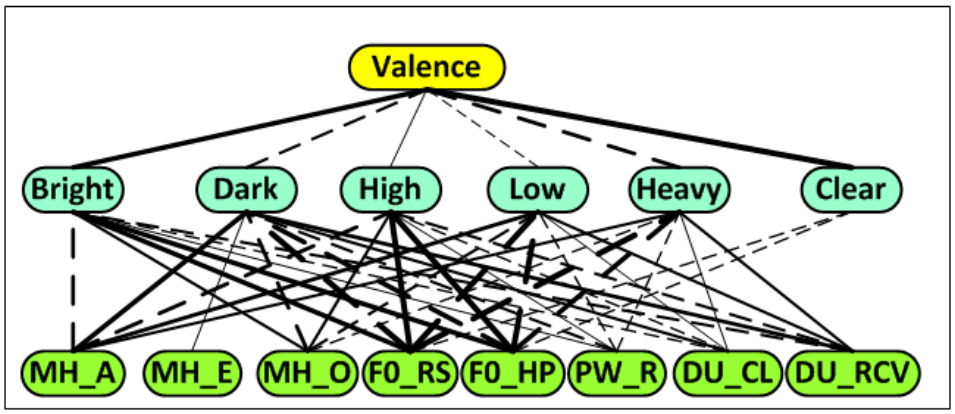


Figure 4.7: The perceptual model for valence dimension from Japanese database.

In order, to estimate the valence dimension using the perceptual model in Fig. 4.7, a bottom-up method was used to estimate the values of the sex semantic primitives in the middle layer from the eight acoustic features in the bottom layer. As mentioned in the previous section that FIS is multi-input one output, therefore seven FISs systems are required to estimate valence emotion dimension as follow: sex FISs were needed: one to

estimate each semantic primitive from acoustic features. In addition, one FIS is needed to estimate the value of valence dimension from the sex estimated semantic primitives. Figure 4.8 illustrates the seven FISs used to estimate valence for Japanese language.

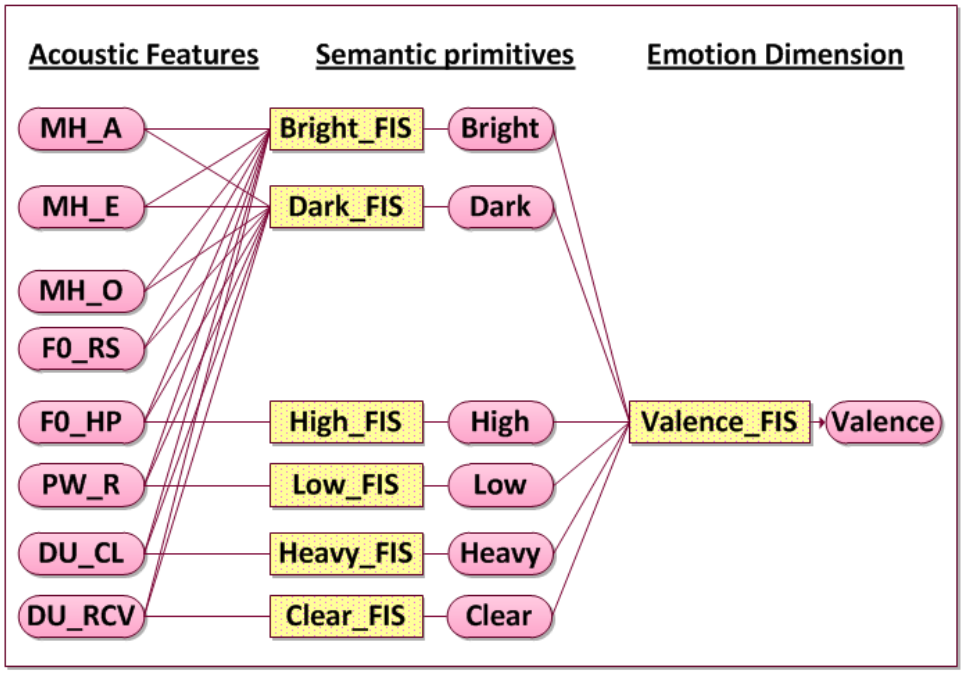


Figure 4.8: Valence dimension estimation using a three layer model.

4.5 Semantic primitives estimations using ANFIS

In this study, ANFIS was used to construct FISs models that connect the elements of our recognition system. The ANFIS model under consideration is a multi-input single-output system with eight inputs and one output. Six ANFIS were used to build six FISs for estimating each semantic primitives in the middle layer from the eight acoustic features in the bottom layer. Figure 4.9 shows the input and the output for estimating Bright semantic primitive, eight acoustic features (MH_A, MH_E, MH_O, F0_RS, F0_HP, PW_R, DU_CL, DU_RCV) and one output Bright semantic primitive. Similarly other semantic primitives can be estimated using these acoustic features.

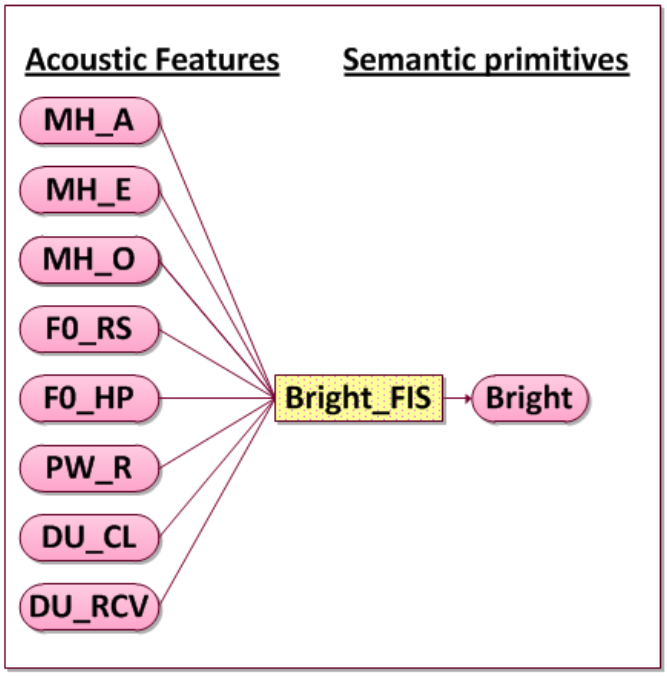


Figure 4.9: Bright semantic primitive estimation form acoustic features using FIS.

The data were obtained from the Japanese database the data set consists of 179 utterances as introduced in chapter 3. The ANFIS structure generated in this study utilizes fuzzy clustering of the input and output data sets as well as the Gaussian Bell shape membership function. Thus the number of rules is equal to the number of output clusters. In order to minimize the over fitting of the model developed, the complete data set was split into a training (80%) and testing data set (20%). The ANFIS model was

first trained using the training data set followed by validation process using the remaining data. The errors associated with the training and checking processes are recorded. ANFIS training was found to converge after training with 150 epochs for High, Clear, and Heavy, moreover, it converge after 250 epochs for Low and Dark, while it converges after 300 epochs for Bright as shown in Figure 4.10.

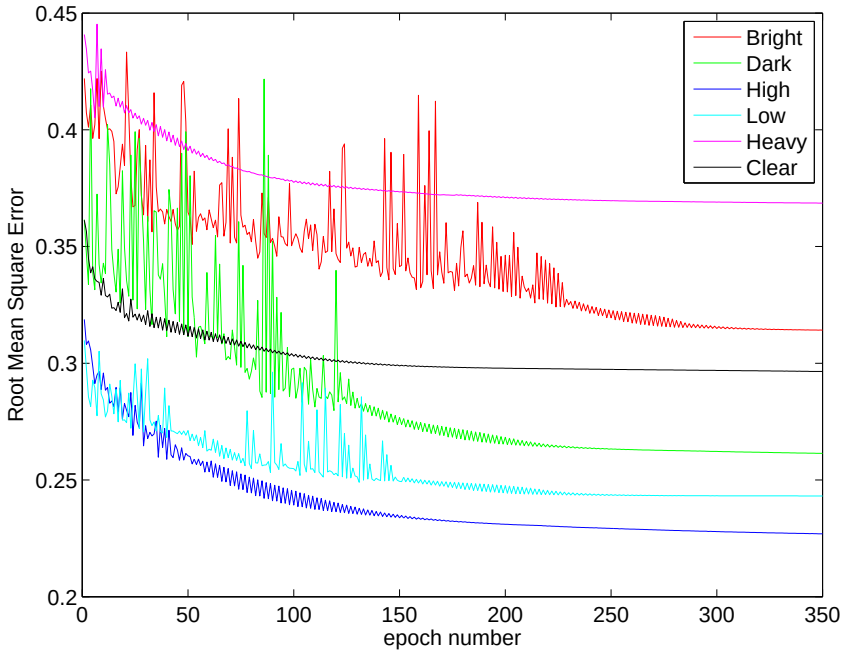


Figure 4.10: ANFIS training RMSE for (Bright, Dark, High, Low, Heavy, Clear).

Root mean square error (RMSE) for both the training and testing of ANFIS are very small which reflects the ability of ANFIS to capture the essential components of underlying dynamics governing the relationships between the input and the output variables. The computation of membership functions (MFs) parameters is facilitated by a gradient descent vector.

Fuzzy reasoning which is made up of fuzzy if-then rules together with fuzzy membership functions is the main feature of fuzzy inference systems. Fuzzy reasoning derives conclusions from the set of rules which are either data driven or provided by experts. Figure 4.11 shows the reasoning procedure for a first order Sugeno fuzzy model. Each rule has a crisp output and the overall output is a weighted average. This figure shows the IF-then rules derived by ANFIS for estimating Bright semantic primitives from eight

acoustic features in case of Japanese database. There are three rules which defining the relations between the input and output as a linguistic variables.

Figure 4.12 shows the a sample of the rules using crisp values for example when MA_A is very low and MH_E is very low and MH_O is very high, and F0_RS is very high, and F0_HP is very high, and PW_R is low, and DU_CL is very high, and DU_RCV is very low Then Bright is very high (Bright range from 1 to 5) and this value is 4.84 is very high.

Using the same procedure the other five semantic primitive were estimated from the eight acoustic features. This prediction for semantic primitives will be used to estimate the valence dimension as described in the next section.

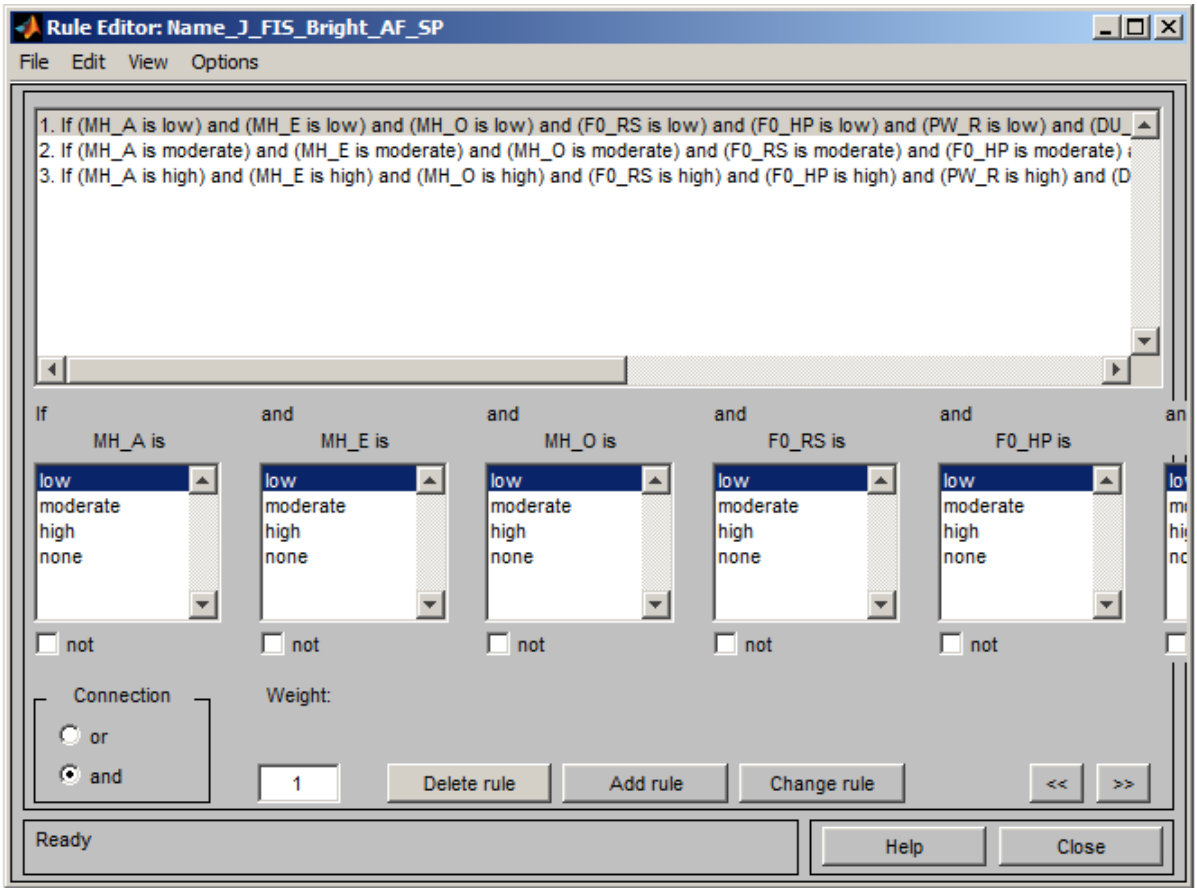


Figure 4.11: If-Then rules derived by ANFIS used for estimating Bright.

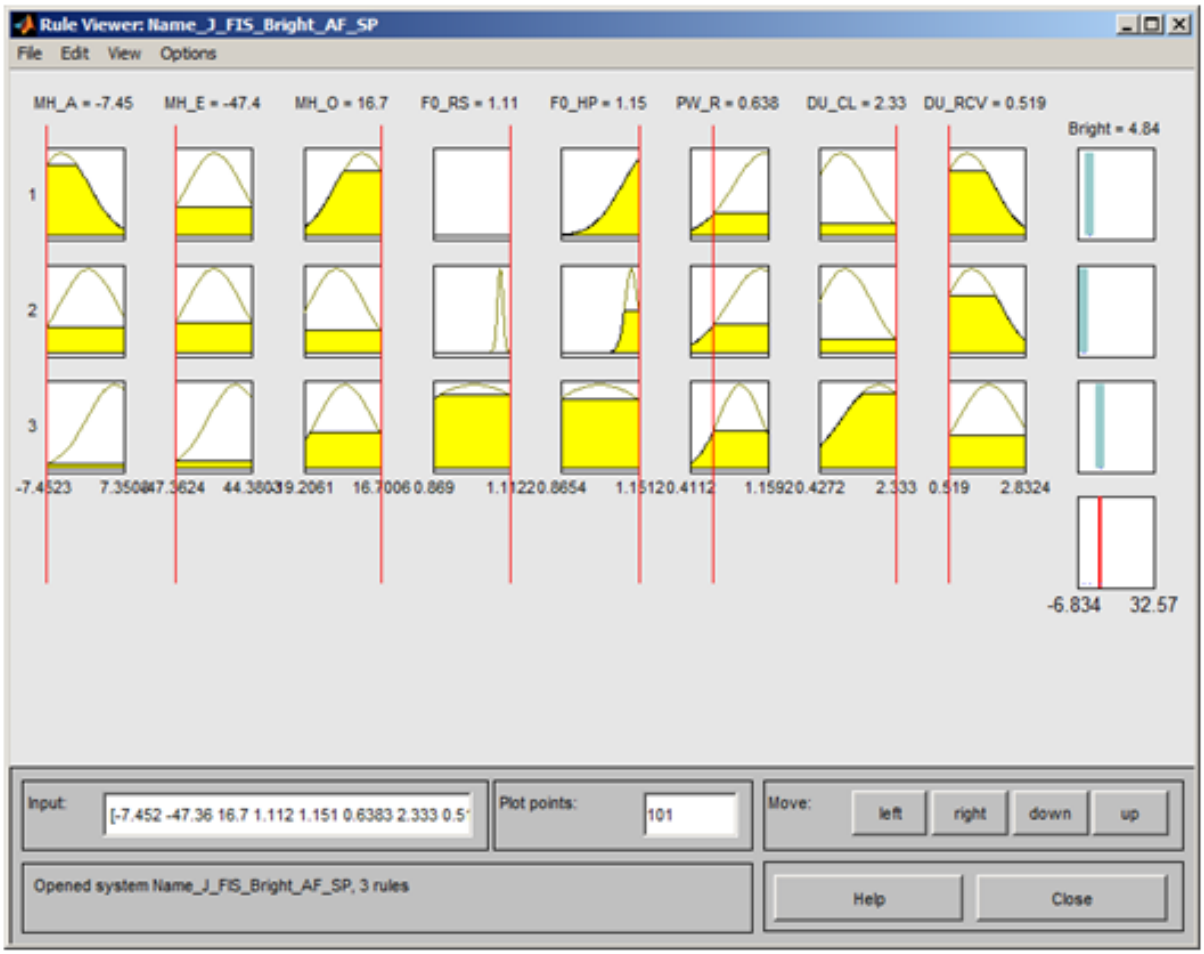


Figure 4.12: Sample of rule set of an ANFIS model Bright=+4.84, very large

4.5.1 Dimension estimations using ANFIS

The second step is to estimate emotion dimension from the estimated semantic primitives. One ANFIS was built for estimating valence dimension from the six estimated semantic primitives the input and the output for this system are shown in Figure 4.13.

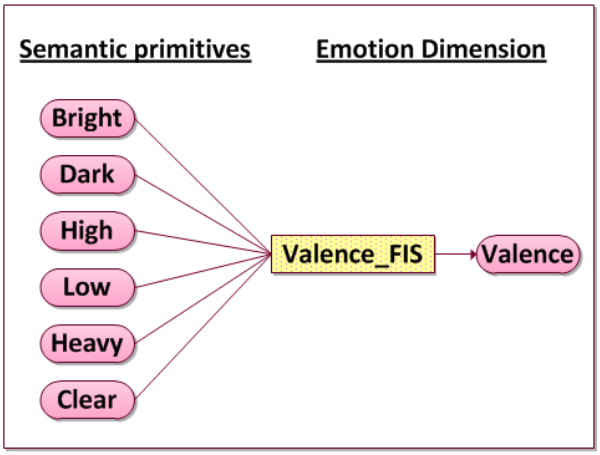


Figure 4.13: Valence dimension estimation from semantic primitives.

The ANFIS model was first trained using the training data set followed by validation process using the remaining data. The errors associated with the training and checking processes are recorded. ANFIS training was found to converge after training with 150 epochs for valence, while it converges after 200 epochs for activation and dominance as shown in Figure 4.14.

For training the system the input and the output were the human evaluation. The same 80% which was used for training the ANFIS for estimating semantic primitives in the previous section were used as a training set, while the testing set is the set of estimated semantic primitives from the first step of recognition. In the previous studies usually emotion dimension estimated from acoustic features directly, while in this study semantic primitives are used as a bridge between the acoustic features and emotion dimensions.

Figure 4.15 shows the IF-then rules derived by ANFIS for estimating valence dimension from six semantic primitives for Japanese database. There are four rules which defining the relations between the input and output as a linguistic variables.

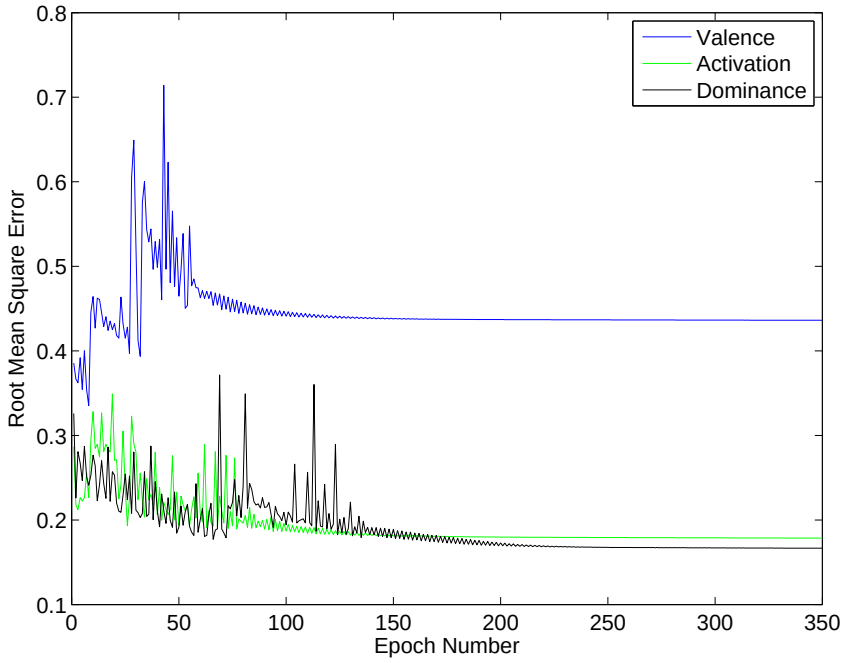


Figure 4.14: ANFIS training RMSE for (Valence, Activation, Dominance).

