博士論文

Research on Emotion Recognition using Speech and Physiological Signals

（音声・生体情報を用いた感情識別手法に関する研究）

張　皓

Table of Contents

1 Introduction 1

1.1 Social background 1
1.2 Research background 3
1.3 Remaining issues 5
1.4 Research purpose 5
1.5 Structure of this thesis 7

2 Valence and arousal recognition using EEG signals 11

2.1 Research background 11
2.1.1 Emotion theory 11
2.1.2 Emotion model 13
2.1.3 Emotion recognition using EEG signals 15
2.2 Adopted methods 18
2.2.1 Wavelet analysis 18
2.2.2 Statistical parameters 20
2.2.3 Genetic algorithm 21
2.2.4 Principal component analysis (PCA) 22
2.2.5 Probabilistic neural network (PNN) 23
2.3 Proposed EEG-based emotion recognition method 25
2.3.1 Cross-level wavelet feature extraction 25
2.3.2 Classification 28
2.4 Experiments 29
2.4.1 Valence elicitation 29
2.4.2 Experimental protocols 31

2.4.3 EEG dataset 32
2.5 Results 33
2.5.1 Valence recognition results 33
2.5.2 Comparisons 39
2.5.3 Study on the effectiveness of EEG electrodes 41
2.6 Discussion 43
2.7 Application on emotional arousal detection 46
2.8 Summary 48

3 Emotional speech database construction 49

3.1 Research background and remaining issues 49
3.2 Procedure of experiments 51
3.2.1 Online survey 51
3.2.2 Onsite experiment 52
3.3 Description of signals 54
3.4 Description of data 60
3.5 Speech data selection 64
3.5.1 Strategy for speech data selection 66
3.5.2 Results 68
3.6 Summary 69

4 Purely segment-level speech emotion recognition 70

4.1 Research background 70
4.2 Defining terminology of emotions 75
4.3 Utterance level or segment level? 77
4.4 Segment level based speech emotion recognition 79
4.4.1 Experimental data for evoking emotions 79
4.4.2 Methodology 82

4.4.3 Results 89
4.4.4 Discussion 93
4.5 Real-world speech emotion recognition by proposed methods 98
4.5.1 Results of four-emotion recognition 98
4.5.2 Comparison results 99
4.5.3 Results of database evaluation 100
4.5.4 Further validation of four-emotion recognition 101
4.5.5 Discussions 102
4.6 Application perspective: Emotion strength analysis 102
4.6.1 Experimental data 103
4.6.2 Results 105
4.6.3 Discussions 105
4.7 Summary 106

5 Conclusions 107

References 111

Acknowledgements 129

List of Figures

Figure 1. The structure of this thesis..... 10
Figure 2. Illustration of James-Lange theory. 12
Figure 3. Illustration of Cannon-Bard theory. 12
Figure 4. Illustration of Schachter-Singer Two-Factor theory. 13
Figure 5. Emotion dimensional model. 15
Figure 6. Illustration of wavelet decomposition. 19
Figure 7. PNN structure. 24
Figure 8. Schematic of cross-level wavelet feature estimation from EEG signals. 29
Figure 9. Indication of pictures selected from IAPS for valence stimulation (Males). The red dots represent the pictures selected from IAPS among all pictures, shown as gray dots. 30
Figure 10. Indication of pictures selected from IAPS for valence stimulation (Females). The red dots represent the selected pictures among all pictures, shown as gray dots. 30
Figure 11. Emotional valence and arousal stimulation procedure. 31
Figure 12. Test cases of brain activity in different areas with their notations. ... 32
Figure 13. Visualization of feature space using PCA. (a)-(d) were calculated from raw EEG signals, (e) MLWF was the std calculated using the best performance level after wavelet decomposition, and (f) CLWF was the std calculated using the cross-level decomposed signals from discrete wavelet decomposition selected using GA..... 34
Figure 14. Average results for valence detection on 50 participants using different feature groups and EEG sets. a, Average accuracy for 3-level

valence detection using 12 different feature groups (mean ± s.e.m). b, Comparisons of effectiveness for 3-level valence detection using EEG signals in different brain areas by proposed features (CLWF); ****p<0.00001, *****p<0.000001 by anaysis of variance (ANOVA) plus Tukey's Honestly Significant Difference (HSD) test. c, Comparison of 2-level (L1: 98.4% and L3: 97.8% respectively) and 3-level (L1: 94.0%, L2: 86.6%, and L3: 90.2% respectively) overall valence classification accuracy using proposed features (CLWF); *****p<0.000001 by paired t-test. 36
Figure 15. Results of comparing valence recognition performance with std using dynamic statistical parameter selection. 39
Figure 16. Comparison results of valence recognition with std using SVM and PNN. 40
Figure 17. Comparisons of different feature selection methods. 41
Figure 18. GA based EEG electrodes reduction. a, GA-reduced EEG electrodes distribution; no. 1-16: Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6. b, Comparison results of GA-reduced EEG sets; ****p<0.00001 by ANOVA plus Tukey's HSD throughout the figure. c, Selected electrodes' brain area summary. 42
Figure 19. Indication of pictures selected from IAPS for arousal stimulation (Males). The red dots represent the pictures selected from IAPS among all pictures, shown as gray dots. 46
Figure 20. Indication of pictures selected from IAPS for arousal stimulation (Females). The red dots represent the selected pictures among all pictures, shown as gray dots. 47
Figure 21. Results of arousal recognition. 47
Figure 22. Environment setting for experiment from viewpoint of coordinator. 52
Figure 23. Procedure for eliciting emotions by recalling experiences. 53
Figure 24. Placement of sensors and signals that were collected. 54