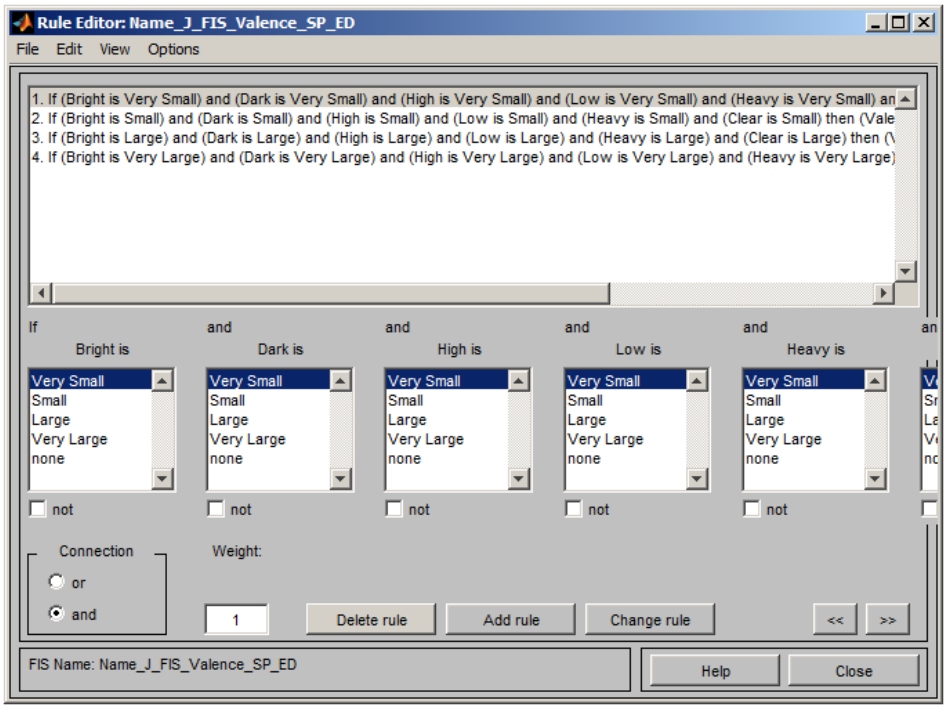


Figure 4.15: If-Then rules derived by ANFIS used for estimating Valence.

The scale for all semantic primitives were { 1, 2, 3, 4, 5 }, and the scale for all emotion dimension were {-2, -1, 0, +1, +2 }. Figure 4.16 shows a sample of the rules using numerical values for example when $Bright = 3.05$ (moderate), and $Dark = 3$ (moderate), and $High = 2.86$ (moderate), and $Low = 5$ (very large), and $Heavy = 4.91$ (very large), and $Clear = 1.27$, (low), Then $Valence = -2$ (very negative), which mean Valence is very negative i.e. the emotional state could be Anger or Sadness the final decision for the emotional state will be determined from the value of the activation. For instance if the Activation is positive then the utterance will be classified as Anger, On the other hand, if the Activation is negative then the emotional state of that utterance will be Sad. The advantages of the dimensional representation is the degree of each emotional state i.e. if the valence is very negative and the activation is very negative then the emotional state will be very Sad. On the other hand if the valence is very positive and the activation is very positive then the emotional state will be very angry.

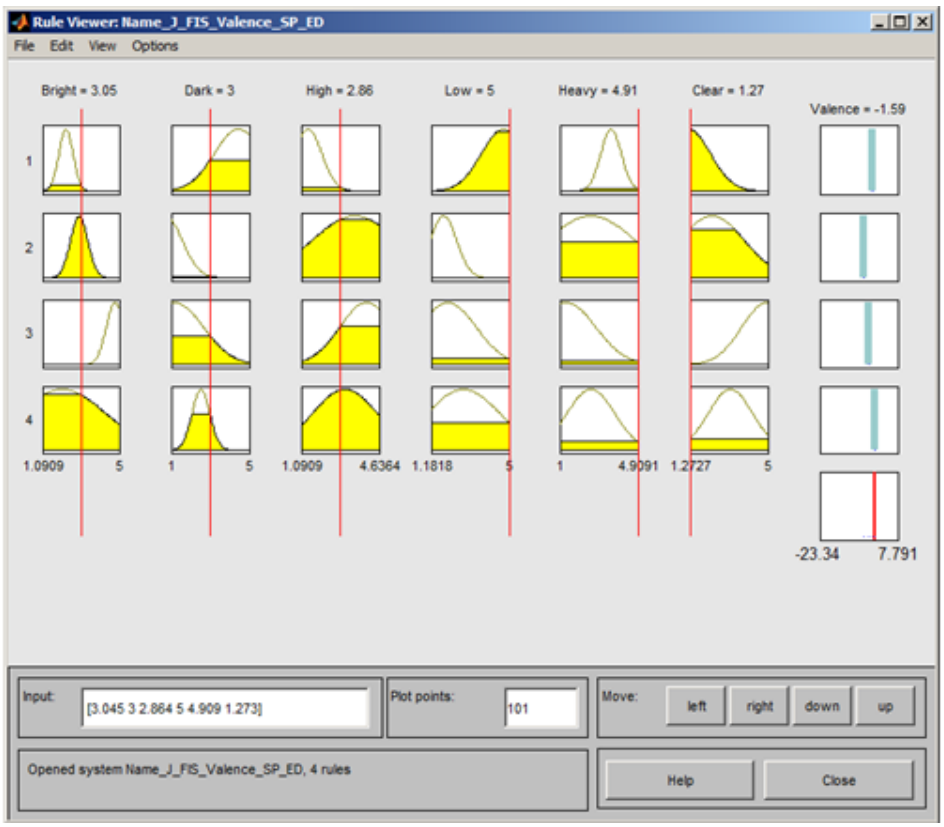


Figure 4.16: Sample of rule set of an ANFIS model for Valence=-2.

Figure 4.17 shows another sample of the rules using numerical values in this figure, the position of the red lines for semantic primitives are changed in order to obtain $Valence = 2$, this way can be used for emotion syntheses i.e in order to make the output Sad what is the required modification for the semantic primitive and for the acoustic features. Therefore, we tried to adjust the red lines for each membership function to obtain $Valence = 2$. The final values for semantic primitives are $Bright = 5$ (very large), and $Dark = 1$ (very small), and $High = 2.86$ (moderate), and $Low = 1.18$ (small), and $Heavy = 1$ (very small), and $Clear = 4.09$, (very large), Then $Valence = +2$ (very positive).

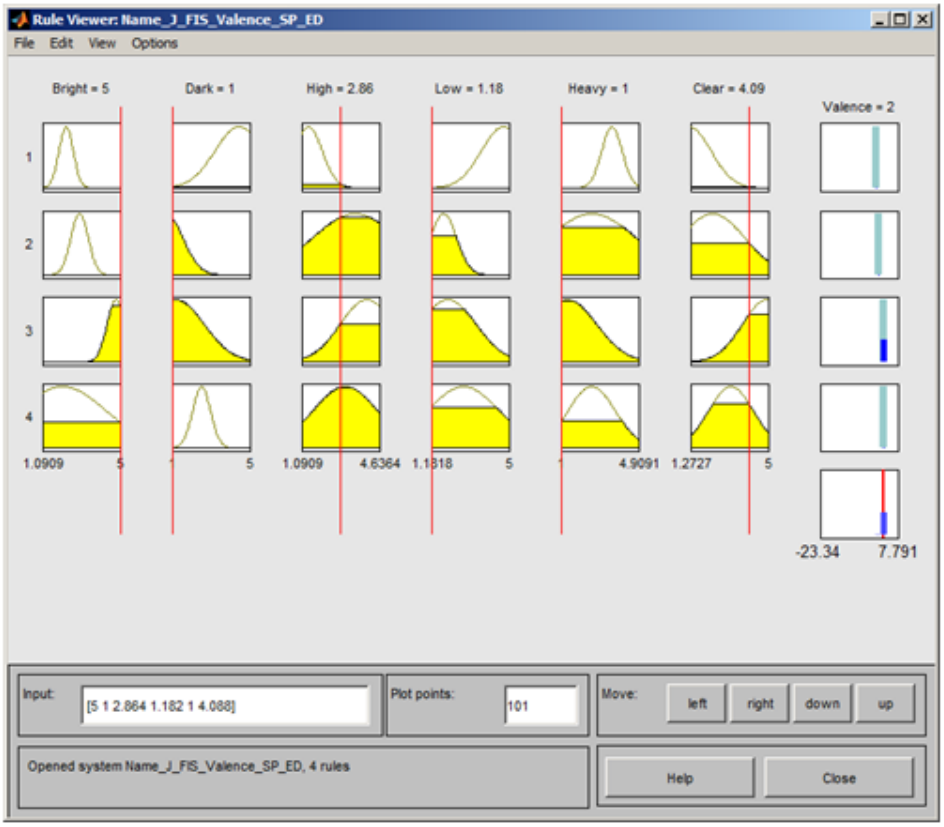


Figure 4.17: Sample of rule set of an ANFIS model for Valence=+2.

4.6 Summary

The aims of this chapter are: attempt to answer the most the following question, what are the most related acoustic features for each emotion dimensions?, try to improve the prediction of emotion dimensions values by constructing a speech emotion recognition system. The proposed idea to improve the prediction of emotion dimensions in this study can be done by imitating the process of human perception for recognizing the emotional state from a speech signal.

Therefore, this study proposes a three-layer model to improve the estimating values of emotion dimensions from acoustic features. The proposed model consists of three layers: emotion dimensions (valence, activation, and dominance) constitute the top layer, semantic primitives the middle layer, and acoustic features the bottom layer.

To answer the first question, we proposed a top-down feature selection method to select the most related acoustic features based on the three-layer model. By firstly, selecting the highly correlated semantic primitives for emotion dimension, then selecting the set of all acoustic features which are highly correlated with the selected semantic. The set of selected acoustic features are considered the most related to the emotion dimension in the top layer.

The most important result is that, using the proposed three-layer model for feature selection, the number of relevant acoustic features to emotion dimensions increases. For example, the number of relevant features for the most difficult dimension valence increases from one to nine using the proposed method. Moreover, the number of features increased from eight to nine for activation and from eight to ten for dominance. The three-layer model outperform the traditional two-layer model for acoustic features selection.

In this chapter the implementation of the proposed model into an automatic speech emotion recognition system is introduced. The bottom-up method was used to estimate emotion dimensions. The input of the proposed system are the selected acoustic features. Fuzzy inference system FIS was used to connect the elements of the proposed system. Firstly one FIS was used to estimate each semantic primitive in the middle layer form

the acoustic features in the bottom layer. Then another FIS was used to estimate each emotion dimensions from the estimated semantic primitives. The detailed evaluation for the proposed system will be introduced in the next chapter.

Chapter 5

Evaluation of the proposed system

5.1 Introduction

The previous chapter introduced a top-down method to find the most relevant acoustic features for each emotion dimension on the bases of a three-layer model for human perception. Moreover, the previous chapter also presented the implementation of the proposed emotion recognition system to estimate emotion dimensions (valence, activation, and dominance) based on a three-layer model. The proposed method for acoustic feature selection and the implementation of the proposed emotion recognition system was based on our assumptions: imitating human perception can guide the selection of new acoustic features with better discrimination for all emotion dimensions, and these selected acoustic features will have a large impact for predicting values of emotion dimensions, especially for valence the most difficult dimension.

In this chapter, we investigate whether our assumptions are satisfied or not. Therefore, we try to answer the following two questions: first, whether the selected acoustic features are effective for predicting emotion dimensions? second, whether the proposed emotion recognition system improve the estimation accuracy of emotion dimensions (valence, activation, and dominance) comparing with the conventional system?

To investigate the first question, first, the most relevant acoustic features for each emotion dimension, which was selected using the feature selection method, were used as inputs of the proposed emotion recognition system, to estimate values of emotion dimensions. Then, the estimation results of emotion dimensions using the most correlated acoustic features are compared with those of estimation using the lower correlated acoustic features, and all-acoustic features.

Furthermore, to investigate the second question which is how effectively our proposed system improve emotion dimensions estimation? Therefore, the performance of the proposed system was compared with that of the conventional two-layer system, using two different languages Japanese and German, with two different tasks (speaker-dependent task and multi-speaker task).

5.2 Evaluation measures

The outputs of the proposed emotion recognition system are the estimated values of emotion dimension valence, activation, and dominance, not a classification into one of a finite set of categories. Therefore, the performance of the proposed system can not measured by recognition accuracy. The estimation of emotion dimensions is preformed by training an automatic emotion recognition system using acoustic features as inputs and annotated emotion dimensions by human subjects as an output. Then, the trained system can be used to estimate emotion dimensions for a new utterance. Therefore, the performance of that emotion recognition system is measured by how close the estimated values using the proposed system with the annotated values by human subjects.

In most of the previous studies, the used metrics to measure the performance of the emotion dimensions estimation was the mean absolute error (MAE) between the estimated values of emotion dimensions and the evaluated values by human subjects. The MAE is the most common parameter to measure the machine learning algorithms performance on estimation tasks, as in our case.

However, to imitate human perception, the proposed emotion recognition system includes two estimation tasks for each emotion dimension, for example for valence dimensions: the first task is to estimate the most related semantic primitives for valence from most related acoustic features for valence; the second task is to estimate valence dimension from the estimated semantic primitives for valence. Therefore, to imitate the human error, the MAEs are evaluated for the estimated semantic primitives and the estimated emotion dimensions. The MAE is used to measure the distance between the estimated values by the proposed system and the annotated value by human subjects. The smaller the MAE the closer estimated value to the human evaluation.

The MAE is calculated for each semantic primitive and each emotion dimension as follows: let $\widehat{x} = \{\widehat{x}_i\}(i = 1, 2, \ldots, N)$ is the sequence of the estimated values of one semantic primitive or one emotion dimension using the proposed system, moreover, let $x = \{x_i\}(i = 1, 2, \ldots, N)$ is the sequence of evaluated values by human subjects for

the corresponding semantic primitive or emotion dimension. Where N is the number of utterances in our database. Then, the mean absolute error MAE is calculated according to the following equation:

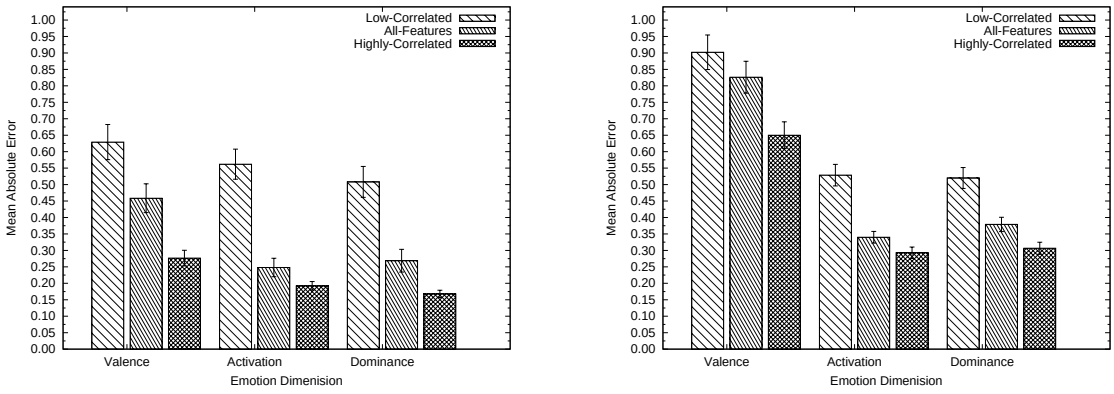
$$MAE = \frac{\sum_{i=1}^{N} |\widehat{x}_i - x_i|}{N} \tag{5.1}$$

5.3 Effectiveness of the selected acoustic features

This section aims to investigate whether the selected acoustic features using the proposed method in Chapter 4 will improve emotion dimensions estimation or not. To accomplish this task, the proposed automatic emotion recognition system was tested using three different groups of acoustic features, for each emotion dimension: (1) highly correlated acoustic features (absolute values of their correlations with semantic primitives is ≥ 0.45), (2) lower correlated acoustic features, and (3) the all-acoustic features. The proposed acoustic features selection method presented in Section 4.3.1 is used to divide all acoustic features into two groups highly correlated and low correlated acoustic features.

To measure the importance of acoustic feature groups, the mean absolute error (MAE) between the predicted values of emotion dimensions using the proposed system and the corresponding average value given by human subjects is used as a metric of the discrimination associated with each group. The MAE is calculated for all emotion dimensions (valence, activation, and dominance) in accordance with Eq. 5.1. The accuracy of the proposed system in terms of five-fold cross validation was calculated for the two databases.

Figures 5.1(a) and 5.1(b) show the MAEs for estimating (valence, activation, and dominance), for Japanese and German database, respectively, for the three groups of acoustic features. The error bars in these figures represent the standard errors for the absolute differences between human evaluation and system estimation for emotion dimensions. Analysis of variance (ANOVA) was used to test whether the three groups are statistically different with respect to the use of correlated acoustic features for emotion dimensions



(a) Japanese Database (Single Speaker)

(b) German Database (Multi Speaker)

Figure 5.1: Mean Absolute Error (MAE) between human evaluation and estimated values of emotion dimensions.

estimation. For Japanese database, at level 0.001, a significant discrimination among the three groups was observed, for valence ($F[2, 534] = 29.30$, $p \leq 0.001$), for activation ($F[2, 534] = 59.28$, $p \leq 0.001$), and for dominance ($F[2, 534] = 51.14$, $p \leq 0.001$). For the German database the results were significant for all emotion dimensions at level 0.001, the information of the F-test were as follows: valence $F[2, 597] = 6.95$, $p \leq 0.001$), activation ($F[2, 597] = 20.06$, $p \leq 0.001$) and dominance ($F[2, 597] = 17.80$, $p \leq 0.001$).

For both databases, the results reveal that, the MAEs by using the selected acoustic features group as an inputs for the proposed system are the smallest compared with the other two of groups features, these results indicate that, the selected acoustic features improve the prediction of all emotion dimensions. Therefore, the selected acoustic feature is effective to improve emotion dimension estimation.

5.4 System Evaluation

In this study, a three-layer model was proposed to improve emotion dimension estimation. In Chapter 4, based on the three-layer model an automatic speech emotion recognition system was implemented. This section presents the evaluation results for the proposed emotion recognition system. In order to explicitly investigate whether the proposed system effectively improve emotion dimensions estimation or not, the performance of the proposed

system was compared with that of the conventional two-layer system by using two different languages: Japanese and German, using two different tasks (1) speaker-dependent task, and (2) multi-speaker task.

To accomplish these tasks, the proposed and the conventional emotion recognition system were constructed. The selected acoustic features group was used as input for both the proposed system and the conventional system. The proposed system was constructed based on the three-layer model of human perception as follows: one FIS was used to estimate each semantic primitive from the selected acoustic features, then one FIS was used to estimate each emotion dimension from the estimated semantic primitives as described in Chapter 4. For constructing the conventional system based on the two-layer model, one FIS was used to estimate each emotion dimension from the selected acoustic features directly as described in Chapter 2.

Section 5.4.1 introduces the evaluation using the speaker dependent task for both Japanese and German language. Moreover, Section 5.4.2 presents the evaluation results in the multi-speaker task for German language.

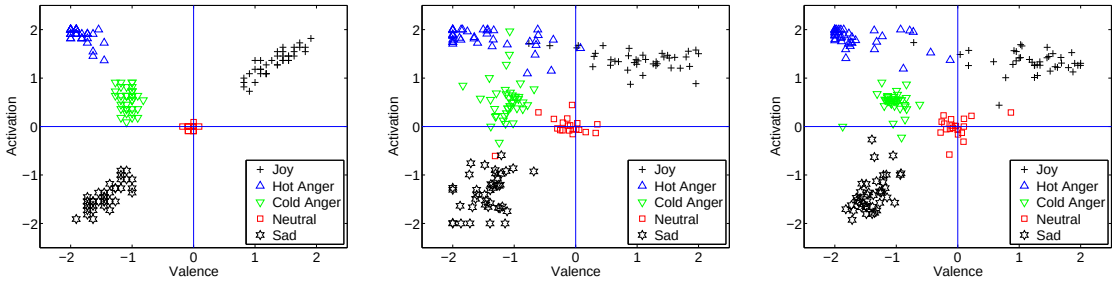
5.4.1 Evaluation Results for Speaker-dependent Task

In the speaker-dependent task, the automatic emotion recognition system was trained and tested using utterances for one speaker.

5.4.1.1 System evaluation for Japanese database

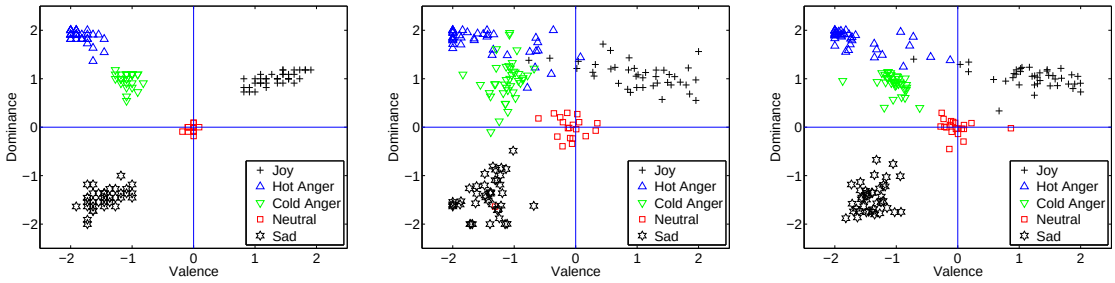
Japanese database is a single speaker database therefore, both conventional and proposed emotion recognition systems were validated using all 179 utterances included in the Japanese database. The 5-fold cross validation was used to evaluate the automatic systems.

The distribution of output of the automatic systems as well as the human evaluation are shown in scatter-plot of Valence-Activation, Valence-Dominance, and Activation-Dominance in Figures 5.2, 5.3, and 5.4 respectively.



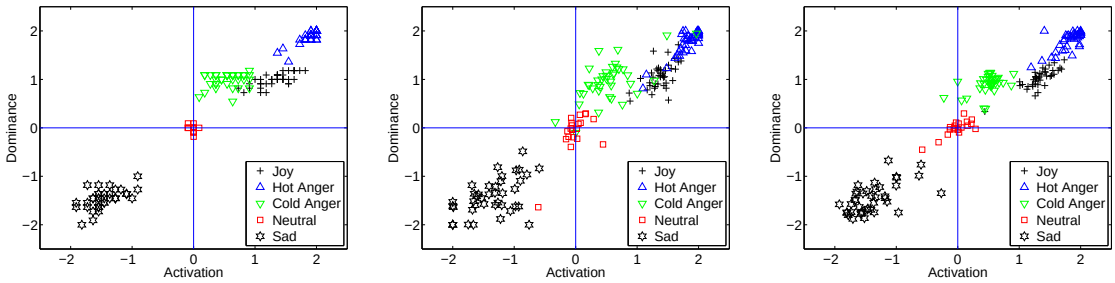
(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.2: The distribution of Japanese database in the Valence-Activation space.



(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.3: The distribution of Japanese database in the Valence-Dominance space.



(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.4: The distribution of Japanese database in the Activation-Dominance space.

For each space there are three panels (a), (b), (c); the most left panel shows the distribution of human evaluation, and the middle panel (b) shows the estimation of the conventional system, while (c) presents the distribution of the proposed system. The emotional state of each utterance is represented by one point in the dimensional space. From the activation-dominance space, it is observed that the estimated values of activation and dominance are equivalent, the reason behind this is that our databases does not include fear emotional state and mainly dominance dimensions is used to distinguish between the anger and fear emotional state, therefore the dominance dimension is considered redundant dimension.

Our proposed emotion recognition system is based on the three-layer of human perception. Therefore, the MAEs are evaluated for the estimated semantic primitives and the estimated emotion dimensions, to imitate the human error for estimating semantic primitives and emotion dimensions, respectively. Figure 5.5 shows the MAEs for the most related semantic primitives for the valence dimension. From this figure it is clearly that the MAEs for all semantic primitives were very close to zero, which means that the estimated values using the proposed system are very close to the evaluated values using human subjects the maximum MAE value was 0.3 which achieved by clear semantic primitive.

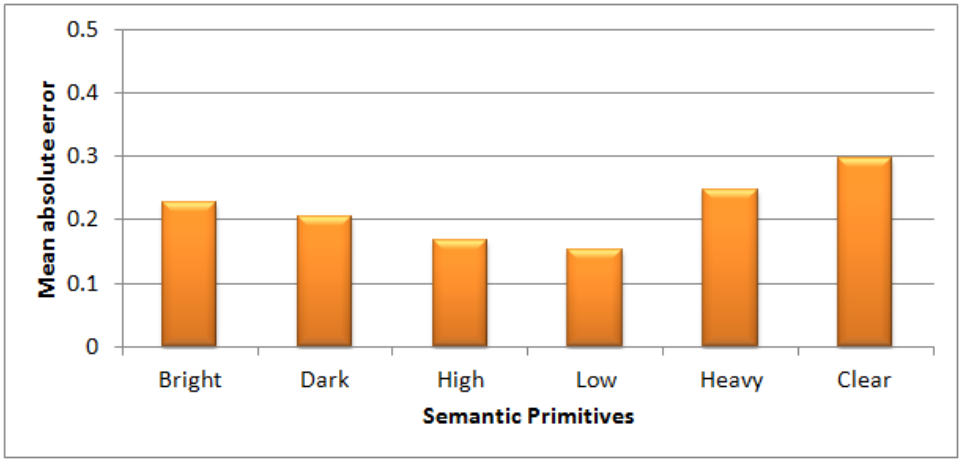


Figure 5.5: MAE for the most related semantic primitives for valence estimated from the most related acoustic features for valence for Japanese database (Single-speaker).

In order to asses the performance of the proposed emotion recognition system the final output for conventional and the proposed system are compared, therefore, the MAEs for emotion dimensions estimation from the both of them were compared as follows: the MAEs for emotion dimensions (valence, activation, and dominance) between the two systems output and human evaluation are calculated using Eq. 5.1.

The results are shown in Fig. 5.6. The error bars represent standard errors. Using t-test, at level 0.05, the results for all emotion dimensions are as follows: valence ($t(178)=3.16$, $p\leq 0.05$), activation ($t(178)=2.47$, $p\leq 0.05$), and dominance ($t(178)=4.99$, $p\leq 0.05$). These results are statistically significant for all emotion dimensions.

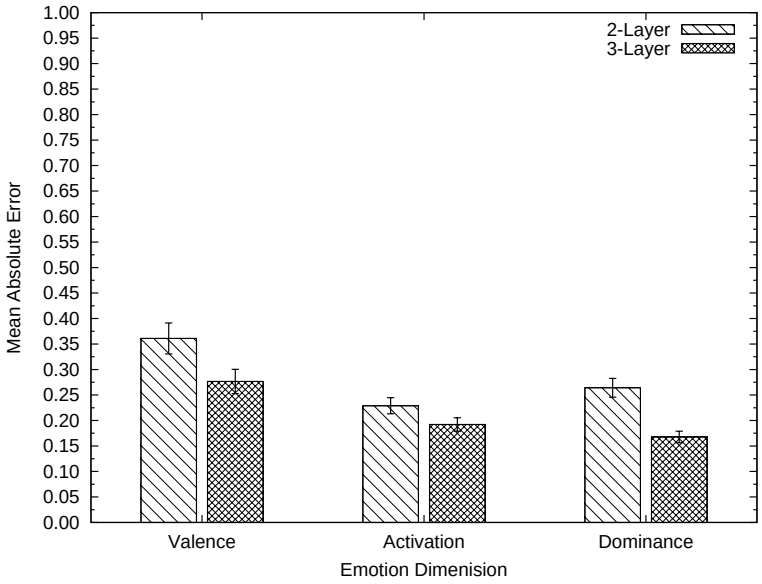


Figure 5.6: MAE between human evaluation and two systems outputs (two-layer and three-layer system) for Japanese database (Single-speaker).

These results suggest that the proposed system outperform the conventional two-layer system, for all emotion dimensions. This results reveals that the proposed system outperform the conventional one for evaluating the valence and dominance. The most important results The result is that the MAEs values for emotion dimensions valence, activation, and dominance were 0.28, 0.19, and 0.17, respectively. These results indicate that emotion dimensions estimation using the proposed system is very close to human evaluation.

Table 5.1: Number of utterances used for each speaker from Berlin database In the first column is the speaker ID M03 means male, 03 is the speaker code used in the database

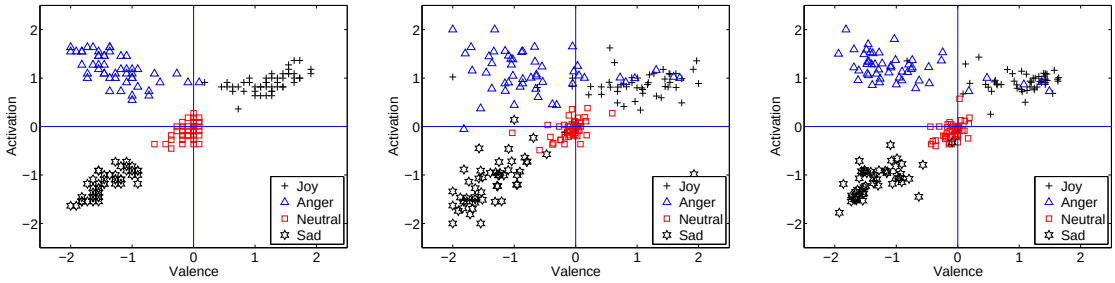
Speaker ID	Total
M03	24
F08	27
F09	15
M10	14
M11	25
M12	15
F13	20
F14	17
M15	22
F16	21

5.4.1.2 System evaluation for German database

The German database contained ten speakers: five male and five female. The number of utterances for each German speaker are small. Table 5.1 shows the number of utterance for each speaker in the selected German database. Therefore, leave-one-out-cross-validation (LOOCV) was used for evaluation for each German speaker. The proposed system and the conventional two-layer system were used to estimate emotion dimensions by training the systems and testing them using the utterances for each speaker individually.

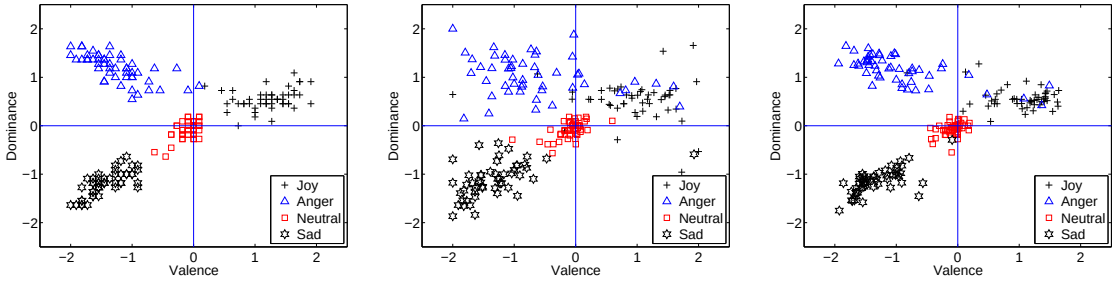
The distribution of emotion dimension estimation for all utterances for all speakers are presented together in the valence-activation space, valence-dominance space, and activation-dominance space as shown in Figures 5.7, 5.8, and 5.9, respectively.

The MAEs for the estimated semantic primitives and the estimated emotion dimensions, are evaluated for each speaker individually. Figure 5.10 shows the average of MAEs for the most related semantic primitives for the valence dimension, from the ten German speakers. These results indicate that the MAEs for all semantic primitives were very close to zero, which means that the estimated values using the proposed system are very close to the human evaluation. The maximum MAE value was 0.24 which achieved by heavy semantic primitive.



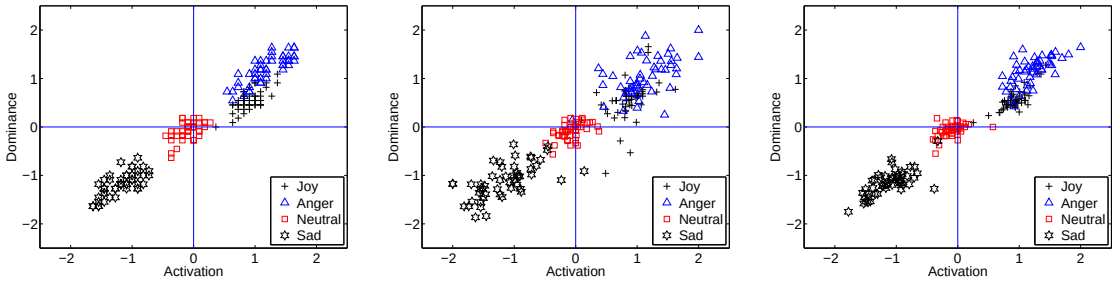
(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.7: The distribution of all German speakers' utterances in the Activation-Dominance space.



(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.8: The distribution of all German speakers' utterances in the Activation-Dominance space.



(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.9: The distribution of all German speakers' utterances in the Activation-Dominance space.

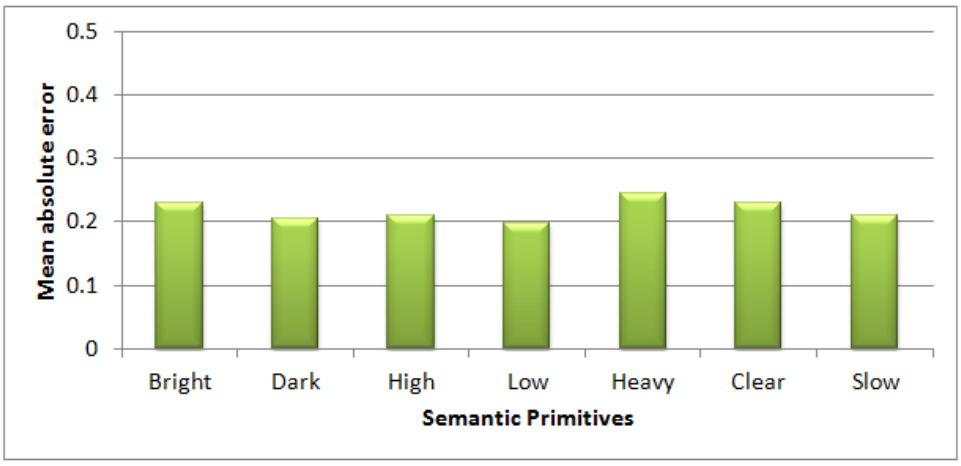
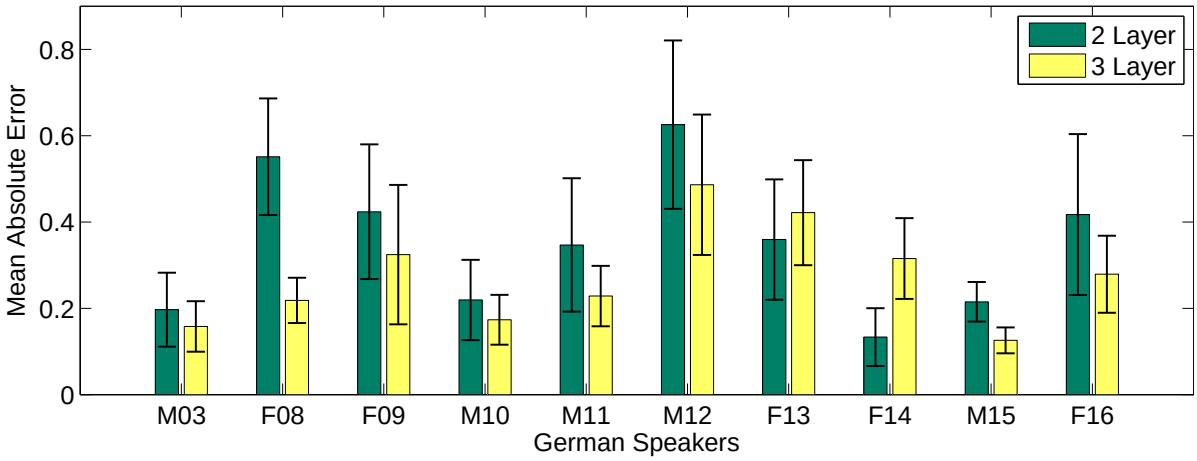


Figure 5.10: The average of MAEs for the most related semantic primitives for the valence dimension, from the estimation using ten German speakers. (Speaker-dependent).

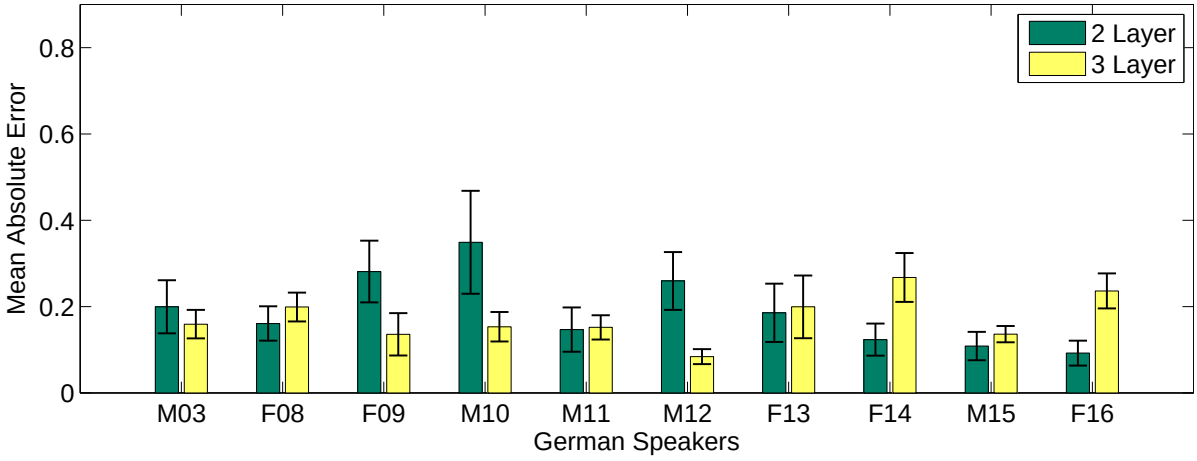
In order to asses the performance of the proposed emotion recognition system the final output for the the proposed and conventional system are compared, therefore, the MAEs for emotion dimensions estimation from the both of them were calculated by training and testing the system using the ten German speakers. Figure 5.11 shows the MAEs for valence, activation, and dominance for the ten speakers, in panel (a), (b), and (c), respectively.

Finally, the average of all MAEs from all speakers for each emotion dimension was calculated. The results are presented in Figure 5.12. In order to compare between the two-layer system and the proposed system the paired t-tests was used, the results are statistically significant for valence ($t(199)=2.09$, $p \leq 0.05$) and dominance ($t(199)=1.78$, $p \leq 0.05$), but there is no significant differences for activation between the two-layer and the three-layer model ($t(199)=0.23$, p-value=0.41). These results reveal that the proposed system outperform the conventional one for evaluating the valence and dominance.

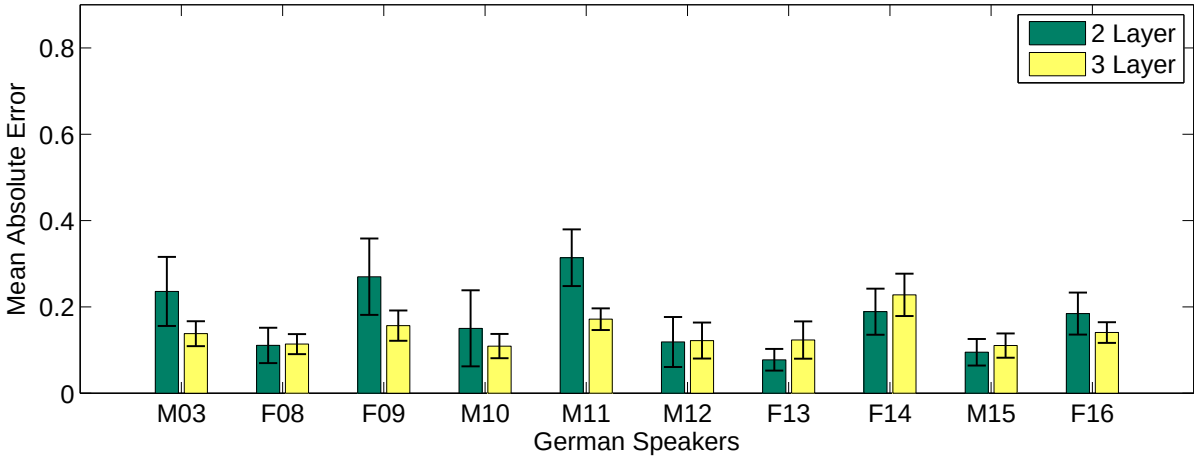
The most important result is that, the estimation results using the proposed three-layer system are very close to human evaluation, as demonstrated by the small values of MAEs for all emotion dimensions, as can be seen from Figure 5.12. The results of MAEs were 0.27, 0.17, and 0.14 for valence, activation, and dominance, respectively.