Figure 25. Image of bioplux. 55
Figure 26. Example of collected speech signals. 56
Figure 27. Example of collected EEG signals. 57
Figure 28. ECG signal with QRS positions. 58
Figure 29. Respiration signal with explanations. 58
Figure 30. EMG signal illustration. 59
Figure 31. Skin temperature signal. 59
Figure 32. Signal from blood volume pulse. 60
Figure 33. Data distribution related to gender and age. 61
Figure 34. Summary of self-assessments. 61
Figure 35. Summary of self-assessment results according to gender. Data (y-axis) indicate percentage of confidence levels no less than three. 62
Figure 36. Summary of self-assessment results according to two age groups of less and greater than 30. Data (y-axis) indicate percentage of confidence levels no less than three. 62
Figure 37. Illustration of data selection. 68
Figure 38. Illustration of data selection results. 68
Figure 39. Feature extraction schemes illustrations of a short and a long utterances. 73
Figure 40. Illustrations of segmentation schemes. 83
Figure 41. Illustration of proposed segmentation approach. 84
Figure 42. Fixed length segment positions illustration using proposed segmentation approaches (20-segment selecting situation is shown and the positions are represented using grey lines). 'S' is included in the abbreviation to represent the purely segment-level concept. 88
Figure 43. Illustration of segment-level classification concept for decision model. 89

Figure 44. Comparison of emotion recognition accuracy between existing segmentation schemes using segment features with global features. 90
Figure 45. Speech emotion recognition accuracy based on segment-level analysis using absolute interval segment features (SATI)..... 92
Figure 46. Comparison of emotion recognition accuracy between different segmentation schemes using purely segmented features and PNN (σ=0.2).93
Figure 47. Confusion matrix of four-emotion recognition in segment level (10-fold cross validation). 99
Figure 48. Confusion matrix of four-emotion recognition in utterance level (10-fold cross validation). 99
Figure 49. Comparison of conventional and proposed methods. 100
Figure 50. Illustration of original and selected speech databases. 100
Figure 51. Emotion recognition performance using different databases. 101
Figure 52. Confusion matrix of four-emotion recognition..... 101
Figure 53. Defined emotion strength based on IAPS. 103
Figure 54. Experimental protocols. 104
Figure 55. Statistical analysis for emotion components using segment-level speech emotion analysis for all speech samples. 105

List of Tables

Table 1. Top ten reasons for death in Japan. 2
Table 2. Frequency range of different wavelet decomposition levels with EEG frequency band information..... 26
Table 3. Average results (mean ± std) for two-level and three-level valence detection under different stimulations (50 subjects). 38
Table 4. Feature selection methods for control study 40
Table 5. Summary of existing emotion databases (DB). 51
Table 6. The advantages and disadvantages of different brain activity sensing technologies. 66
Table 7. Representation of emotions in dimensional space. 67

1 INTRODUCTION

In this thesis, an emotion recognition method by using electroencephalography (EEG) is proposed with high accuracy and it is applied on evaluating collected real-world emotional speech data for constructing a Japanese emotional database. An emotion recognition method using speech based on segment level feature analysis is proposed and validated using proposed database. In this chapter, background, research purpose and targets are introduced.

1.1 SOCIAL BACKGROUND

In recent years, most developed countries are facing serious issues with the increasing number of lifestyle related diseases. In Japan, according to the statistics from the Ministry of Health, Labour and Welfare, about 55% of people death is due to cancer, heart diseases, and cerebro-vascular diseases, which are greatly related to unhealthy lifestyles [1]–[3], more detailed statistics from the Ministry of Health in Japan for the year of 2009 are shown in Table 1. It can be easily inferred that changing circumstances readily affect human perspective and subjective feelings, negative and inconstant emotions directly or indirectly cause variety of diseases.

Table 1. Top ten reasons for death in Japan.

Cause of death	Mortality	Mortality rate	Percentage in all death
Cancer	344105	273.5	30.1
Heart disease	180745	143.7	15.8
Cerebro-vascular disease	122350	97.2	10.7
Pneumonia	112004	89.0	9.8
Caducity	38670	30.7	3.4
Accident	37756	30.0	3.3
Suicide	30707	24.4	2.7
Nephropathy	22743	18.1	2.0
Hepatopathy	15969	12.7	1.4
Chronic pulmonary disease	15359	12.2	1.3

Health psychologists explain a different health status of an individual in aspects of experiencing positive and negative emotions. Negative emotions have deleterious effects on health. This hypothesis is started with a phenomenon of cardiac disorders of soldiers; it is found that negative emotional experiences such as combat experiences have long-term effects on human health in later life [4]. Then it follows a large number of researches indicate that negative emotions are related to coronary heart disease or cardiovascular disorders. Anger and aggressive behavior play an essential role in the hypothesis that emotions influence physical health and emotions are a precipitating factor in Menieres disease [5].