(a) Valence.



(b) Activation.



(c) Dominance.

Figure 5.11: Mean Absolute error between human evaluation and the automatic systems estimation for 10 German Speakers individually.

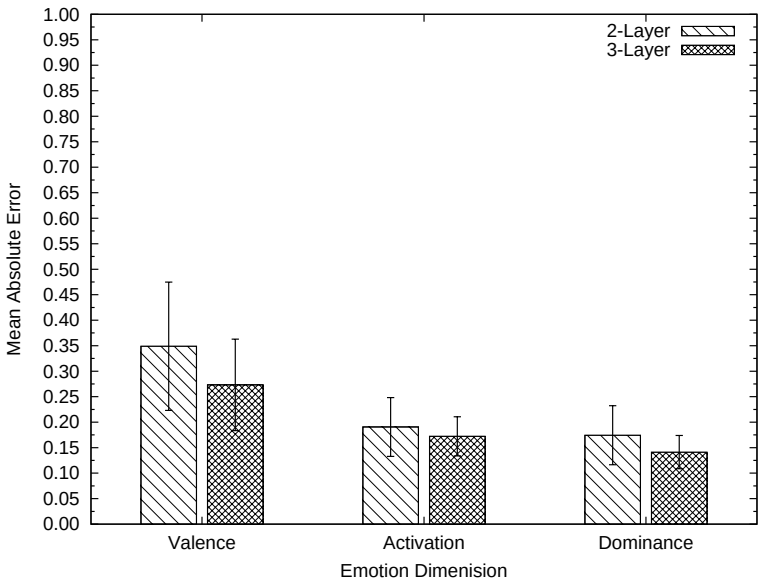


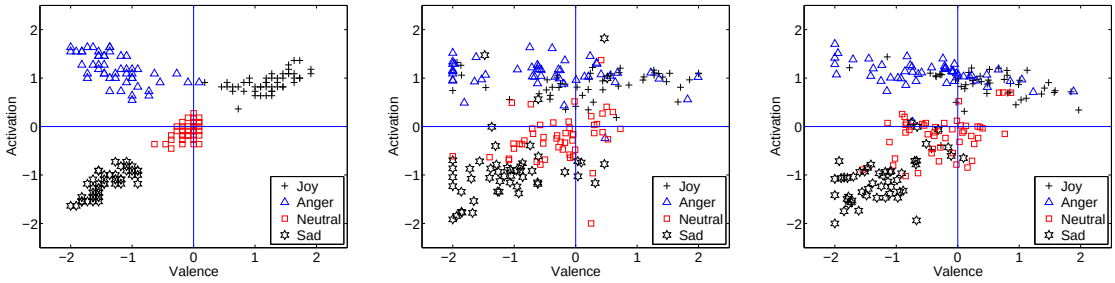
Figure 5.12: MAE between human evaluation and two systems outputs (two-layer and three-layer system) for German database (speaker-dependent).

5.4.2 Evaluation Results for Multi-Speaker Task

In the previous section the proposed emotion recognition system was used to estimate emotion dimensions for speaker-dependent task. The purpose of this section is to investigate whether the the proposed system is also effective in case multi-speaker task, and what is the performance comparing to the traditional two-layer system.

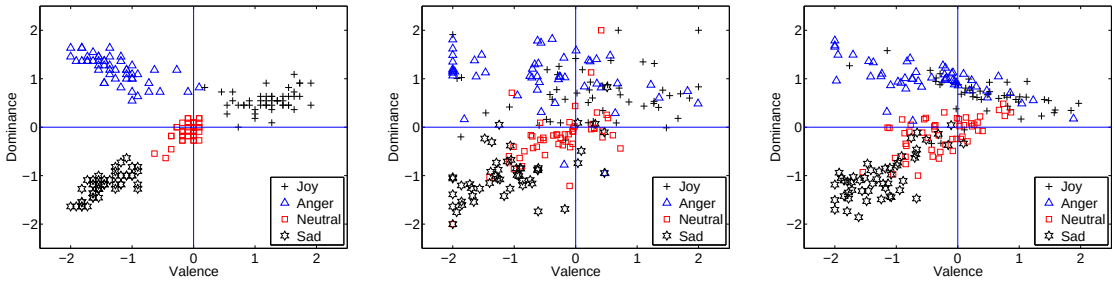
German database contains 200 utterances selected from the Berlin database uttered by 10 speakers. Therefore, German database was used to investigate the multi-speaker effect on emotion dimension estimation. Thus, the proposed system was validated using the whole database i.e. all 200 utterances were used to train and test the emotion recognition system. Five-fold cross validation was used to evaluate this system, i.e, the German database is divided into 80% training set (160 utterances for training) and 20% testing set (40 utterances for testing).

The distribution of human evaluation, traditional method estimation, and the proposed system estimation are shown in Figures 5.13, 5.14, and 5.15 in the valence-activation space, valence-dominance space, and activation-dominance space respectively.



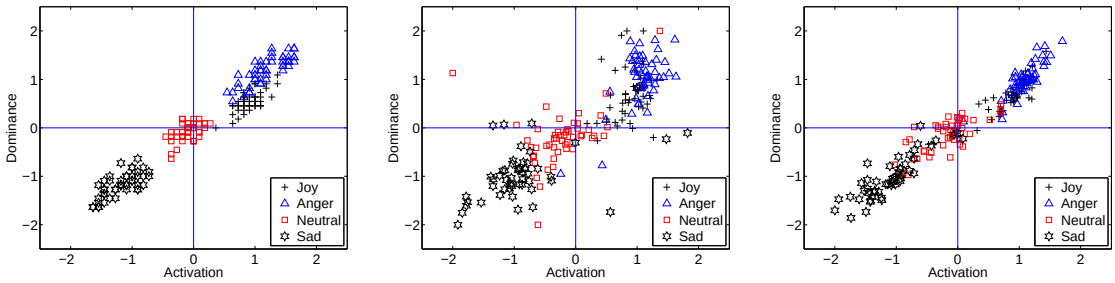
(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.13: The distribution of German database in the Valence-Activation space.



(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.14: The distribution of German database in the Valence-Dominance space.



(a) Manually labeled by Human. (b) Estimated using Two-Layer. (c) Estimated using Three-Layer.

Figure 5.15: The distribution of German database in the Activation-Dominance space.

The MAEs are calculated for the estimated semantic primitives and the estimated emotion dimensions. Figure 5.16 shows the MAEs for the most related semantic primitives for the valence dimension in case of multi-speaker task. The values of MAEs for all semantic primitives ranged from 0.36 to 0.46, obtained for low and heavy semantic primitive, respectively. These results are higher then those results achieved by the speaker dependent task. These results indicate that the estimated values of semantic primitives are not so close to the evaluated values by human subjects.

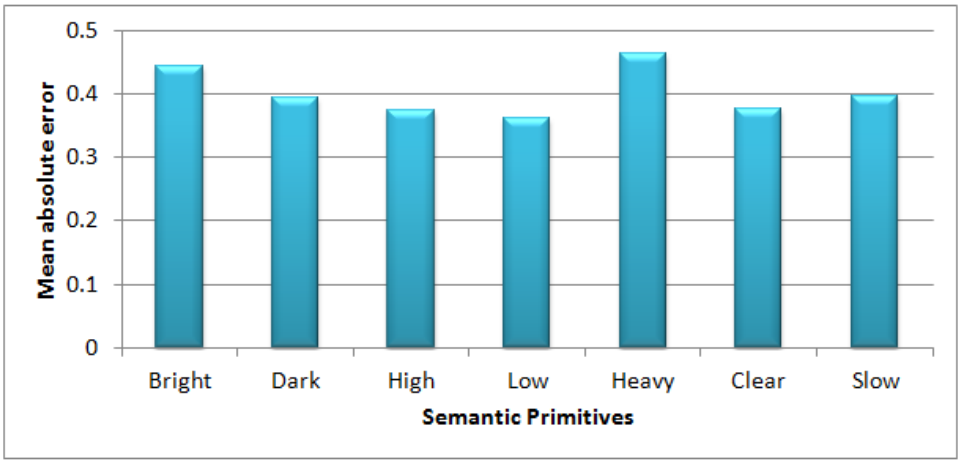


Figure 5.16: MAE for the most related semantic primitives for valence estimated using the most related acoustic features for valence for German database (multi-speaker).

In order to asses the performance of the proposed emotion recognition system, the estimated emotion dimensions for conventional two-layer system and the proposed system are compared. The MAEs for all emotion dimensions from the both of systems were compared as follows. The results for multi-speaker evaluation are shown in Figure 5.17. The error bars represent standard errors. The results of the paired t-test at 0.05 significant level were as follows: valence ($t(199)=2.83$, $p\leq 0.05$), activation ($t(199)=1.93$, $p\leq 0.05$), and dominance ($t(199)=3.38$, $p\leq 0.05$). These results are statistically significant for all emotion dimensions. These results reveal that the proposed system outperforms the conventional one in the multi-speaker task.

The results of MAEs for emotion dimensions valence, activation, and dominance were 0.65, 0.19, and 0.17, respectively. The MAE for valence was very high comparing to the

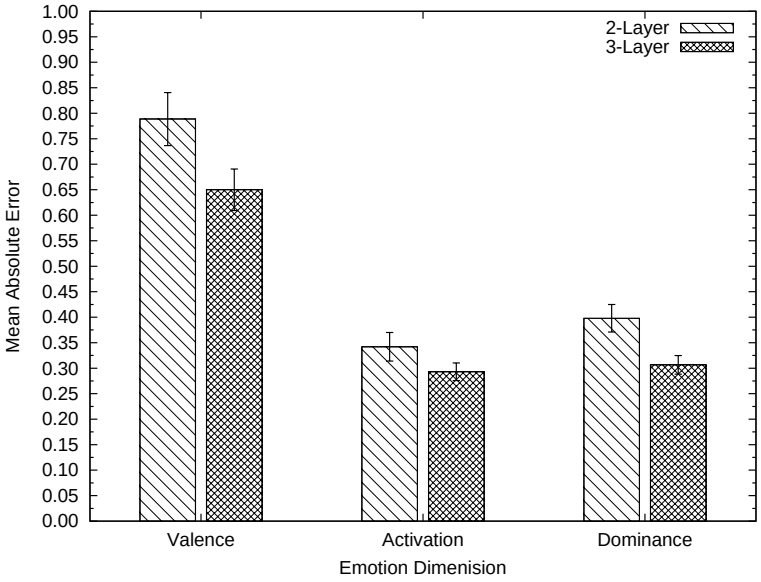


Figure 5.17: German Database (multi-speaker): MAE between human evaluation and two systems' output.

results obtained in case of speaker-dependent. However, the MAEs for activation and dominance still very close zero, which reveal that the estimation results for activation and dominance are very close to human evaluation, as shown in Figure 5.12.

5.4.3 Discussion

In order to investigate whether the selected acoustic features are effective for emotion dimension estimation, first, the most relevant acoustic features were selected for each emotion dimensions, for Japanese and German databases as described in Chapter 4. Then, the proposed emotion recognition system was tested using three different groups of acoustic features: (most relevant, not relevant, and all) acoustic features. The highest performance for all emotion dimensions were achieved using the group of the most relevant acoustic features, demonstrated by the smallest values of MAEs for both databases.

To investigate whether the proposed system improve the estimation results of emotion dimensions, the performance of the proposed system was compared with that of the conventional two-layer system, using two different languages Japanese and German, with two different tasks (speaker-dependent task and multi-speaker task).

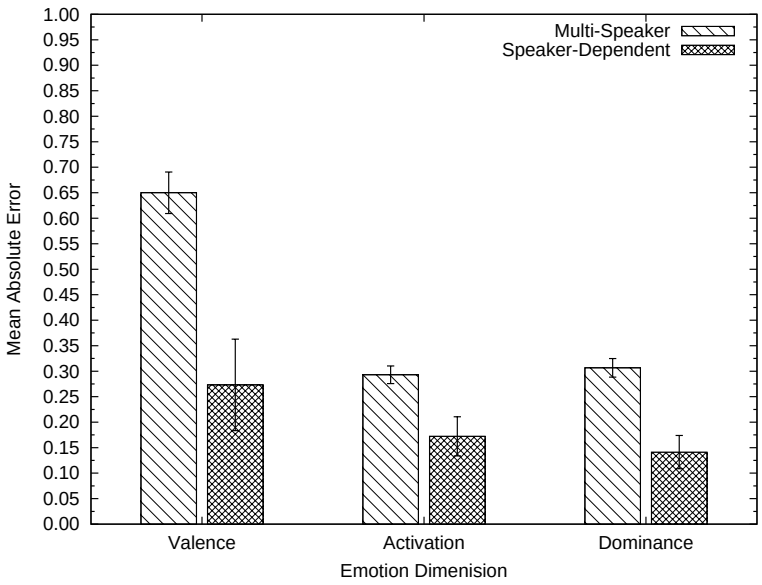


Figure 5.18: Comparison between MAE between human evaluation and two systems' output for multi-speaker task and Speaker-dependent task.

The results of speaker-dependent task were as follows: The MAEs for Japanese database were 0.28, 0.19, and 0.17 for valence, activation, and dominance, respectively. Moreover, for German database, the MAEs were 0.27, 0.17, and 0.14 for valence, activation, and dominance, respectively. These values indicate that the error between human evaluation and system outputs are very small and close to zero, which means that the estimated values using the proposed emotion recognition system are very close to human evaluation.

However, the MAEs of multi-speaker emotion recognition task for German database, were 0.65, 0.29, and 0.31 for valence, activation, and dominance, respectively. Figure 5.18 shows the MAEs for German database for both speaker-dependent and multi-speakers. It clearly from this figure that the accuracy of all emotion dimensions is greatly improved when using speaker-dependent emotion recognition. The MAE for valence dimension decreased from 0.65 using the multi-speaker emotion recognition to 0.27 using speaker-dependent emotion recognition, while activation decreased from 0.19 to 0.17, and the dominance decreased from 0.17 to 0.14, the great improvement was for the valence dimension.

5.5 Summary

The aim of this chapter is to investigate whether the following assumptions are satisfied or not: (1) whether the selected acoustic features using the feature selection method are effective for predicting emotion dimensions? (2) whether the proposed emotion recognition system improve the estimation accuracy of emotion dimensions (valence, activation, and dominance) comparing with the conventional two-layer system?

First, the most relevant acoustic features were selected for each emotion dimensions, for Japanese and German databases as described in Chapter 4. Then, the proposed emotion recognition system was tested using three different groups of acoustic features: (most relevant, not relevant, and all) acoustic features. The results revel that the best performance for all emotion dimensions were achieved using the selected acoustic features for both databases.

Furthermore, the performance of the proposed system was compared with that of the conventional two-layer system, using two different languages Japanese and German, with two different tasks (speaker-dependent task and multi-speaker task). The results reveal that the proposed three-layer emotion recognition system is effective and gives the best results for all emotion dimensions for both speaker-dependent and multi-speaker task. However, for both language databases, the highest performance achieved for speaker-dependent task, demonstrated by the very small values of MAEs for all emotion dimensions.

Chapter 6

Cross-lingual Speech Emotion Recognition System

6.1 Introduction

Speech is the most natural and important means of human-human communication in our daily life, when we use the same language. Even without the understanding of one language, we can still judge the expressive content of a voice, such as emotions. An interesting question to ask is whether emotional states can be recognized universally or not. Several studies have indeed shown evidence for certain universal attributes for both speech [7, 46] and music [83, 60], not only among individuals of the same culture, but also across cultures. Therefore, it is interesting to build an automatic speech-emotion recognition system that has the ability to detect the emotional state regardless of the input language.

Most of the previous studies for automatic speech emotion recognition were based on detecting the emotional state working on mono-language, i.e. training and testing the automatic emotion recognition system using only one language database. However, in order to develop a generalized emotion recognition system, the performance of these systems must be analyzed in mono-language as well as cross-language. The ultimate goal of this chapter is to construct a cross-lingual emotion recognition system that has the ability to estimate emotion dimensions for one language by training the system using another language. Therefore, the question we try to answer in this chapter is; whether our proposed automatic emotion recognition system described in Chapter 4 is able to estimate emotion dimensions valence, activation, dominance cross-lingually?

6.2 Cross-language emotion recognition system

In order to accomplish this task, we follow these steps: firstly, a variety of acoustic features were extracted for two different language databases: Japanese and German language, as described in Chapter 2. Then, the proposed feature selection method was used to select the most relevant acoustic features for each emotion dimension for the two databases as described in Chapter 4. In this chapter, we investigate whether the two databases share

some common acoustic features and semantic primitives which allow us to estimate emotion dimensions cross-lingually as described in Section 6.2.1. Then, a cross-language speech emotion recognition system based on the three-layer model was constructed, the input of this system are the common acoustic features between the two languages as presented in Section 6.2.2. Finally, the proposed cross-language emotion recognition system was validated by training the system using one language and testing using the second language as introduced in Section 6.3. 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 the proposed cross-language emotion recognition system, then the trained system is used to estimate emotion dimensions for the Japanese database, and vice versa. To avoid multi-speaker variation, the proposed system is trained using all utterances for one speaker from one language, and tested using utterances for another speaker from the second language.

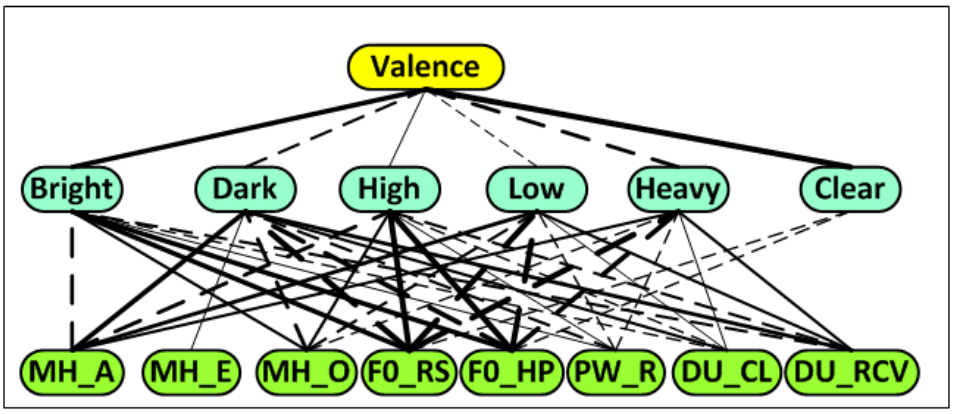
6.2.1 Feature selection for the cross-language emotion recognition system

This section investigates whether their are some common acoustic features and semantic primitives for the two languages. Therefore, a perceptual three-layer model was constructed for each emotion dimension in case of cross-language. To construct this model, a perceptual three-layer model was constructed for each emotion dimension individually for both language, then the common acoustic features between each model for the two languages were selected to constitute the bottom layer for the cross-language perceptual model. Moreover, the common semantic primitives between each model for the two-languages were selected, to constitute the middle layer for the cross-language model.

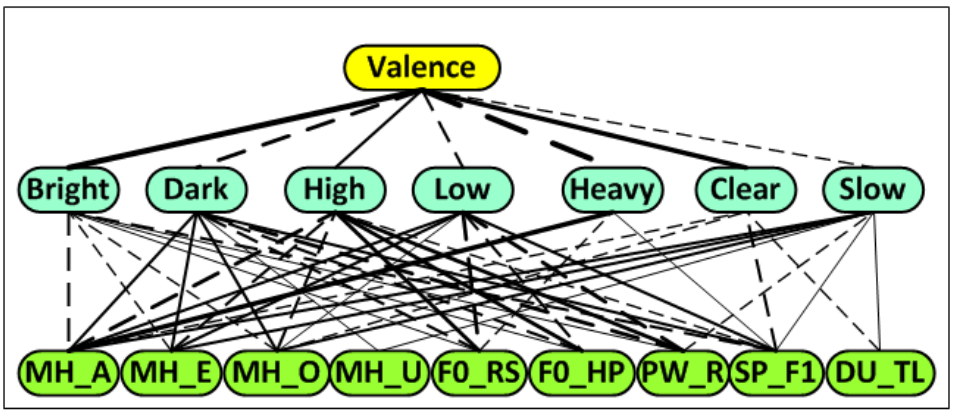
6.2.1.1 Acoustic feature and semantic primitives selection

Using feature selection method described in Chapter 4, firstly, the most relevant semantic primitives were selected for each emotion dimension. Secondly, the most relevant acoustic

features for each semantic primitive were selected. Finally, a perceptual three-layer model was constructed for each emotion dimension as follows: the emotion dimensions in the top layer, and the most relevant semantic primitives for this dimension are in the middle layer, while the relevant acoustic features are in the bottom layer. Fig. 6.1 illustrates the valence perceptual models for Japanese and German database, respectively, where the solid lines in this figure represent a positive correlation, and the dashed ones indicate a negative correlation. The thickness of each line indicates the strength of the correlation; the thicker the line is, the higher the correlation.



(a) Japanese database.



(b) German database.

Figure 6.1: The perceptual three-layer model for valence.

The valence perceptual model for German and Japanese language are compared as follows: For both languages, the valence dimension is found to be positively correlated with Bright, High and Clear semantic primitives, while it is negatively correlated with Dark, Low, and Heavy semantic primitives. Therefore, the two languages not only share

Table 6.1: The elements in the perceptual three-layer model for Valence dimensions for cross-language emotion recognition system, using Japanese and German language, the first indicate the position of the layer in the model, the second column is the elements in each layer, the third is the number of elements in each layer

Layer	Elements in each layer	Number
Top layer	Valence	1
Middle Layer	Bright, Dark, High, Low, Heavy, Clear	6
Bottom Layer	MH_A, MH_E, MH_O, F0_RS, F0_HP, PW_R	6

six semantic primitives but also share similar correlations between the emotion dimensions and the corresponding semantic primitives.

In addition, comparing the relationship between semantic primitives and acoustic features, it is found that the six semantic primitives that were shared by both German and Japanese have a similar correlations with six common acoustic features (MH_A, MH_E, MH_O, F0_RS, F0_HP, and PW_R). This finding suggests the possibility of some type of universality of acoustic cues associated with semantic primitives. Therefore, the proposed method can be used effectively to select the most relevant acoustic features for each emotion dimension regardless the used language.

Using these common acoustic features and semantic primitives for valence dimension, we can easily build the valence perceptual model for the cross-language mode. The elements in each layer for the cross-language perceptual three-layer model for valence dimension are listed in Table 6.1 as follow: valence dimension in the top layer, six common semantic primitives in the middle layer, and six common acoustic features in the bottom layer.

6.2.2 The proposed cross-language speech emotion recognition system

In the previous section a cross-language perceptual three-layer model were constructed for each emotion dimension, by selecting the common acoustic features and the common semantic primitives between the two languages. This section introduces the proposed cross-language emotion recognition system using a bottom-up method for the perceptual

models which were constructed in the previous section.

Figure 6.2 shows the block diagram of the proposed cross-language emotion recognition system for estimating the valence dimension. To estimate the valence dimension the following steps are preformed: estimating values of the six semantic primitives in the middle layer from the six acoustic features in the bottom layer using six FISs one for estimating each semantic primitive, as shown in Fig. 6.2. In addition, one FIS was needed to estimate the value of the valence dimension from the six estimated semantic primitives. In a similar way, the activation and dominance can be estimated using FIS for each semantic primitive, and one FIS for the activation and dominance, respectively.

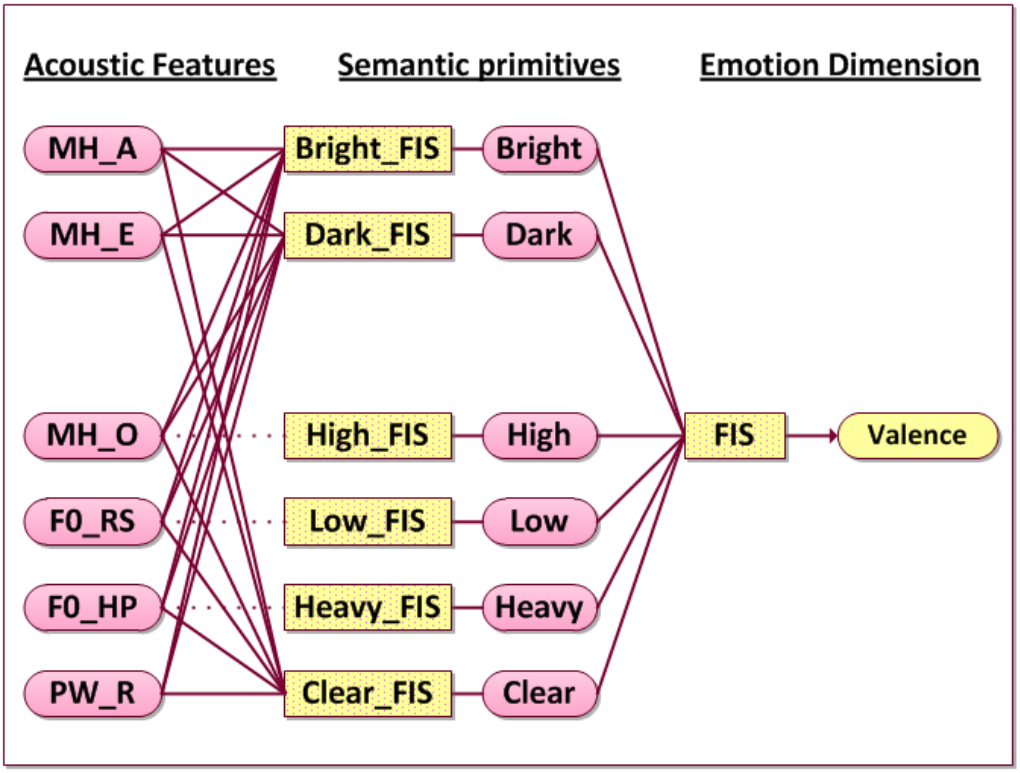


Figure 6.2: Block diagram of the proposed cross-language emotion recognition system for estimating valence dimension.

6.3 System Evaluation

The aim of this chapter, is to investigate whether an automatic emotion recognition system trained using one language has the ability to detect the emotion dimension from

another languages. To accomplish this task, the common acoustic features and semantic primitives between Japanese and German databases were investigated, as explained in section 6.2.1. For example, in the case of the valence dimension, it was found that their were six common acoustic features and also six common semantic primitives for both databases. These acoustic features were used as input to the proposed system, as shown in Fig. 6.2.

To estimate emotion dimensions for Japanese language from German language, the proposed cross-language emotion recognition system is trained using the acoustic features, semantic primitives, and emotion dimensions for the German language. Then, the acoustic features for Japanese language are used as an input to the trained system to estimate emotion dimensions for Japanese language, and vice-versa. The mean absolute error MAE between the predicted values of emotion dimensions and the corresponding average value given by human subjects is used as a metric of the discrimination associated with each case. The MAE is calculated according to the following equation

$$MAE^{(j)} = \frac{\sum_{i=1}^{N} |\widehat{x}_i^{(j)} - x_i^{(j)}|}{N} \tag{6.1}$$

where $j \in \{Valence, Activation, Dominance\}$, $\widehat{x}_i^{(j)}$ is output of the emotion recognition system, and $x_i^{(j)}$ is the values evaluated by the human subjects.

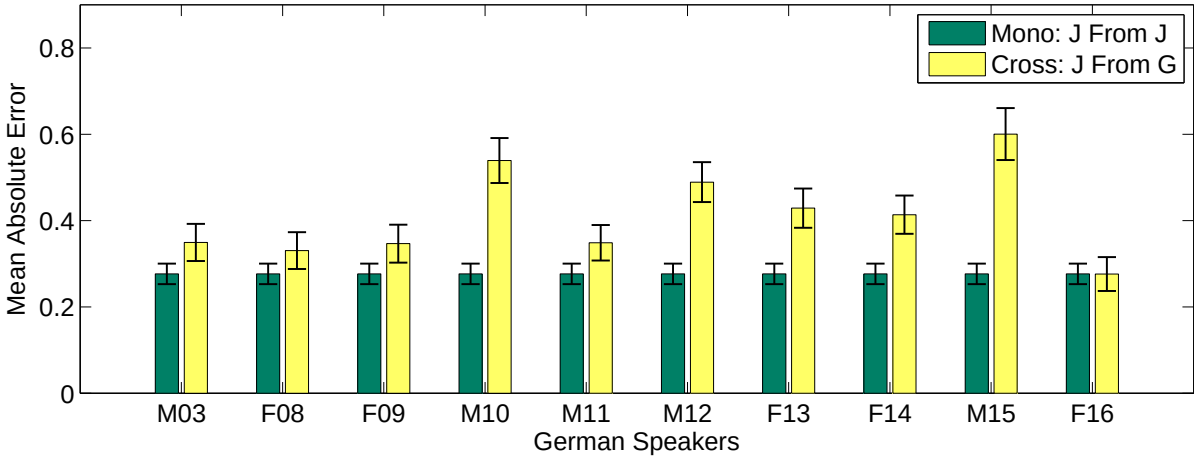
The results of estimating emotion dimensions using the cross-language emotion recognition system were compared with those of the mono-language emotion recognition system evaluated in Chapter 5. Section 6.3.1 explains the results of estimating Japanese emotion dimensions from a trained cross-language emotion recognition system using German database, while Section 6.3.2 presents the results of estimating German emotion dimensions from a trained cross-language emotion recognition system using German database. To avoid the multi-speaker variation, the evaluation of cross-language emotion recognition system was conducted by training the system using one speaker from one language, and testing the system using one speaker from the other language.

6.3.1 Japanese emotion dimensions estimation from German database

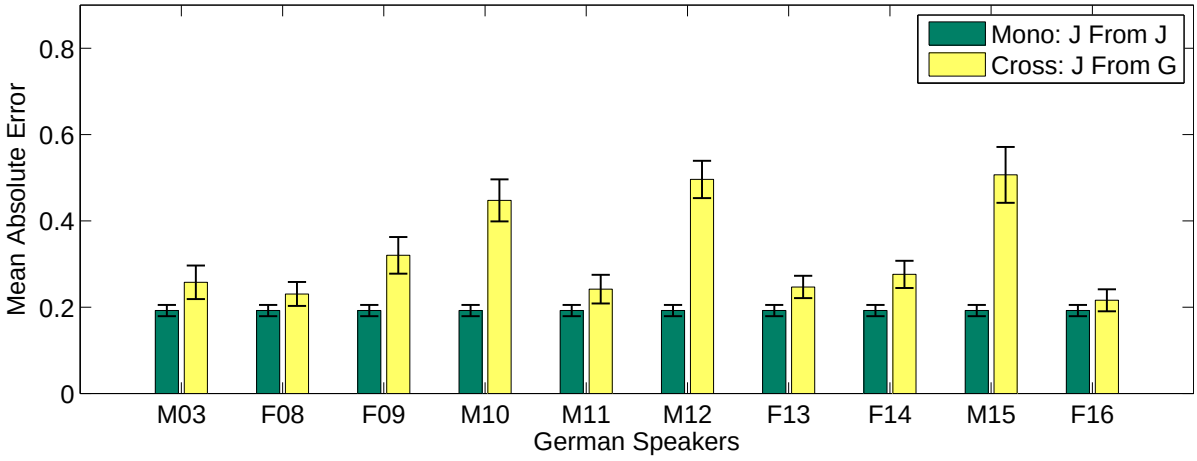
For estimating Japanese emotion dimensions from German database, the cross-language emotion recognition system was trained 10 times, one for each German speaker. Moreover, the Japanese database was tested 10 times, one time for each German speaker. For a comparative analysis of the performance of the proposed cross-language emotion recognition system, the results were compared with those of the mono-language emotion recognition system, which was trained and tested using the Japanese database. The mean absolute error MAE for all emotion dimensions, for the mono-language (Japanese-from-Japanese) emotion recognition system and the cross-language (Japanese-from-German) emotion recognition system are illustrated in Figures 6.3(a), 6.3(b), and 6.3(c). These figures show that the MAEs between human evaluation and estimated emotion dimensions from the 10 German speakers as well as the MAEs for estimating emotion dimensions from Japanese database.

From Figures 6.3(a) and 6.3(b) it is clearly that the estimation of valence and activation dimension from 10 German speakers were very close to the estimation from Japanese database except in three different speakers M10, M12, and M15, the difference between MAE were the highest. Therefore, we can conclude that using the cross-language emotion recognition system the MAE is increased for estimating Japanese emotion dimensions, however this increment were very small.

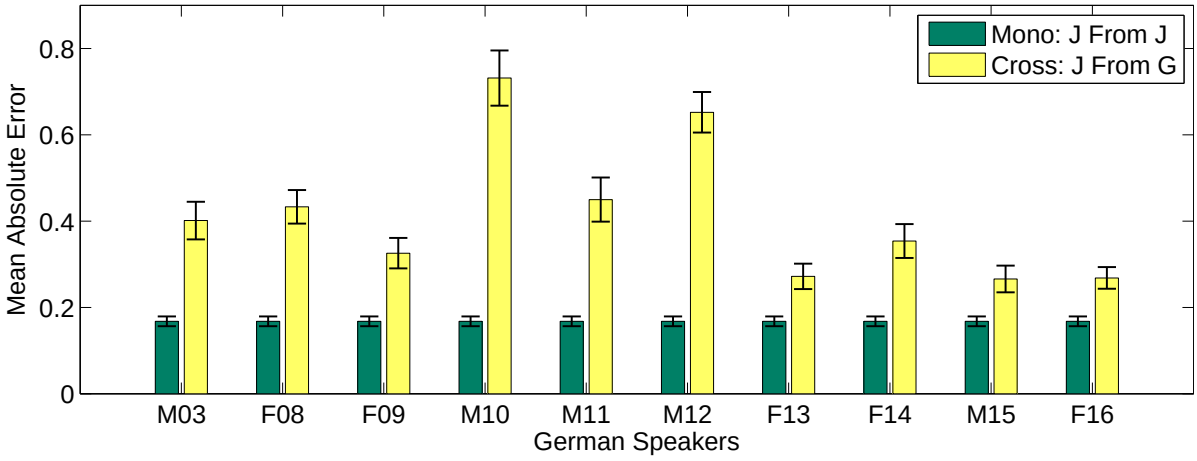
Finally, in order to determine the over all estimation for the whole estimation from all German speakers, each utterance in the Japanese database is estimated 10 times for each emotion dimensions (valence, activation, and dominance), one from each German speaker. For each emotion dimension, the average value from the 10 estimations was calculated for each utterance. Moreover, the MAEs for the average Japanese emotion dimensions from all German speakers were determined and the results are presented in Figure 6.4. From this figure, the increment of the MAE for estimating Japanese emotion dimensions from German database is as follows: the MAE for valence increased from 0.28 using the mono-



(a) Valence.



(b) Activation.



(c) Dominance.

Figure 6.3: Mean absolute error (MAE) for estimating Japanese emotion dimensions (valence, activation, and dominance) using (1) a mono-language emotion recognition system trained using Japanese database and (2) a cross-language emotion recognition system trained using 10 German speakers individually.

language emotion recognition system to 0.41 using the cross-language emotion recognition system, the MAE for activation increased from 0.19 to 0.32, and for dominance increased from 0.17 to 0.42. In all cases, the mean absolute error of emotion dimensions increased, however these increments do not constitute a large difference. Therefore, we can conclude that the cross-language emotion recognition system that trained using German database has the ability to estimate Japanese emotion dimensions as good as using the Japanese language for training the system.

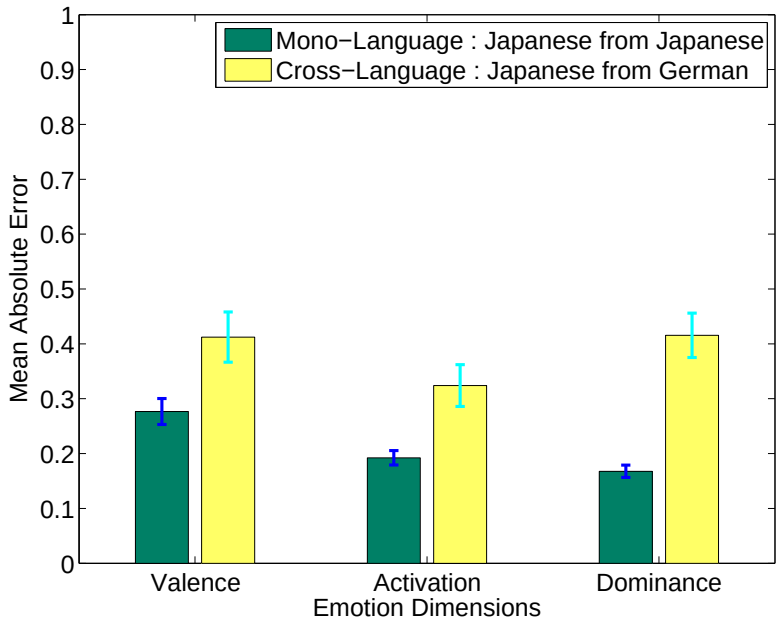
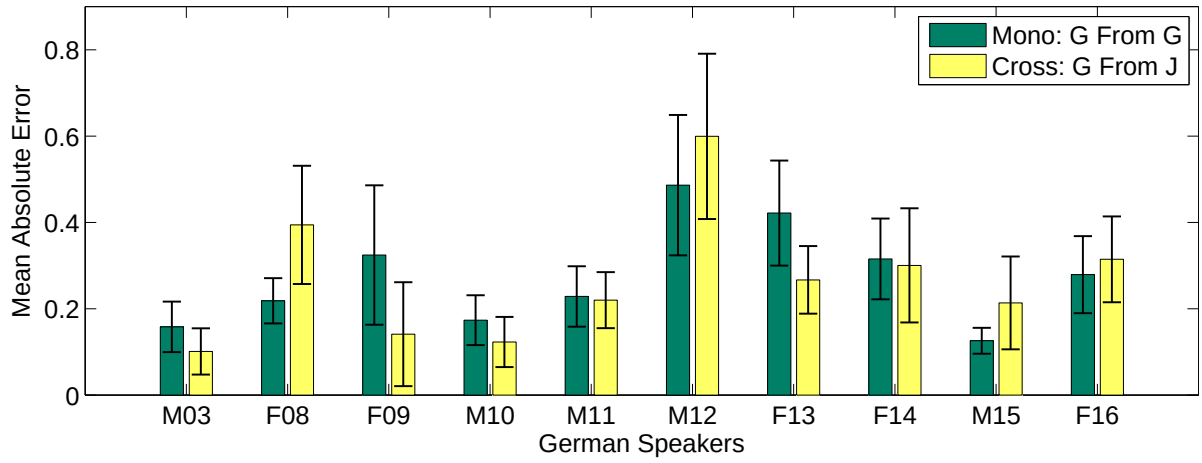


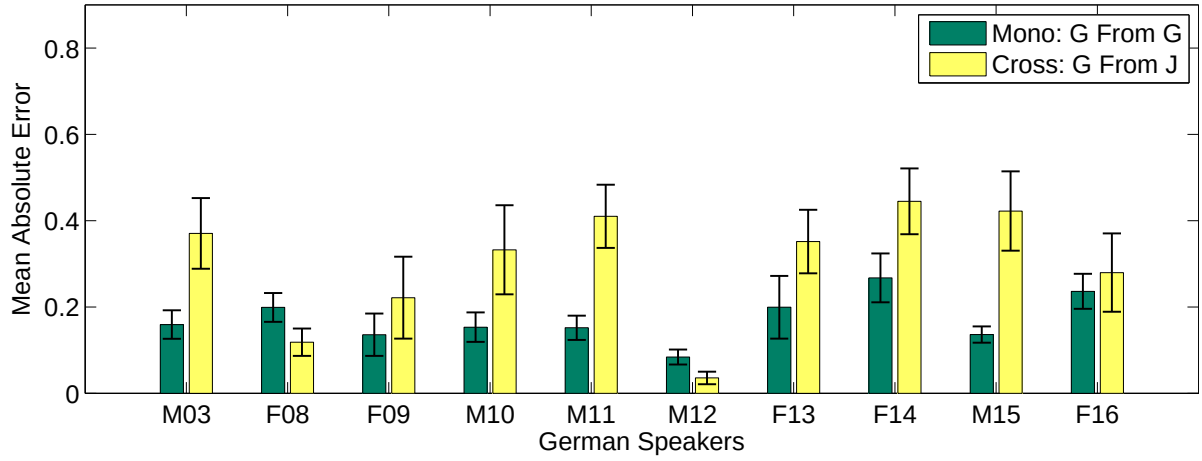
Figure 6.4: Mean absolute error (MAE) for (1) the estimated values of emotion dimensions using mono-language emotion recognition system trained using Japanese database and (2) the average of estimated values of emotion dimensions using cross-language emotion recognition system.