Other than negative emotions, positive emotions play a protective role in keeping people from diseases. In a study of relationship between positive emotions and health based on 2 years' medical data from 1014 patients, it indicates that higher levels of positive emotions were associated with decreased likelihood of diseases such as hypertension and diabetes [6]. Another study shows that the happiest group of people had few diseases such as hypochondriasis and depression [7]. Furthermore, positive affect such as happiness has been argued to be distinguishing factor of depression [8].

To summarize, there exists much phenomenon or experimental data reflecting the fact that the positive emotions are a protective factor for human health while negative ones can increase risks for getting various diseases. Besides the direct influence of physical or mental health, they also engage in behaviors that damage one's health such as alcoholism, smoking, drug abuse, etc.

1.2 RESEARCH BACKGROUND

Recently, many studies on engineering methodologies to recognize emotions automatically have been proposed due to the growing needs of improving service quality in healthcare, human communication, commercial, etc. [9]–[13]. Emotion recognition is an interdisciplinary research evolving many research fields such as neuroscience, engineering, design, computer science, and others. Though many researchers are interested in exchanging and discussing ideas and making effort

to share a roadmap, there are difficulties in standardization and different focuses for emotion research regarding different purposes in each research field. In late 1990s, researchers initially tried to use facial and vocal expressions for emotion recognition and it's known as emotional sensing technology. Lots of discussions and disputes still exist in emotion research; a very important reason is that both psychology and cognitive sciences have not reached to the common conclusion about how emotions are formed. However, it comes to a trace in the research of emotion recognition methods where machines are able to detect human emotions for improving human life using information technologies.

The debate on emotion categories or whether emotions can be distinguished is far from being resolved in psychology and neurophysiology. As for human beings, we try to understand others' emotions by their ways of speaking, facial expressions, gestures, and context information. Basically, human receive all possible information and make a proper guess of others' emotion. And even professionals such as analysts and doctors cannot absolutely identify others' emotional status. In addition, as for engineering, there is no research trace of mechanism of emotion generations. In this situation, there is no evidence for proving exactly a certain emotion or make absolute link to existing evidences based on theoretical knowledge. Indeed, people might want to build an implicit theory of emotions or implicit personality theory until the theory development

came to a point of realizing the role of cognitive evaluations in emotions [14]–[16]. The scientific theory of emotion is still open. So it is crucial to set reasonable emotion categories, gather and validate data by referring as well as improving previous research results. More detailed information is introduced in each chapter.

1.3 REMAINING ISSUES

Emotion monitoring is an important research issue for healthcare, and speech monitoring is the most convenient and natural way to realize it. However, emotion recognition accuracy using speech signals still needs improvements. In order to evaluate the proposed emotion recognition methods, new database elicited by real experiences is necessary but not assessable. Moreover, current emotion evaluation method for speech, which is called as other persons' assessment, is not effective for real-world emotions. Alternative data assessment method is necessary for data selection.

1.4 RESEARCH PURPOSE

The purpose of this thesis is to propose a new emotion recognition method using speech signals. It has been tackled through the following three targets.

1. Propose a new classification approach of emotion recognition using speech.

Novel method based on segment-level feature extraction and classification for speech emotion recognition is proposed and examined. The proposed method is validated using a large-scale emotional speech database constructed in this research.

2. Construct an emotion database with natural emotional speech elicited by real experiences.

Experiments have been designed for collecting speech signals under different emotional states. In order to build a robust emotion recognition method, a natural speech database is necessary for validating the performance. In this thesis, I focus on real-world emotions, which is more realistic speech data collected from recalling real life situations, instead of acted emotions, to avoid incorrect conclusions. Additional method has been applied to select data that is included in the following speech emotion recognition method using EEG signals.

3. Propose a new emotional speech data evaluation method using EEG signals.

By improving the accuracy from proposing novel method, personal emotional states can be recognized using only EEG signals, and this technology enables EEG signals to be used as references of emotions.

1.5 STRUCTURE OF THIS THESIS

The structure is demonstrated as follows and the relationship of major chapters is illustrated in Figure 1.

Chapter 1 Introduction

Previous researches are reviewed and summarized. In addition, research purpose and targets, and the structure of this thesis are introduced.