6.3.2 German emotion dimensions estimation from Japanese database

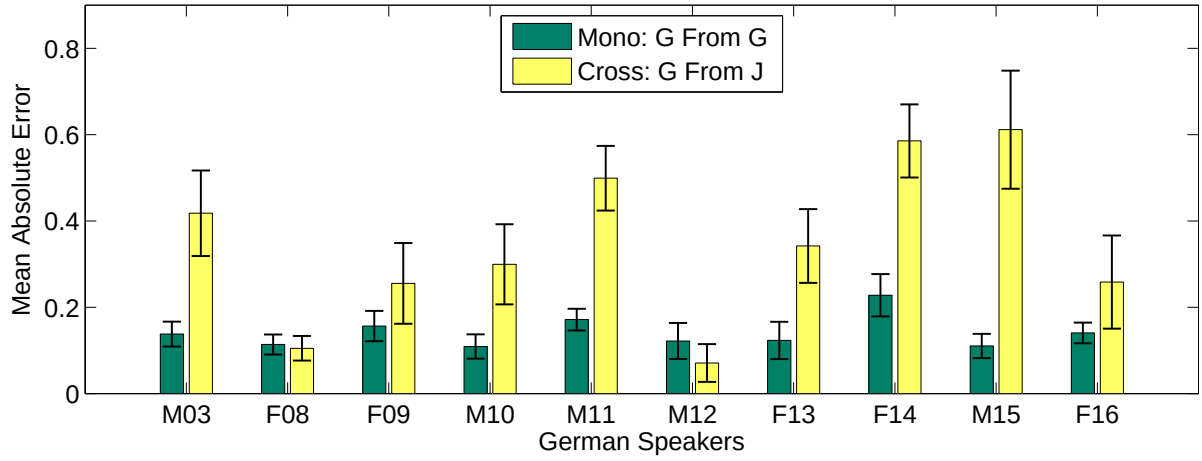
On the other hand, in order to estimate German emotion dimensions from Japanese database, a cross-language emotion recognition system was constructed and trained using Japanese database. This system was tested using the utterances from German database, for each German speaker individually.



(a) Valence.



(b) Activation.



(c) Dominance.

Figure 6.5: Mean absolute error (MAE) for estimating German emotion dimensions (valence, activation, and dominance) for 10 German speakers individually using: (1) a mono-language emotion recognition system trained using each German speaker dataset individually and (2) a cross-language emotion recognition system trained using Japanese database.

Figures 6.5(a), 6.5(b), and 6.5(c) show the MAEs for estimating German emotion dimensions (valence, activation, and dominance) for 10 German speakers, using mono-language emotion recognition system and cross-language emotion recognition system. The results for each speaker individually revel that the estimation using the cross-language system were close to the estimation using mono-language system for all speakers.

The combination of all speaker utterances constitute the whole German database. The MAE for emotion dimensions estimation of the whole database using the cross-language emotion recognition system was calculated and compared with the estimation of emotion dimensions using the mono-language emotion recognition system, as shown in Figure 6.6.

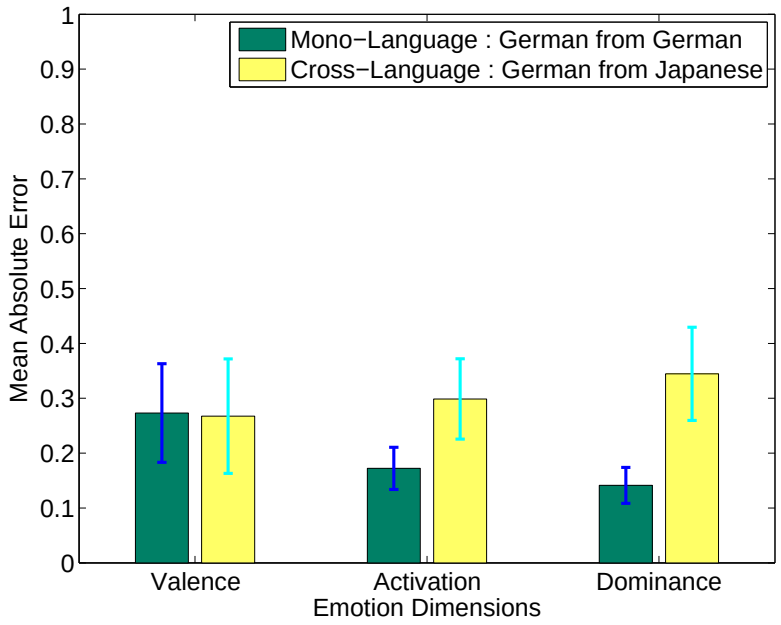


Figure 6.6: Mean absolute error (MAE) for estimating emotion dimensions using: (1) a mono-language emotion recognition system trained using each all German speakers and (2) a cross-language emotion recognition system trained using Japanese database.

From this figure, the difference between MAE for estimating German emotion dimensions using the mono-language emotion recognition system and the cross-language emotion recognition system is as follows: The MAE for valence is unchanged for both mono-language and cross-language system, the MAE for activation increases from 0.17 using the mono-language emotion recognition system to 0.30 using the cross-language emotion recognition system, and the MAE for dominance increases from 0.14 to 0.34. In

cases of the activation and dominance, the mean absolute error of the emotion dimension increases, however these increments do not constitute a large difference.

These results revel that our cross-language emotion recognition system trained using Japanese database has the ability to estimate German emotion dimensions with a little bit higher error than the estimation using the mono-language emotion recognition system. Therefore, we can conclude that the emotion dimension for the German database can be detected from a speech emotion recognition system trained with the Japanese database with a small error.

6.4 Summary

The aim of this chapter is to investigate whether emotion dimensions (valence, activation, and dominance) can be estimated cross-lingually or not? In order to accomplish this task, we work with two language databases Japanese and German language. First, we investigate whether their are common acoustic features between the two languages. Second, we construct a cross-language emotion recognition system based on human perception three-layer model to accurately estimate emotion dimensions.

For both languages, our proposed feature selection method was used to select the most relevant acoustic features for each emotion dimension. Then, we investigate whether the two databases share some common acoustic features and semantic primitives which allow us to estimate emotion dimensions cross-lingually. For each emotion dimension, it was found many acoustic features and semantic primitives were shared by both database. These common acoustic features and semantic primitives allow us to construct our proposed cross-language emotion recognition system based on the three-layer model. The input of this system are the common acoustic features and the outputs are the estimated emotion dimensions: valence, activation, and dominance.

For estimating emotion dimensions, the proposed cross-language emotion recognition system was trained using one language and testing using the second language. For instance, Japanese emotion dimensions were estimated form German database by train-

ing the system using acoustic features, semantic primitives, and emotion dimensions for each German speaker dataset individually, then the trained system was used to estimate Japanese emotion dimensions using Japanese acoustic features as inputs, in a similar way the German emotion dimensions were estimated from Japanese database.

These results revel that our cross-language emotion recognition system trained using one language database has the ability to estimate emotion dimensions for the other language database as good as the estimation using the mono-language emotion recognition system. Therefor, our assumption that emotion dimensions can be estimated cross-lingually is confirmed i.e. values of emotion dimensions for Japanese language can be estimated from a cross-language emotion recognition system trained with German database, and vice-versa.

Chapter 7

Mapping the estimated emotion dimensions into emotion categories

7.1 Introduction

Emotion dimensions and emotion categories are closely related, i.e. by detecting the emotional content using one of these two approaches, we can infer its equivalents in the other scheme. Using the dimensional approach, emotion categories are represented by regions in an n-dimensional space, for example, in the two-dimensional space valence-activation, happy is represented by a region which lies in the first quarter, in which valence is positive, and activation/arousal is high as shown in Figure 7.1.

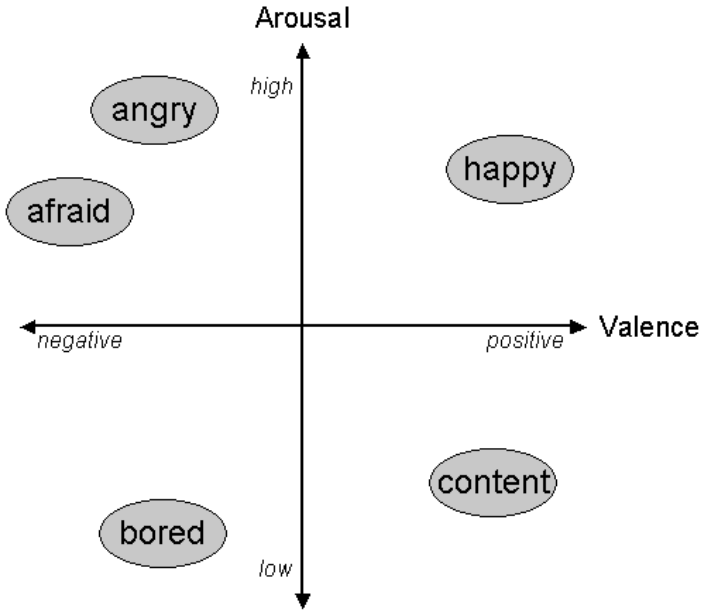
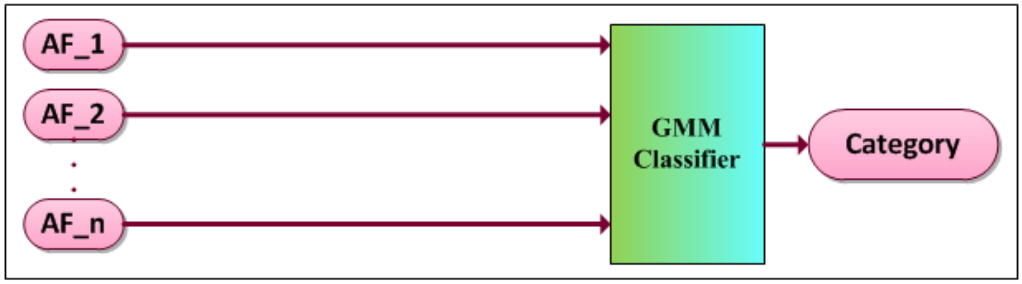


Figure 7.1: Basic emotions are marked as areas within the Valence-Arousal space.

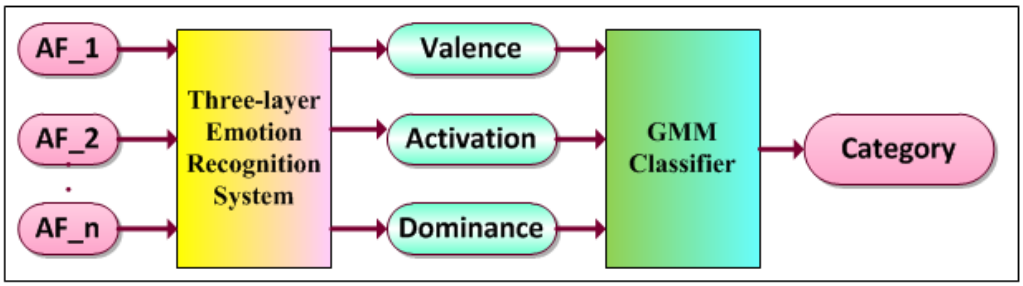
For instance, if an utterance is estimated with positive valence and high activation it could be inferred as Happy, and vice versa. Thus, we can easily map emotion categories into the dimensional space and vice, versa. Therefore, any improvement in dimensional approach will leading to an improvement in the categorical approach. In this chapter, we investigate whether using the estimated emotion dimensions as inputs to the emotion recognition classifier will improve the categorical classification or not.

Gaussian Mixture Model (GMM) traditional used for classifying emotion category by mapping acoustic features to emotion category as show in Figure 7.2(a). However, in this

chapter the estimated values of emotion dimensions (valence, activation, and dominance) were used as inputs for GMM to predict the corresponding emotional category as show in Figure 7.2(b).



(a) Using acoustic features for emotion classification



(b) Using estimated emotion dimensions for emotion classification

Figure 7.2: Emotion classification using Gaussian Mixture Model (GMM) as classifier and the input are as follows: (a) acoustic features (b) estimated emotion dimensions.

To measure the improvement for the proposed method, we compared the classifications into emotion categories using (1) acoustic features directly, with the classification using (2) the estimated values of emotion dimensions.

7.2 Classification into emotion categories

Gaussian Mixture Model GMM classifier was widely used for emotion classification from acoustic features into emotion categories. In this study GMM was used to detect the emotion category but not from the acoustic features instead the estimated emotion dimensions were used as the input for the emotion recognition system based on GMM. Figure 7.3 shows the used procedure for classifying emotion categories by mapping acoustic features and estimated emotion dimensions into emotion categories.

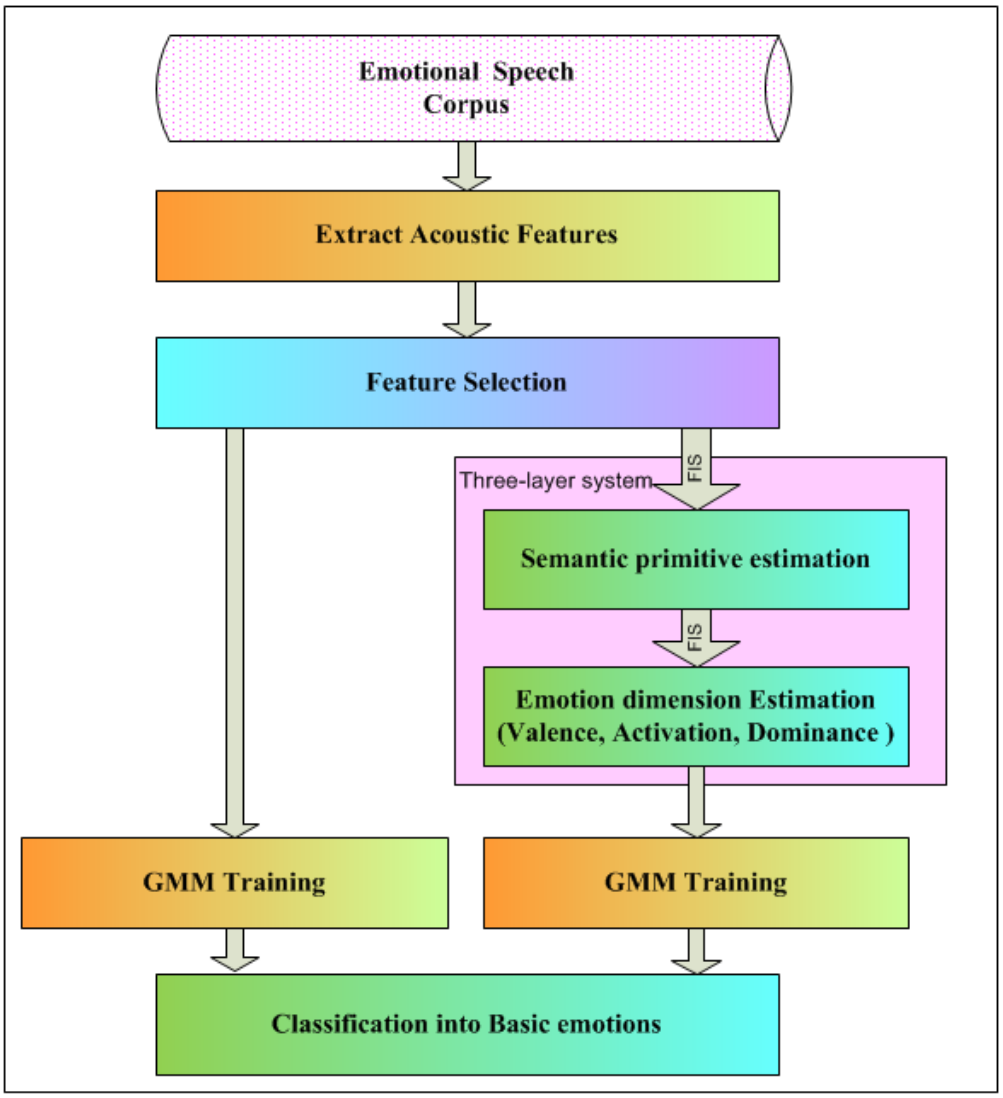


Figure 7.3: Emotion classification using acoustic features directly and estimated emotion dimensions.

The most relevant acoustic features which were used for estimating emotion dimensions as described in Chapter 5, were used to predict the emotional categories using GMM for the two databases. However, in order to predict the emotional categories using the proposed method, the most relevant acoustic features for each emotion dimensions were used to estimate semantic primitives using FIS, then the estimated semantic primitives were used to estimate emotion diminutions using FIS, finally, GMM is used to map the values of emotion dimensions to the corresponding emotion category.

Table 7.1: Classification results for Japanese database.

(a) Mapping acoustic features directly to emotion categories using GMM classifier (Ave. 53.9%)

	Neutral	Joy	Cold Anger	Sad	Hot Anger
Neutral	**30.0%**	15.0%	45.0%	5.0%	5.0%
Joy	2.5%	**40.0%**	12.5%	2.5%	42.5%
Cold Anger	7.7%	12.8%	**71.8%**	5.1%	2.6%
Sad	0.0%	7.5%	12.5%	**77.5%**	2.5%
Hot Anger	2.5%	45.0%	2.5%	0.0%	**50.0%**

(b) Mapping the estimated emotion dimensions for speaker-dependent task to emotion categories using GMM classifier (Ave. 94.0%)

	Neutral	Joy	Cold Anger	Sad	Hot Anger
Neutral	**80.0%**	10.0%	5.0%	5.0%	0.0%
Joy	0.0%	**97.5%**	2.5%	0.0%	0.0%
Cold Anger	0.0%	0.0%	**100.0%**	0.0%	0.0%
Sad	0.0%	0.0%	0.0%	**100.0%**	0.0%
Hot Anger	0.0%	2.5%	5.0%	0.0%	**92.5%**

(c) Evaluation using listening test experiment by human subjects, from the work of Huang et al. [33] (Ave. 92.0%).

	Neutral	Joy	Cold Anger	Sad	Hot Anger
Neutral	**98.0%**	0.0%	2.0%	0.0%	0.0%
Joy	12.0%	**87.0%**	1.0%	0.0%	0.0%
Cold Anger	10.0%	0.0%	**86.0%**	4.0%	0.0%
Sad	5.0%	0.0%	3.0%	**92.0%**	0.0%
Hot Anger	1.0%	0.0%	2.0%	0.0%	**97.0%**

7.2.1 Classification for Japanese Database

For the Japanese database, first, the acoustic features were used as input to train the GMM classifier to classify the Japanese database into five categories: Neutral, Joy, Hot Anger, Sadness, and Cold Anger. Moreover, the estimated values of emotion dimensions were used as input to train GMM to classify the values of every point in the space Valence-Activation-Dominance into one emotion category.

The confusion matrix of the results is shown in Table 7.1(a) for mapping acoustic features into categories and in Table 7.1(b) for mapping values of emotion dimensions into emotion categories. In these tables, the numbers are the percentages of recognized utterances of the category in the top line versus the number of utterances for emotions in the left column.

Sad and cold anger achieved the best recognition results both of them achieved 100%. The classification error was highest for neutral, joy, and hot anger using the categorical approach. In contrast, joy and neutral achieved the highest improvement, neutral increased from 30% using the categorical approach to 80% using the dimensional approach, moreover, joy rate increased from 40% to 97.5%.

Tables 7.1(b) and 7.1(c) show the classification rate for the proposed emotion recognition system and the human evaluation using a listening test. From the listening test [33], as shown in Table 7.1(c): neutral and hot anger achieved the best recognition rate they achieve 98% and 97%, respectively, however, these two categories achieved 80%, 92.5% using the proposed system. The overall classification rate using the proposed system is close to the human perception, the results were 94% and 92% , respectively.

It can be summarized that the emotion dimensions estimation lends itself well to emotion categorization as shown in Figure 7.4.

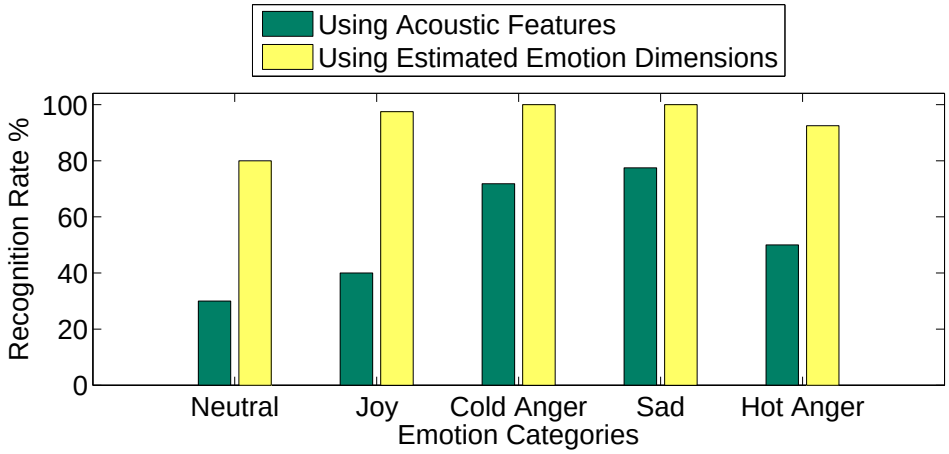


Figure 7.4: Recognition rate for emotion categories (Neutral, Joy, Cold Anger, Sadness, Hot Anger) for Japanese database using GMM classifier by mapping (1) acoustic features and (2) the estimated emotion dimensions from the speaker-dependent task.

From this figure we can easily see that using the dimensional approach the classification rate increased for all emotion categories. The average classification rate increased from 53.9% to 94% for categorical approach and dimensional approach, respectively.

Table 7.2: Classification results for German database.

(a) Mapping acoustic features directly to emotion categories using GMM classifier (Ave. 60.0%)

	Neutral	Happy	Anger	Sad
Neutral	**66.0%**	16.0%	0.0%	18.0%
Happy	12.0%	**54.0%**	32.0%	2.0%
Anger	2.0%	42.0%	**54.0%**	2.0%
Sad	16.0%	6.0%	12.0%	**66.0%**

(b) Mapping the estimated emotion dimensions for multi-speaker task to to emotion categories using GMM classifier (Ave. 75.0%)

	Neutral	Happy	Anger	Sad
Neutral	**74.0%**	10.0%	4.0%	12.0%
Happy	6.0%	**62.0%**	32.0%	0.0%
Anger	2.0%	18.0%	**80.0%**	0.0%
Sad	16.0%	0.0%	0.0%	**84.0%**

(c) Mapping the estimated emotion dimensions for speaker-dependent task to emotion categories using GMM classifier (Ave. 95.5%)

	Neutral	Happy	Anger	Sad
Neutral	**98.0%**	0.0%	2.0%	0.0%
Happy	0.0%	**94.0%**	6.0%	0.0%
Anger	0.0%	8.0%	**92.0%**	0.0%
Sad	2.0%	0.0%	0.0%	**98.0%**

(d) Using speaker-dependent emotion recognition using probabilistic neural network, from the work of Iliou et al. [36] (Ave. 94.8%).

	Neutral	Happy	Anger	Sad	Others
Neutral	**92.4.0%**	0.0%	0.0%	1.3 %	6.3 %
Happy	0.0%	**93.0%**	0.0%	0.0%	7.0%
Anger	0.8%	0.8%	**95.2%**	0.0%	3.2%
Sad	1.60%	0.0%	0.0%	**98.4%**	0.0%

7.2.2 Classification for German Database

In order to investigate the speaker-dependency impact on emotion classification, German database was used because it is a multi-speaker database. Therefore, we applied the classification from estimated emotion dimensions to both the individual speakers of the German database (speaker-dependent task) and the combined set of all sentences across all speakers in the database (multi-speaker task).

The results of classification of the German database into four categories (Neutral, Happy, Angry, and Sad) are as follows: the confusion matrix of the results is shown in

Table 7.2(a) for mapping acoustic feature into categories, Table 7.2(b) for mapping emotion dimensions into categories for multi-speaker estimation, and Table 7.2(c) for mapping emotion dimensions into categories for speaker-dependent estimation.

From these tables, we can conclude that the recognition rate increased of all emotion categories when using the estimated emotion dimensions instead of acoustic features directly. Happy and anger achieved the highest improvement, happy increased from 54% using the categorical approach to 94% using emotion dimensions form speaker-dependent task, moreover, anger rate increased from 54% to 92%. The reason for these improvement for happy and anger classification rate was due to the high performance for the valence dimension using the proposed approach.

Tables 7.2(c) and 7.2(d) show the classification rate for the proposed system and the probabilistic neural network [36], respectively, for speaker-dependent task. As can seen from these tables that, the classification rate for the two systems were very close for all emotion categories. Moreover, the overall classification rate using the proposed system is very close to those of the probabilistic neural network, the results were 95.5% and 94.8%, respectively.

Figure 7.5 shows the improvement for each emotion categories for German database. It is clearly that using the dimensional approach the classification rate increased for all emotion categories. The average classification rate are as follows: 60%, 75%, and 95.5% form acoustic features, estimated emotion dimensions using the multi-speaker task, and speaker-dependent task, respectively. The best recognition rate was achieved for the speaker-dependent task. This result is consistent with most of previous studies indicating that speaker-dependent training of the estimator achieves the most accurate emotion classification results.

7.3 Discussion

In this chapter we investigate whether the estimated emotion dimensions valence, activation, and dominance can be used to improve emotion classification for categorical ap-

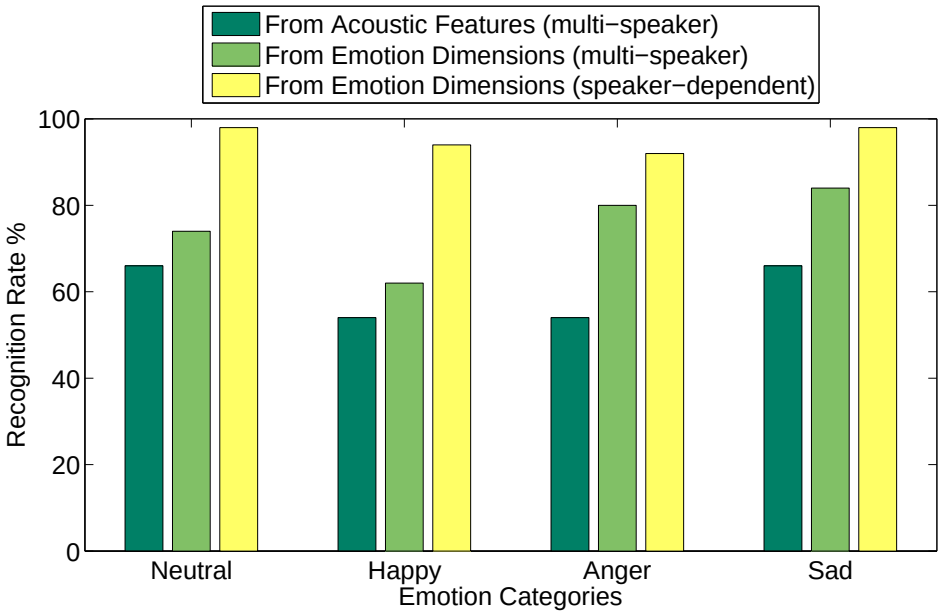


Figure 7.5: Recognition rate for emotion categories (Neutral, Happy, Angry, Sad) for German database using GMM classifier by mapping (1) acoustic features, (2) the estimated emotion dimensions from the multi-speaker task, and (3) the estimated emotion dimensions from the speaker-dependent task.

proach. The performance of the proposed emotion recognition system greatly improved when using the dimensional approach for both language databases. These improvement in the dimensional representation can be reflected to the categorical representation by mapping the estimated emotion dimensions into emotion categories.

To accomplish this task, the estimated emotion dimensions were mapped using Gaussian Mixture Model (GMM) into emotion categories as follows: for Japanese database GMM was used to map the emotion dimensions to emotion categories (Neutral, Joy, Cold-Anger, Sadness, Hot Anger), and in German database, GMM was used to map emotion dimensions to the emotion categories (Neutral, Joy, Anger, Sadness).

The results for speaker-dependent task were as follows: in case of Japanese database; sad and cold anger achieved the best recognition results, they achieved 100% as shown in Table 7.1(b). In addition, for German database, neutral and sad achieved the best recognition rate, both of them achieved 98% as shown in Table 7.2(c). the overall recognition rate for speaker-dependent task were, 94% and 95.5% for Japanese and German database, respectively.

However, in case of multi-speaker task for German database, sad category achieved the highest recognition rate 84% as presented in Table 7.2(b). The over all classification rate for multi-speaker task for German database was 75%. These, reveal that, sad is the most easily recognized emotional category among all categories for both speaker-dependent and multi-speaker task, for both language. The most important result is that the highest performance achieved for speaker-dependent task, in both languages.

7.4 Summary

The aim of this chapter is to investigate whether the estimated emotion dimensions can to improve the categorical classification or not. Therefor, the emotion dimensions values are mapped into the given emotion categories using a GMM classifier, for Japanese and German database. To measure the improvement of using the dimensional approach, the classification rate of emotion categories from acoustic features directly was compared with that of using the estimated values of emotion dimensions.

For the Japanese database, the overall recognition rate was 53.9% using direct classification using acoustic features and up to 94% using emotion dimensions. For the German database, the rate of classification directly from acoustic features was 60%, which was increased by up to 75% and 95.5% using emotion dimensions for multi-speaker and speaker-dependent tasks, respectively. The result reveals that the recognition rate in speaker-dependent tasks is higher than in multi-speaker tasks. This corresponds with previous studies indicating that speaker-dependent training of the estimator achieves the most accurate emotion classification results. The most important results is that, the classification using emotion dimensions instead of acoustic features improves the recognition rate for both database.

Chapter 8

Summary and Future Work

This work is motivated by the long-term goal to construct an automatic speech emotion recognition system that has the ability to accurately estimate emotion dimensions valence, activation and dominance from a speech signal. Our focus, in the dimensional approach is to improve the estimation results of the valence dimension. It was found in most of the previous studies that the acoustic features related to the valence dimension are very few, very weak and inconsistent. Due to these limitations, it was very difficult to predict this dimension. This study investigate the answer of the following important questions for constructing emotion recognition system:

- **the first question is:** what are the acoustic features relevant to emotion dimensions valence, activation and dominance?
- **the second question is:** how to develop the model or the relationship between acoustic features and emotion dimensions to improve the estimation results for emotion dimensions?
- **the third question is:** whether there are common acoustic features between languages? which allow us to build an automatic emotion recognition system to estimate emotion dimensions for one language by training the system using another language.

Acoustic features are very important for building an automatic speech emotion recognition system. They are used as an input for automatic emotion recognition system. As far as the input more discriminative the best output results will be obtained for the system. Most of acoustic feature selection technique were based on the correlation between acoustic features and emotion dimensions as a two-layer model. Using the conventional two-layer model the activation, and dominance could be predicted with high accuracy, while valence was poorly estimated in most of them. This problem not only for the dimensional approach but also for categorical approach. For instance, some emotion categories such as happy and angry share the same acoustic features which make it difficult for the learning algorithm to discriminate between these emotions. This is the reasons why acoustic discriminate ability for valence still problematic: there are no strong discriminate

acoustic features available to discriminate between positive (e.g., happiness) and negative (e.g., anger), however, these emotions are usually not hard to distinguish for humans. All these studies suggest that finding relevant features to discriminate in the valence domain is one of the main challenges in speech emotion recognition. For all of these reasons, our motivation in this study is to investigate the most related acoustic feature for each emotion dimensions, especially the most challenging dimension valence.

To improve the estimation results for emotion dimensions valence, activation, and dominance from a speech emotion recognition system, this study propose the following assumptions:

- **the first assumption is:** human perception is a three-layer model not two-layer model, therefor, constructing a speech emotion recognition system based on a three-layer model which imitate human perception will help us to find the most related acoustic feature to each emotion dimension, moreover, using these acoustic features will improve the estimation accuracy for emotion dimensions valence, activation, and dominance.

- **the second one is:** human has the ability to detect the emotional state of a speaker even without understanding the language of the speaker, therefore, automatic emotion recognition system could detect the emotional state regardless of the language.

The conventional two-layer model has limited ability to find the most relevant acoustic features for each emotion dimension, especially valence, or to improve the prediction of emotion dimensions from acoustic features. However, this model does not imitate human perception, this is reason behind the poor estimation of valence dimension. Human perception is a multi-layer process as described by Scherer [79]. Huang and Akagi 2008, assume that human perceive emotional speech not directly from a change of acoustic features, but rather from a composite of different types of smaller perceptions that are expressed by semantic primitives or adjectives.

In this thesis, the proposed idea to improve the prediction of emotion dimensions

can be done by imitating the process of human perception for recognizing the emotional state from a speech signal. Therefore, to overcome the limitations of the two-layer model, this study proposes a three-layer model for human perception to improve the estimating values of emotion dimensions from acoustic features. Our proposed model consists of three layers: emotion dimensions (valence, activation, and dominance) constitute the top layer, semantic primitives the middle layer, and acoustic features the bottom layer. A semantic primitive layer is added between the two conventional layers acoustic features and emotion dimensions.

The following are the details of constructing the proposed emotion recognition system in this study:

8.1 The elements of the proposed system

In Chapter 3 the elements of the proposed three-layer emotion recognition system were collected, ranging from the used databases, over acoustic feature extraction, to the experimental evaluation for emotion dimensions and semantic primitives using two listening tests by human subjects. Two databases were selected to validate the proposed system one Japanese and the other German database.

The input of our automatic emotion recognition system are the acoustic features, therefore, 21 acoustic features were extracted for each utterance in the two databases as initial set of acoustic features. Semantic primitives are adjectives describing emotional voice, this is the new layer we added between the two traditional layers: acoustic features and emotion dimensions. 17 semantic primitives are used to represent the new layer as follow: (Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow). Three emotion dimensions (valence, activation, and dominance) are constitute the top layer which are the final outputs for the proposed system.

In order to build the perceptual model for each dimension, two listening experiments were conducted to evaluate all elements of the semantic primitive layer and emotion

dimensions layer, for the two databases. Finally, inter-rater agreement was preformed in order to obtain a reliable data by excluding the subjects who have very low correlation coefficient among all subjects.

8.2 Selecting the most relevant features for each emotion dimension

The first half of Chapter 4 attempt to answer the first question of this study; what are the most relevant acoustic features for each emotion dimension? Based on the first assumption of this study, we assume that the acoustic features that are highly correlated with semantic primitives will have a large impact for predicting values of emotion dimensions, especially for valence. Therefore, a top-down feature selection method was proposed to select the most related acoustic features based on the three-layer model. By firstly, selecting the highly correlated semantic primitives for emotion dimension, then selecting the set of all acoustic features which are highly correlated with the selected semantic. The set of selected acoustic features are considered the most related to the emotion dimension in the top layer.

Having identified the most relevant acoustic features and semantic primitives for each emotion dimension, a perceptual three-layer model was constructed for each emotion dimension. The perceptual three-layer model for each emotion dimensions consists of: the desired emotion dimension in the top layer, the most relevant semantic primitives in the middle layer, the most relevant acoustic features in the bottom layer.

The most important result is that, using the proposed three-layer model for feature selection, the number of relevant acoustic features to emotion dimensions increases. For example, the number of relevant features for the most difficult dimension valence increases from one using the conventional method to nine using the proposed method. Moreover, the number of features increased from eight to nine for activation and from eight to ten for dominance.

The proposed three-layer model based on human perception assumption allow us to find a set of acoustic features for each emotion dimension, especially for valence which was the most difficult dimension. Therefore, the three-layer model outperform the traditional two-layer model for selecting acoustic feature.

8.3 System Implementation

The second half of Chapter 4, tried to improve the exciting emotion recognition system in order to accurately estimate emotion dimensions from acoustic features. The constructed perceptual three-layer model for each emotion dimension was used to improve emotion dimensions estimation using a bottom-up method. The bottom-up method imitating human perception process for estimate emotion dimensions. This method was used to construct our emotion recognition system as follows: the input of the proposed system are the acoustic features in the bottom layer, the output of are the emotion dimensions valence, activation, and dominance. Fuzzy inference system FIS was used to connect the elements of the proposed system. Firstly one FIS was used to estimate each semantic primitive in the middle layer form the acoustic features in the bottom layer. Then one FIS was used to estimate each emotion dimensions from the estimated semantic primitives.

8.4 System Evaluation

Chapter 5 investigate whether the first assumption is satisfied or not. Therefor, we try to answer the following two questions: whether the selected acoustic features are effective for predicting emotion dimensions? second, whether the proposed emotion recognition system improve the estimation accuracy of emotion dimensions (valence, activation, and dominance) or not?

In order to assess the performance of the proposed system, mean absolute error (MAE) is used to measure the distance between the estimated dimensions by the proposed system and the evaluated emotion dimensions by human listeners. The smaller MAE the closer

estimated value to the human evaluation.

To investigate the first question, the most relevant acoustic features for each emotion dimension, which was selected using the feature selection method, were used as inputs of the proposed emotion recognition system, to estimate values of emotion dimensions. Then, the estimation results of emotion dimensions are compared with those of estimation using the non-relevant acoustic features and all acoustic features. For both databases, the results reveal that, the MAEs by using the selected acoustic features group as an inputs to the proposed emotion recognition system were the smallest compared with the other two of groups features, these results indicate that, the selected acoustic features improve the prediction of all emotion dimensions.

Furthermore, to investigate the second question which mean is how effectively our proposed system improve emotion dimensions estimation. Therefore, the performance of the proposed system was compared with that of the conventional two-layer system, using two different languages Japanese and German, with two different tasks (speaker-dependent task and multi-speaker task).

To accomplish these tasks, two emotion recognition system were constructed the first system was constructed based on the proposed approach and the other based on the conventional approach. The selected acoustic features group was used as input for both the proposed system and the conventional system. The proposed system was constructed based on the three-layer model of human perception as follows: one FIS was used to estimate each semantic primitive from the selected acoustic features, then one FIS was used to estimate each emotion dimension from the estimated semantic primitives. For constructing the conventional system which based on the two-layer model, one FIS was used to estimate each emotion dimension from the selected acoustic features directly.

For both Japanese and German database, The MAEs for all dimensions were very small which indicate that the proposed three-layer system is effective and gives the best results for all emotion dimensions (valence, activation, and dominance) for both speaker-dependent and multi-speaker task. However, the MAEs for the multi-speakers task were higher than those for the speaker-dependent task.

For German and Japanese databases, the overall best result is achieved for all emotion dimensions using speaker-dependent task. These results suggest that the valence dimension estimation is speaker dependent, while activation and dominance is may be speaker independent. The multi-speaker variation have a great effect for valence dimension estimation results.

Therefore, from this study it was evident that the valence dimension estimation is improved by using the proposed model. Therefore, the most important results is that the proposed automatic speech emotion recognition system based on the three-layer model for human perception was superior to the conventional two-layer system.

8.5 Cross-language emotion recognition System

Most of the previous studies for automatic speech emotion recognition were based on detecting the emotional state working on mono-language, i.e. training and testing the automatic emotion recognition system using only one language database. However, in order to develop a generalized emotion recognition system, the performance of these systems must be analyzed in mono-language as well as cross-language. The goal of Chapter 6, is to construct a cross-lingual emotion recognition system that has the ability to estimate emotion dimensions for one language by training the system using another language. Therefore, the question we try to answer the third question in this study, whether there are common acoustic features between two languages? which allow us to build an automatic emotion recognition system to estimate emotion dimensions for one language by training the system using another language. Therefore, we investigate whether our proposed automatic emotion recognition system is able to estimate emotion dimensions valence, activation, dominance cross-lingually?

To accomplish this task, first, we investigate whether their are common acoustic features between the two languages. Second, we construct a cross-language emotion recognition system based on human perception three-layer model to accurately estimate emotion dimensions.

For both languages, our proposed feature selection method was used to select the most relevant acoustic features for each emotion dimension. For each emotion dimension, it was found that many acoustic features and semantic primitives were shared by both database. These common acoustic features and semantic primitives allow us to construct our proposed cross-language emotion recognition system based on the three-layer model. The input of this system are the common acoustic features and the outputs are the estimated emotion dimensions: valence, activation, and dominance.

For estimating emotion dimensions, the proposed cross-language emotion recognition system was trained using one language and testing using the second language. For instance, Japanese emotion dimensions were estimated form German database by training the system using acoustic features, semantic primitives, and emotion dimensions for each German speaker dataset individually, then the trained system was used to estimate Japanese emotion dimensions using Japanese acoustic features as inputs, in a similar way the German emotion dimensions were estimated from Japanese database.

These results revel that our cross-language emotion recognition system trained using one language database has the ability to estimate emotion dimensions for the other language database as good as the estimation using the mono-language emotion recognition system. Therefor, our assumption that emotion dimensions can be estimated cross-lingually is confirmed i.e. values of emotion dimensions for Japanese language can be estimated from a cross-language emotion recognition system trained with German database, and vice-versa.

8.6 Mapping estimated emotion dimensions into emotion categories

Emotion dimensions and emotion categories are closely related, i.e. by detecting the emotional content using one of these two approaches, we can infer its equivalents in the other scheme. For instance, if an utterance is estimated with positive valence and high

activation it could be inferred as Happy, and vice versa. Thus, we can easily map emotion categories into the dimensional space and vice, versa. Therefore, any improvement in dimensional approach will leading to an improvement in the categorical approach.

Chapter 7 investigates whether the estimated emotion dimensions can be used as inputs to the emotion recognition classifier to improve the categorical classification or not. Therefor, the emotion dimensions values are mapped into the given emotion categories using a GMM classifier. For Japanese and German database used in this study, the results of classifying into emotion categories using acoustic features directly and the estimated values of emotion dimensions was compared to measure the improvement of using the dimensional approach.