Chapter 2 Valence and arousal recognition using EEG signals

Emotion theories and models are reviewed, EEG signals are known as one of the most reliable signals for emotion recognition, and they are used as references in this study for selecting high quality emotional speech data. For achieving high accuracy for emotion recognition using EEG signals, signals from different EEG electrodes are considered independently in order to find an optimum combination through different levels of wavelet coefficients based on the genetic algorithm (GA). A new set of features named the cross-level wavelet feature group (CLWF) is proposed. The procedure of selecting one level from eight decomposed wavelet coefficients based on GA can be considered as feature selection with prior knowledge guidance. In addition, a reduced EEG set using less number of electrodes are studied and discussed, the conclusions from analytical results also

show the constancy with previous researches; improvements can be made by using different statistical parameters according to the distribution of EEG signals.

Chapter 3 Emotional speech database construction

In order to evaluate the robustness of proposed speech emotion recognition method with emotional speech data elicited by real experiences, I designed experiments for collecting speech signals under different emotional states. Self-assessments are conducted after emotional experience recalling of each emotion, and EEG-assessments are conducted after the onsite experiments. Four emotions are defined on the arousal-valence space so that emotions during speaking can be identified using EEG signals by arousal and valence recognition method developed in Chapter 2.

Chapter 4 Purely segment-level speech emotion recognition

Previous speech emotion recognition schemes are reviewed. And the reasons of adopting segment-level approach are presented. I address the quantitative analysis of various analytical schemes related to segment-level speech emotion recognition, and propose an automatic approach for selecting a number of the most representative samples in order to improve the classifier generalization ability. I propose several segmentation strategies, entropy-based ATIR (eATIR),

mutual information-based ATIR (miATIR), and correlation coefficients based ATIR (crATIR).

I established a model using these segment-level speech frames at selected positions. The decision for determining the emotion of an utterance is based on the prediction of its segments from a classifier by applying the majority vote method. Testing on two and four labels emotion recognition has been carried out both on a 50-person emotional speech database. Significant improvements in the level of accuracy have been achieved.

Chapter 5 Conclusions

This thesis is concluded in this chapter.

Chapter 6 Application perspective: Emotion strength analysis

Finally, the application perspective of emotion strength analysis is demonstrated and discussed.

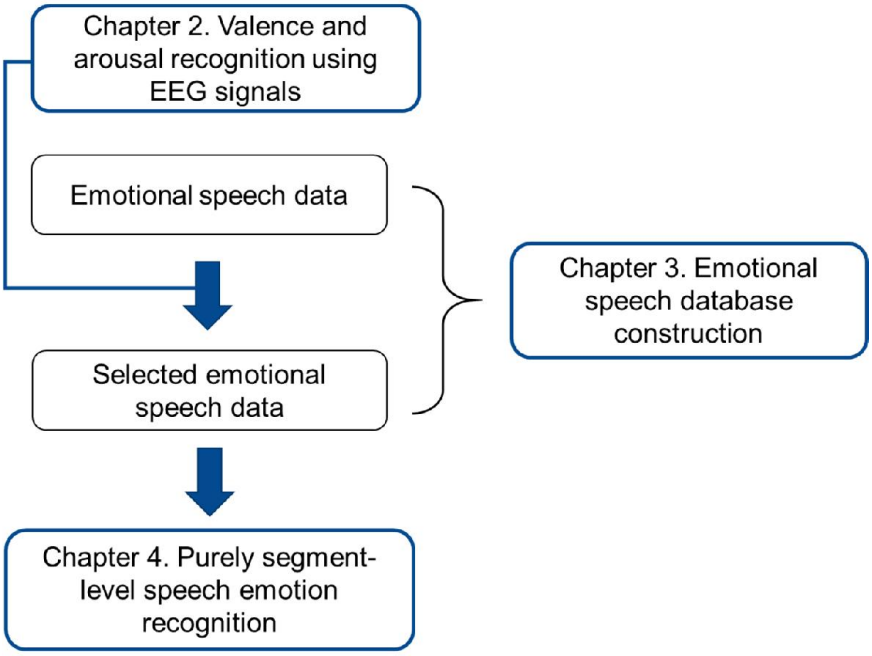

Figure 1. The structure of this thesis

2 VALENCE AND AROUSAL RECOGNITION USING EEG SIGNALS

In this chapter, a novel EEG-based emotion recognition method is proposed for better performance and understanding unknown aspects of relationship between brain activities and human emotions.

2.1 RESEARCH BACKGROUND

2.1.1 Emotion theory

Psychologists and ecologists have introduced different theories of emotions. Back to 19th century, Charles Darwin argued that emotions have universal cross-cultural counterparts and emotions exist because their beneficial to human or animal's survival [17]. This theory is from the perspective of evolutionary theory and it's the basis for discrete model of emotions. Paul D. MacLean further developed this theory by introducing more abstract reasoning and more instinctive responses [18]. Major theories are James-Lange theory [19], [20], Cannon-Bard theory [21], and Schachter-Singer Two-Factor theory [22], [23].

James-Lange theory argues that physiological change is primary and emotion is secondary. Emotion is experienced after the brain receives information from the nervous system. Physiological changes are known as changes in heart rate,

perspiration, muscular tension, dryness of mouth, etc. Without physiological sensation, the emotion will not be existed according to this theory. The schematic is shown as follows in Figure 2.