For the Japanese database, the overall recognition rate was 53.9% using direct classification using acoustic features and up to 94% using emotion dimensions. For the German database, the rate of classification directly from acoustic features was 60%, which was increased by up to 75% and 95.5% using emotion dimensions for multi-speaker and speaker-dependent tasks, respectively. The result reveals that the recognition rate in speaker-dependent tasks is higher than in multi-speaker tasks. The most important results is that, the classification using emotion dimensions instead of acoustic features improves the recognition rate for both database.

8.7 Contributions

Compared with the conventional approach for feature selection, the proposed approach take into account human perception which helps us,

- to find the many acoustic features related to the valence dimension, which the most challenging dimension in all previous study, as well as to find new acoustic features for activation and dominance.

- to improve the estimation results of emotion dimensions especially valence dimension, using the acoustic features determined by this approach. Moreover, to improve

the activation and dominance dimensions.

- to investigate the common acoustic features related to emotion dimensions among different language (German/Japanese).

- to construct a cross-language emotion recognition system to estimate emotion dimensions cross-lingually, using the common acoustic features between the two languages selected by the proposed approach.

8.8 Future Work

In the future my focus is on emotional speech modification. The question I try to answer how neutral speech should be modified in order to perceived as emotional speech. In other words, we investigate the speech acoustic features that are effective for perception of emotions, and propose an emotion modification model to transform neutral speech into emotional speech. The modification can be achieved by modifying acoustic features according the relationship between the acoustic feature and emotion dimensions. Using the dimensional approach for modification will make the emotional state of the transformed speech more natural. The proposed acoustic feature selection method can be used to find the most related acoustic feature for each emotional state. The immediate application, for the proposed neutral to emotion transformation system can be used in the field of text-to-speech (TTS) synthesis.

Bibliography

[1] Akagi, M. "Analysis of Production and Perception Characteristics of Non-linguistic Information in Speech and Its Application to Inter-language Communications," Proceedings : APSIPA ASC 2009 : Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference: pp. 513-519 (2009)

[2] Albrecht, I. and Schroder, M. and Haber, J. and Seidel, H.-P. "Mixed feelings: Expression of non-basic emotions in a muscle-based talking head," *Virtual Reality*, **8(4)**, pp. 201-212, (2005).

[3] Jo-Anne Bachorowski and Michael J. Owren. "Sounds of Emotion," Annals of the New York Academy of Sciences, 1000(1), pp. 244-265, January (2006)

[4] Breazeal, C. "Emotion and sociable humanoid robots," International Journal of Human-Computer Studies, 59, 119-155, (2003)

[5] Burkhardt, F. Paeschke, A. Rolfes, M. Sendlmeier, W. and Weiss, B. "A Database of German Emotional Speech," Proc. of Interspeech, 2005.

[6] Busso, C. and Rahman T. "Unveiling the Acoustic Properties that Describe the Valence Dimension," INTERSPEECH 2012.

[7] Banse, R. and Scherer, K. R. "Acoustic profiles in vocal emotion expression," Journal of personality and social psychology, 70(3), pp. 614-36, March (1996)

[8] Batliner, A. Steidl, S. Schuller, B. Seppi, D. Vogt, T. Wagner, J. Devillers, L. Vidrascu, L. Aharonson, V. Kessous, L. and Amir, N. Whodunnit - searching for

the most important feature types signalling emotion-related user states in speech, Comput. Speech Lang., vol. 25, no. 1, pp. 4-28, 2010.

[9] Boersma, P., and Weenink, D., "Praat: doing phonetics by computer," (Version 5.0.07) Computer program, 2008. Retrieved February 1, 2008, from http:www.praat.org. (2008).

[10] Brunswik, E., Historical and thematic relations of psychology to other sciences. Vol. 83. 1956: Scientific Monthly. 151-161.

[11] Burkhardt, F. Paeschke, A. Rolfes, M. Sendlmeier, W. F. and Weiss, B. "A database of German emotional speech," In Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH), Lisbon, Portugal, September 2005, pp. 1517-1520, (2005).

[12] Cowie, R. "Describing the emotional states that are expressed in speech," *Speech and Emotion*, pp. 11-18, (2000).

[13] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., "Emotion recognition in humancomputer interaction," IEEE Signal Process. Mag. 18 (1), pp. 32-80, (2001).

[14] Devillers, L., and Vidrascu, L. "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," Interspeech (2006).

[15] Dang, J., Dang, J., Li, A. Erickson, D. Suemitsu, A. Akagi, M. Sakuraba, K. Minematsu, N. Hirose, K. "Comparison of Emotion Perception among Different Cultures," APSIPA,Sapporo, Japan, (2009)

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

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

[18] Ekman, P., Sorenson, E. R., and Friesen, W. V. . "Pan-cultural elements in facial displays of emotions," Science, 164, pp. 86-88, (1969)

[19] Ekman, P., and Friesen, W. V. "Constants across cultures in the face and emotion. Journal of Personality and Social Psychology," 17, pp. 124-129. (1971)

[20] Ekman, P., and Friesen, W. V. "Pictures of facial affect. Palo Alto, CA: Consulting Psychologists Press," (1976)

[21] Ekman, P. "Argument for basic emotions," Cognition and Emotion, 6, pp. 169-200. (1992)

[22] Erickson, D. "Expressive speech: Production, perception and application to speech synthesis," *Acoustical Science and Technology*, **26(4)**, pp. 317-325, July, (2005).

[23] Espinosa, H.P. and Garcia, C.A.R. and Pineda, L.V. "Features selection for primitives estimation on emotional speech," Acoustics Speech and Signal Processing (ICASSP), pp. 5138-5141 (2010).

[24] Espinosa, H.P. Garcia, C.A.R. Pineda, L.V. "Bilingual Acoustic Feature Selection for Emotion Estimation Using a 3D Continuous Model," *Proc. Automatic Face and Gesture Recognition (FG 2011)*, pp. 786-791, (2011),

[25] Espinosa, H.P. Reyes-Garca, C.A. Pineda, L.V. "Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model," *Biomedical Signal Processing and Control*, **7(1)**, pp. 79-87, January (2012).

[26] Fant, G. "The voice source in connected speech," Speech Commun. 22, 125-139, (1997).

[27] Goudbeek, M., Goldman, J. P., and Scherer, K. R. "Emotion dimensions and formant position," INTERSPEECH 2009: 1575-1578, (2009).

[28] Gehm, T. L. and Scherer, K. R. (1988). Factors determining the dimensions of subjective emotional space. In K. R. Scherer (Ed.), Facets of Emotion chapter 5. Lawrence Erlbaum Associates.

[29] Grimm, M. and Kroschel, K. "Rule-Based Emotion Classification Using Acoustic Features," *Proc. Int. Conf. on Telemedicine and Multimedia Communication*, (2005).

[30] Grimm, M. and Kroschel, K. and Mower, E. and Narayanan, S. "Combining categorical and primitives-based emotion recognition," *Proc. EUSIPCO 2006*, (2006).

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

[32] Grimm, M. and Kroschel, K. "Emotion Estimation in Speech Using a 3D Emotion Space Concept," *Emotion*, June, (2007).

[33] Huang, C. F. Erickson, D. and Akagi, M. (2008) "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners.," Acoustics2008, Paris, 2317-2322.

[34] Hanson, H. M. "Glottal characteristics of female speakers: Acoustic correlates," Journal of the Acoustical Society of America 101(1): 466-481, (1997)

[35] Huang, C. and Akagi, M. "A three-layered model for expressive speech perception," *Speech Communication*, **50(10)**, pp. 810-828, October, (2008).

[36] Iliou, T., and Anagnostopoulos, C. N. (2010). Classification on Speech Emotion Recognition-A Comparative Study. International Journal on Advances in Life Sciences, 2(1 and 2), 18-28.

[37] Jang, J-S. , "Self-learning fuzzy controllers based on temporal back propagation," , IEEE Trans Neural Networks, vol 3, no. 5, pp. 714-723. (1992)

[38] Jang, J.S.R. "ANFIS: Adaptive network-based fuzzy inference system. IEEE Transactions on Systems ," *Man and Cybernetics*, **23(3)**, pp. 665-685, (1993).

[39] Jang, J-S. "Input Selection for ANFIS Learning," Proceedings of the IEEE International Conference on Fuzzy Systems, New Orleans. 1996

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

[41] Johnson-Laird, P. N., and Oatley, K. "Basic emotions, rationality, and folk theory," Cognition and Emotion, 6, pp. 201-223, (1992)

[42] Karadogan, S.G. and Larsen, J. "Combining semantic and acoustic features for valence and arousal recognition in speech," Proc. of Cognitive Information Processing, 2012.

[43] Kanluan, I. Grimm, M. and Kroschel, K. "Audio-Visual Emotion Recognition Using An Emotion Space Concept," *Proc. EUSIPCO 2008*, (2008).

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

[45] Kecman, V., "Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models," MIT Press, (2001)

[46] Klaus R. Scherer, Rainer Banse, Harald G. Wallbott, and Thomas Goldbeck. Vocal cues in emotion encoding and decoding. Motivation and Emotion," 15(2), pp. 123-148, June (1991)

[47] Lee, C.M., and Narayanan, S.S. "Emotion recognition using a data-driven fuzzy inference system," Proceedings of InterSpeech (pp. 157-160). Geneva, Switzerland, (2003).

[48] Lee, C.M. and Narayanan, S. "Toward Detecting Emotions in Spoken Dialogs," *IEEE Transactions on Speech and Audio Processing*, **13(2)**, pp. 293-303, (2005).

[49] Mori, H. Satake, T. Nakamura, M. and Kasuya, H. "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," Speech Communication, vol. 53 pp.36-50, 2011.

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

[51] Mamdani, E.H. and Assilian, S. 'An experiment in linguistic synthesis with a fuzzy logic controller', International Journal of Man-Machine Studies, vol 7, no. 1, pp. 1-13. (1975)

[52] Minggang Dong and Ning Wang, "Adaptive network-based fuzzy inference system with leave-one-out cross-validation approach for prediction of surface roughness," Applied Mathematical Modelling, Vol. 35, pp. 1024-1035, (2011).

[53] Murray, I., Arnott, J.: "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," Journal of the Acoustical Society of America 93(2) pp. 1097-1108, (1993).

[54] Nauck, D. Klawonn, F. and Kruse, R. "Foundations of neuro fuzzy systems, " *New York:Wiley*, (1997).

[55] Neiberg, D. Elenius, K. and Laskowski, K. "Emotion recognition in spontaneous speech using GMMs," *Proc. INTERSPEECH 2006*, pp. 809-812, September, (2006).

[56] Nicolaou, M. and Gunes, H. "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *Affective Computing, IEEE*, **2(2)**, pp. 92-105, (2011).

[57] Nose, T. and Kobayashi, T. "A Technique for Estimating Intensity of Emotional Expressions and Speaking Styles in Speech Based on Multiple-Regression HSMM ," *IEICE Transactions on Information and Systems*, **E93-D(1)**, pp. 116-124, (2010).

[58] Nwe, T.L. Foo, S.W. and Silva, L.C.D. "Speech emotion recognition using hidden Markov models," *Speech Communication*, **41(4)**, pp. 603-623, November, (2003).

[59] Pao, T.L. Chen, Y.T. and Yeh J.H. "Comparison of Classification Methods for Detecting Emotion from Mandarin Speech ," *IEICE Transactions on Information and Systems*, **E91-D(4)**, pp. 1074-1081, (2008).

[60] Patrick G. Hunter, E. Glenn Schellenberg, and Ulrich Schimmack. "Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions," Psychology of Aesthetics, Creativity, and the Arts, 4(1) pp. 47-56, (2010)

[61] Patel, S. and Shrivastav, R. "A Preliminary Model of Emotional Prosody using Multidimensional Scaling," *Proc. InterSpeech 2011*, pp. 2957-2960, (2011).

[62] Pereira, C. Dimensions of emotional meaning in speech. In ISCA Workshop on speech and emotion, Belfast, 2000.

[63] Pereira, C. and Watson, C. Some acoustic characteristics of emotion. In ICSLP, Sydney, 1998.

[64] Pierre-Yves, O. "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, **59**, pp. 157-183, July, (2003).

[65] Planet, S. and Iriondo, I. "Comparative Study on Feature Selection and Fusion Schemes for Emotion Recognition from Speech," International Journal of Interactive Multimedia and Artificial Intelligence ; 1(Special Issue on Intelligent Systems and Applications), pp. 44-51 (2012)

[66] Polzehl, T. Schmitt, A. and Metze, F. Approaching multilingual emotion recognition from speech - on language dependency of acoustic/ prosodic features for anger detection, in Proc. of the Fifth International Conference on Speech Prosody, 2010. Speech Prosody (2010)

[67] Russell, J. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11(11), 273-294.

[68] Ramakrishnan, S. and El Emary, I. "Speech emotion recognition approaches in human computer interaction," Telecommunication Systems, pp. 1-12, (2011).

[69] Reiter, S. Schuller, B. Cox, C. Douglas-Cowie, E. Wollmer, M. Eyben, F. and Cowie, R. "Abandoning emotion classes towards continuous emotion recognition with modelling of long-range dependencies," in Proc. 9th Interspeech, pp. 597-600, (2008).

[70] Russell, J. A. "A circumplex model of affect," Journal of Personality and Social Psychology, 39, pp. 1161-1178, (1980)

[71] Russell, J. A., and Bullock, M. . Multidimensional "scaling of emotional facial expressions: Similarity from preschoolers to adults," Journal of Personality and Social Psychology, 48, pp. 1290-1298, (1985)

[72] Shashidhar, G. K. and Sreenivasa, R. "Exploring Speech Features for Classifying Emotions along Valence Dimension," Proc. of Pattern Recognition and Machine Intelligence, India, pp.537-542, (2009)

[73] Scherer, K. R., Dan, E., and Flykt, A. (2006). What determines a feelings position in affective space? A case for appraisal. Cognition and Emotion, 20, 92-113.

[74] Scherer, K. "Vocal communication of emotion: A review of research paradigms," Speech Communication, 40(1-2), pp. 227-256, April (2003)

[75] Schlosberg, H. "Three dimensions of emotion. Psychological Review, 61, 81-88. (1954)

[76] Schroder, M. and Cowie, R. and Cowie, E.D. Westerdijk, M. and Gielen, S. "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," *Proc. Eurospeech 2001*, pp. 87-90, (2001).

[77] Shashidhar G. Koolagudi, K. Sreenivasa Rao "Exploring Speech Features for Classifying Emotions along Valence Dimension," The 3rd international Conference on Pattern Recognition and Machine Intelligence (PReMI-09), IIT Delhi, India, pp.537-542, (2009)

[78] Sugeno, M. and Kang, G.T. , 'Sturcture identification of fuzzy model," Fuzzy Sets and Systems, vol 28, no. 1, pp. 15-33. (1988)

[79] Scherer, K.R., Personality inference from voice quality: The loud voice of extroversion. European Journal of Social Psychology, 1978. 8: p. 467-487.

[80] Truong, K.P. and Raaijmakers, S. "Automatic Recognition of Spontaneous Emotions in Speech Using Acoustic and Lexical Features," In: Popescu-Belis, A., Stiefelhagen, R. (eds.) MLMI 2008. LNCS, vol. 5237, pp. 161-172. Springer, Heidelberg (2008)

[81] Takagi, T. and Sugeno, M. , "Fuzzy identification of systems and its application to modeling and control," IEEE Trans. On Systems, Man and Cybernetics, vol 15, pp. 116-132. (1985)

[82] Tato, R., Santos, R., Kompe, R., Pardo, J.M., "Emotional Space Improves Emotion Recognition," in Proc. Of ICSLP-2002, Denver, Colorado, September (2002).

[83] Fritz, T. Jentschke, S. Gosselin, N. Sammler, D. Peretz, I. Turner, R. Friederici, A. D. and Koelsch, S. "Universal recognition of three basic emotions in music," Current biology : CB, 19(7) pp. 36-57, April (2009)

[84] Tomkins, S. S., and McCarter, R. . "What and where are the primary affect? Some evidence for a theory," Perceptual and Motor Skills, 18, pp. 119-158, (1964)

[85] Tsukamato, Y. , "An approach to fuzzy reasoning method," in M.M. Gupta, R.K. Ragade, R.R. Yager (eds.), Advances in Fuzzy Set Theory and Applications, Elsevier Science Ltd, Amsterdam. (1979)

[86] Valery A. Petrushin, "Emotion in Speech: Recognition and Application to Call Centers," in Proc. ANNIE '99, pp. 7-10, (1999).

[87] Ververidis, D. and Kotropoulos, C. "Emotional speech recognition and synthesis: Resources, features, methods and applications," submitted in Speech Communication Journal, Elsevier.

[88] Vlasenko, B. Philippou-Hubner, D. Prylipko, D. Bock, R. Siegert, I. Wendemuth, A. "Vowels formants analysis allows straightforward detection of high arousal emotions," International Conference on Multimedia Computing and Systems/International Conference on Multimedia and Expo pp. 1-6 2011

[89] Vogt, T. Andr, E. Wagner, J. "Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation," Affect and Emotion in Human-Computer Interaction, pp. 75-91, (2008).

[90] Wang, Y. M. and Elhag, T. M. "An adaptive neuro-fuzzy inference system for bridge risk assessment," Expert Systems with Applications, vol. 34, pp. 3099-3106, (2008).

[91] Wagner, J. Vogt, T. Andr, E. "A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech," ACII 2007: pp. 114-125, (2007).

[92] Wolkenhauer, O., "Data Engineering: Fuzzy Mathematics in Systems Theory and Data Analysis," Wiley, (2001)

[93] Wu, D. and Parsons, T.D. and Narayanan, S. "Acoustic Feature Analysis in Speech Emotion Primitives Estimation," *Proc. InterSpeech 2010*, pp. 785-788, (2010).

[94] Wu, D. Parsons, T. Mower, E. and Narayanan, S. "Speech Emotion Estimation in 3D Space," *Proc. ICME 2010*, (2010).

[95] Wu, S. Falk, T.H. and Chan, W.-Y. "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, **53(5)**, pp. 768-785, May, (2011).

[96] Hao Wu and Dongmei Jiang and Yong Zhao and Sahli, H., "Dimensional emotion driven facial expression synthesis based on the multi-stream DBN model," Proc. of APSIPA ASC, 2012.

[97] Xie, B. Chen, L. Chen, G. C. and Chen, C. "Statistical Feature Selection for Mandarin Speech Emotion Recognition," Springer Berlin Heidelberg, 2005.

[98] Yang, Y. Wang, G. Luo, F. Li, Z. "A Data Driven Emotion Recognition Method Based on Rough Set Theory," RSCTC 2008: pp. 457-464, (2008).

[99] Yu, C. Aoki, P.M. and Woodruff, A. "Detecting User Engagement in Everyday Conversations," *Proc. Eighth Intl Conf. Spoken Language Processing*, (2004).

[100] Zhou, J., Wang, G., Yang, Y. and Chen, P. "Speech Emotion Recognition Based on Rough Set and SVM," In Proceedings of the Firth IEEE International Conference on Cognitive Informatics ICCI, Beijing, China, pp. 53-61, (2006).

[101] Zhang, Q. Jeong, S. Lee, M. "Autonomous emotion development using incremental modified adaptive neuro-fuzzy inference system," *Neurocomputing*, **86**, pp. 33-44, (2012).

Publications

Journal

[1] Elbarougy, R. and Akagi, M. "Improving Speech Emotion Dimensions Estimation Using a Three-Layer Model for Human Perception," *Journal of Acoustical Science and Technology*, (in press).

International Conferences

[2] Elbarougy, R. and Akagi, M. "Cross-lingual Speech Emotion Recognition System Based on a Three-Layer Model for Human Perception," *Proceedings of International Conference (APSIPA2013 ASC)*, Kaohsiung, Taiwan, November 2013.

[3] Elbarougy, R. and Akagi, M. "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model," *Proceedings of International Conference (APSIPA2012 ASC)*, Hollywood, USA, December 2012.

Domestic Conferences

[4] Elbarougy, R. and Akagi, M. "Cross-lingual Speech Emotion Dimensions Estimation Based on a Three-Layer Model," *Proceedings of the Autumn Meeting of the ASJ*, September 2013.

[5] Elbarougy, R. and Akagi, M. "Automatic Speech Emotion Recognition Based on Dimensional Approach," *Proceedings of IEICE Speech Meeting, SP2012-127*, March

2013.

[6] Elbarougy, R., Tokuda, I., and Akagi, M. "Acoustic Analysis of Register Transition between Chest-to-Head Register in Singing Voice," *Proceedings of the Spring Meeting of the ASJ*, March 2013.

[7] Elbarougy, R. and Akagi, M. "Comparison of Methods for Emotion Dimensions Estimation in Speech Using a Three-Layered Model," *IEICE Technical Report, Speech 112(81)*, pp. 19-24, June 2012.