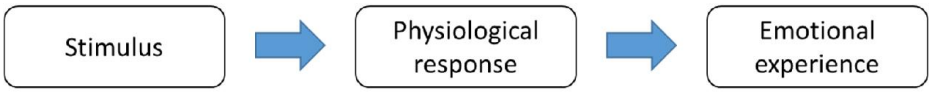

Figure 2. Illustration of James-Lange theory.

Cannon-Bard theory also agrees that physiological responses play a very important role in emotions, but this theory argues that brain is essential to emotions as well as physiological responses generating. It states that an emotional expression result from action of subcortical centers while it doesn't consider the primary role of physiological change. Firstly, it considers that bodily changes and emotional experiences are separate processes, they do not necessarily happen in an absolute order and can happen simultaneously; Secondly, thalamic discharge is essential for emotional experience and also for bodily changes. The schematic is illustrated in Figure 3.

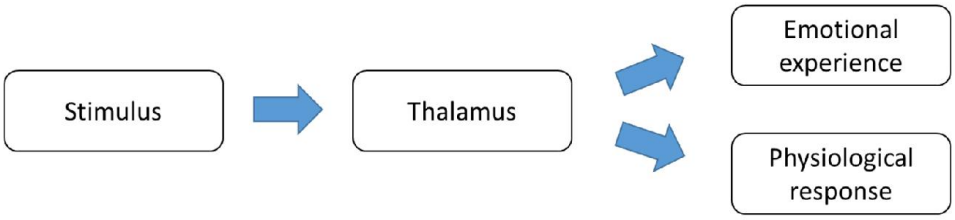

Figure 3. Illustration of Cannon-Bard theory.

Schachter-Singer Two-Factor theory states that physiological arousal and cognitive label are two factors for formulating emotion. Based on this theory, emotions might be misinterpreted by only body's physiological state. Human beings actually feel or label an emotion from cognitive evaluation based on clues from external stimuli. It claims that emotion varies with the same stimuli and different cognition. The schematic is shown in Figure 4.

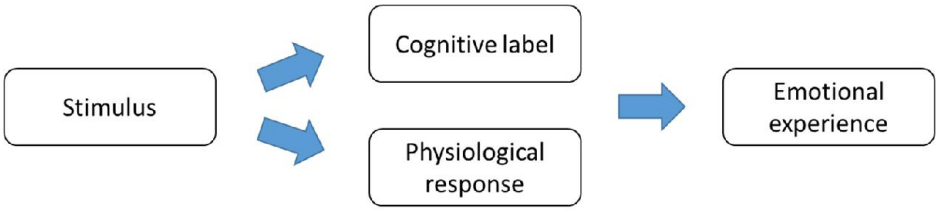

Figure 4. Illustration of Schachter-Singer Two-Factor theory.

By reviewing the major emotion theories, brain activities play important roles across all theories. It can be referred that brain signals are related to human emotions.

2.1.2 Emotion model

There exist multiple strategies for labeling human emotions such as dimensional and discrete emotion models.

Discrete emotion model [24], [25] assumes that there are limited numbers of core emotions while thousands of other emotions can be derived by these core emotions. In Darwin's study [26], this model was initialed by describing several

behavior and physiological processes that are associated with several kinds of emotions of human and animals. Then many researchers further developed this model. William McDougall was the first to believe biological instincts cause emotions. Aristotle claimed that emotions associate with certain types of bodily reactions. William James also believes in discrete model, he considered emotions were made from mental events. However, he thought these events are being broken down to smaller elements that are not a certain emotion. James and Deway considered that the emotions are not only associated with different neutral and physiological processes but also related to different experiences. There exist more researchers believe discrete emotion model and try to define a number of basic emotions [27]–[31]. Human emotions are complex and mixed with many factors such as inner thoughts or personal traits. It is possible, however, that all emotions could be understood by recognizing basic emotions shared across different cultures based on discrete model of emotions. Ekman et al. published a paper describing six basic emotions such as surprise, fear, happiness, sadness, anger and disgust after a series of cross-cultural studies [28].

Another kind of emotion model called dimensional model describing emotions using multiple fundamental properties shared among all emotions [32]. Most researchers agree that emotion has at least two dimensions: valence and arousal [33]–[37]. Both valence and arousal can be defined as subjective experiences.

Valence is a subjective feeling varies from pleasantness to unpleasantness; Arousal is a subjective state of feeling varies from activate to deactivate. It has been reported that human emotions can be reflected individually by subjective experiences of valence and arousal.

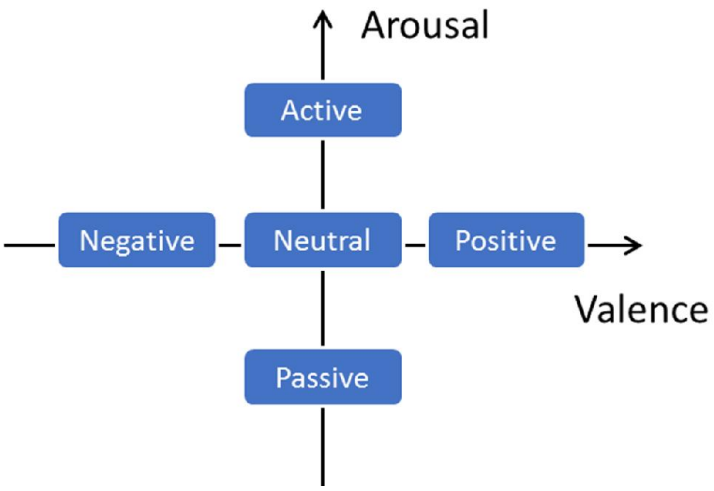

Figure 5. Emotion dimensional model.

2.1.3 Emotion recognition using EEG signals

Emotion recognition from brain activities is a challenging task, but it draws more attention lately because of the latest significant developments in brain-computer-interface (BCI) researches and also the eager of researchers for understanding how brain activities reflect human emotions with the latest developments in data analysis methods and neurosciences [38]–[42]. EEG is one of the most popular brain signals that have many advantages such as less noisy,

non-invasive, high temporal resolution, etc. Even though, it is a relatively new research field of EEG-based emotion recognition [43]–[53]. Previous researches are summarized as follows.

1. Features

Statistical-based features. Many researches adopt statistical-based features for EEG-based emotion recognition such as the mean, the standard deviation, the mean of the absolute values of the first differences of the raw signal, the mean of the absolute values of the first differences of the standardized signal, the mean of the absolute values of the second differences of the raw signal, the mean of the absolute values of the second differences of the standardized signal, and so on.

Frequency domain features. EEG bands have been defined for better understanding brain activities. They are delta, theta, alpha, beta, and gamma; each of them represents different range of frequencies. Power spectral density values or relative power values have been applied extensively on multiple channels for emotion recognition. Moreover, more features such as peak alpha frequency, alpha power have been used.

Time-frequency domain features. EEG signals have been transformed using wavelet transformations. The features including wavelet energy and wavelet entropy are extracted from the transformed signals which contain both time and

frequency domain information. Some researches propose statistical features directly on transformed signals based on wavelet transformation.

2. Channel selection

EEG-based emotion recognition has its own special characteristics other than speech emotion recognition or other physiological signal based emotion recognition. EEG signals are collected from multiple electrodes placing along the scalp; there are typical EEG electrodes set such as International 10-20 system designed for commonly using in different researchers for comparisons of performances. Researchers have special interests in reducing the number of electrodes used for EEG-based emotion recognition. Different areas were analyzed for studying the effects of emotions. For instance, 12 clusters (six cortical zones for each hemisphere) were separated for statistical analysis including anterotemporal, frontal, central, parietal-temporal, parietal, and occipital for each hemisphere [54].

To summarize, current EEG-based emotion recognition performance shows the potential of using EEG signals for emotion recognition; however, further efforts have to be made to improve the performance such as feature selection.

2.2 ADOPTED METHODS

The methods used for proposing novel EEG-based emotion recognition algorithm are demonstrated in this session.

2.2.1 Wavelet analysis

I adopted discrete wavelet decomposition [55] to better understand the frequency and location information of the EEG signals. Wavelets are a mathematical tool that can be used to extract information from many kinds of data, such as those from audio and images. As wavelet transform contains information on both time and frequency domains, it represents a powerful tool to analyze and observe details on non-periodic signals. This technology has also been proved to be powerful when applied to the field of emotion classification [45], [48], [50], [56], [57]. In this study we used a discrete wavelet transformation (DWT) whose definition is

$$T_{m,n} = \int_{-\infty}^{\infty} f(t)\psi_{m,n}(t)\,dt \tag{1}$$

$$\psi_{m,n}(t) = a_0^{-m/2}\psi\left(a_0^{-m}t - nb_0\right) \tag{2}$$

Where $\psi(*)$ is a wavelet function, the integer m controls wavelet dilation and n controls translation. Here, a_0 is a specified fixed dilation step parameter set at a value greater than one, and b_0 is the location parameter, which must be greater than zero. $T_{m,n}$ are the discrete wavelet values given on a

scale-location grid of index m, n and known as wavelet coefficients or detail coefficients.

The decomposition of approximation coefficients into approximation and detail coefficients at subsequent levels can be schematically illustrated as follows in Figure 6; $cA1 - cA(n)$ are approximate coefficients and $cD1 - cD(n)$ are detail coefficients. The signal at each level is decomposed into low and high frequencies.

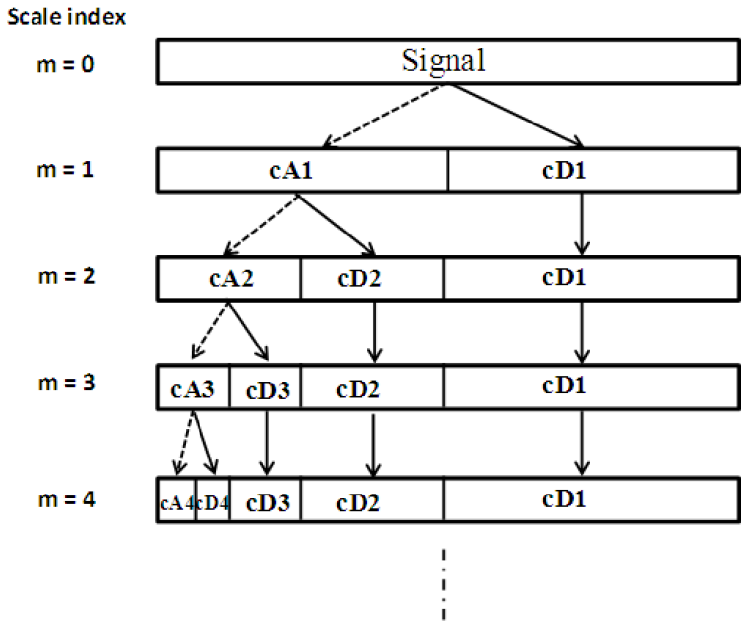

Figure 6. Illustration of wavelet decomposition.

Wavelet analysis has been successfully used for feature extraction and for many classification tasks based on EEG. I selected Daubechies 5 wavelet (Db5)

since it has been adopted successfully for modeling complex signals such as EEG signals [46], and I applied seven levels of wavelet decomposition, which is determined by taking the EEG frequency bands into consideration.

2.2.2 Statistical parameters

Statistical parameters shows effectiveness on demonstrating affective states from physiological signals [58]. Several statistical parameters including the mean, the standard deviation (std), the skewness, and the kurtosis are used in this study to demonstrate the characteristics of transformed EEG signals using discrete wavelet decomposition. The formula is shown as follows.

1. The mean of raw signal

$$\mu X = \frac{1}{N}\sum_{n=1}^{N} X(n) \tag{3}$$

2. The standard deviation of raw signal

$$\delta_X = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(X(n)-\mu_x\right)^2} \tag{4}$$

3. The skewness of raw signal

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} \tag{5}$$

where μ_i is the i_{th} central moment. A moment μ_n of a univariate probability density function $P(x)$ taken about the mean $\mu = \mu_1'$,

$$\mu_n = \int (x-\mu)^n P(x)\,dx \tag{6}$$

4. The kurtosis of raw signal

$$\beta_2 = \frac{\mu_4}{\mu_2^2} \tag{7}$$

where μ_i is the i_{th} central moment.

2.2.3 Genetic algorithm

I adopted a genetic algorithm [59] for feature selection. The GA mimics the process of natural evolution to find beneficial adaptations to a complex environment. A chromosome in GA is an encoding that represents the decision variable of an optimization problem. A finite set of chromosomes in GA is called a population. Each chromosome is rated by the fitness function on its 'fitness', which determines how good it is in solving the optimization problem. Crossover refers to the generation of two new offspring by mating two parental chromosomes. A mutation simply flips randomly the binary value of one or more bits. Crossover and mutation provide opportunities for chromosomes that have higher fitness values to evolve.

The reasons why I adopt GA are shown as follows. According to previous studies, the correlation theory of brain activities asymmetry and emotions were proved by researchers [60]. GA allows us to realize the idea to select appropriate information from each EEG electrode and eventually obtain the overall pattern of brain activities monitored by multiple EEG electrodes. Other data-driven feature selection approaches cannot give attention to the overall pattern of brain activities. The selected features can be only extracted from one or several EEG electrodes so that it fails to demonstrate the whole brain activities asymmetry since there is no priori biological knowledge applied on those algorithms. These considerations are supported by a recent paper proposed in a similar application concerning neuroimaging [61], as they have two findings related to feature selection including data-driven feature selection was no better than adopting whole data and a priori biological knowledge was effective to guide feature selection. As for the implementation of GA with proposed features, I generalized the problem of wavelet level selection according to different EEG positions to an optimization issue of minimizing the fitness function.

2.2.4 Principal component analysis (PCA)

PCA [62] is a common visualization method that is applied through dimensionality reduction by performing a covariance analysis between factors. Technically, a principal component can be defined as a linear combination of

optimally weighted observed variables, the first k principal components capture the greatest variance in the data among all k-dimensional orthonormal linear combinations of the original variables. In this research, PCA was used for feature space visualization to observe the separability of different valence levels by specific features.

Suppose that x is a vector of p variables. The first step is to search linear function $\alpha' x$ of the elements of x having maximum variance, where α_1 is a vector of p constants $\alpha_{11}, \alpha_{12}, \cdots, \alpha_{1p}$ and ' denotes a transpose, so that

$$\alpha_1' x = \alpha_{11} x_1 + \alpha_{12} x_2 + \cdots + \alpha_{1p} x_p = \sum_{j=1}^{p} \alpha_{1j} x_j \tag{8}$$

The next step is to search linear function $\alpha_2' x$, which is uncorrelated with $\alpha_1' x$ having maximum variance; the k_{th} stage, which is linear function $\alpha_k' x$ is found that has maximum variance subject to being uncorrelated with $\alpha_1' x, \alpha_2' x, \cdots, \alpha_{k-1}' x$. The k_{th} derived variable, $\alpha_1' x$, is the k_{th} PC.

2.2.5 Probabilistic neural network (PNN)

Artificial neural networks are effective when used with such signals as EEG since they are very robust in dealing with non-linear and complex signals. Moreover, the fault tolerance of artificial neural networks is necessary in order to

reduce the influence of noise. PNN [63] is one kind of artificial neural network that has been proven to be suitable for classification tasks by many researchers [64]–[66]. The operations are designed into a multi-layered feed-forward network with four layers. The illustration with input features X is shown in Figure 7.

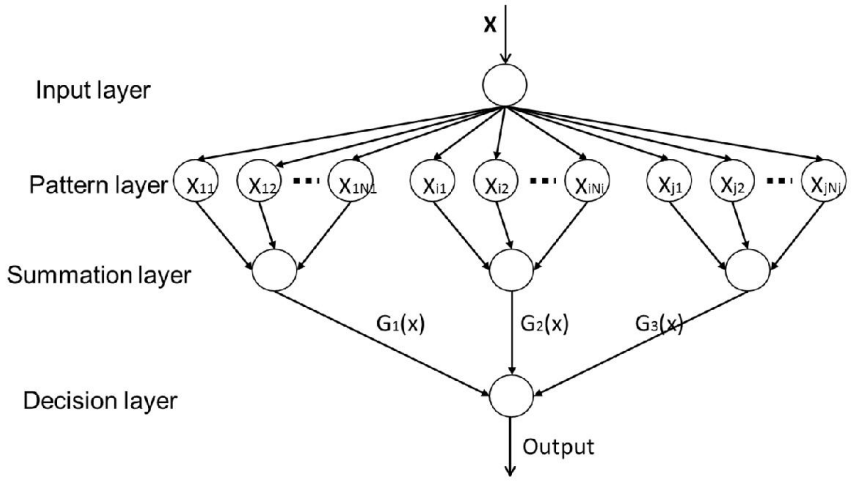

Figure 7. PNN structure.

In this work, PNN is adopted considering its characteristics of fast training process and additional training samples can be added without extensive retraining, which make it practical for model improvement when we have large training database. The inputs of PNN were the optimized features, and the output was the predicted valence level.

2.3 PROPOSED EEG-BASED EMOTION RECOGNITION METHOD

After listing all the methods necessary for the new proposed emotion recognition method, the classification scheme is demonstrated in this section.

2.3.1 Cross-level wavelet feature extraction

The statistical parameters have been shown to be useful for extracting features from raw or processed EEG signals. Although time or frequency domain features from EEG signals failed to provide good accuracy for valence detection [43], [52], [67], we explore new possibilities in time-frequency domain features. A study tries to cluster different emotions by wavelet features from EEG signals using 24 and 63 channels [68] and represented the potential of time-frequency features for identifying emotions. The theory of wavelet decomposition can roughly takes into consideration EEG frequency bands. The frequency information of decomposition levels is listed in Table 2.

Table 2. Frequency range of different wavelet decomposition levels with EEG frequency band information.

Decomposition level	Frequency range (Hz)	EEG frequency band
A7	0-4	Delta
D7	4-8	Theta
D6	8-16	Lower Beta and Alpha
D5	16-32	Upper Beta
D4	32-63	Lower gamma
D3	63-125	Upper gamma
D2	125-250	—
D1	250-500	—

The strategy is to formulate a novel cross-level wavelet feature group (CLWF) for valence detection based on GA, as opposed to the strategy of extracting a mono-level wavelet feature group (MLWF), whose levels are the same through all electrodes for each subject. EEG consists of a multi-channel system that can record a great deal of information with a lot of noise. Many researches have focused on how to omit a number of electrodes and select the most useful ones in order to achieve a moderate level of performance [69]. The studies that have adopted wavelet features have only considered a single decomposition level or have combined levels from all EEG electrodes [56]. However, those studies ignore the fact that useful information for classification may be represented in

different frequency ranges among different EEG electrodes. If we implement this idea by a brute-force search, the order of magnitude for the searching cases will be n^m, where m is the number of EEG electrodes, and n represents the levels of wavelet decomposition. For instance, there will be 152,587,890,625 cases with 16 EEG electrodes and 5 levels of wavelet decomposition, which makes it extremely difficult to find the optimized feature combinations. Thus, GA is introduced in this step to solve this optimization issue. The fitness function is designed as follows.

$$Fitness = 1 - Accuracy \tag{9}$$

Accuracy is calculated using the leave-one-out cross validation (LOOCV) method based on PNN. The input feature for GA is a matrix with n levels of wavelet decomposition for all EEG electrodes. A chromosome is an array of numbers that represent levels of wavelet coefficients related to EEG electrodes; the GA will select one level from each electrode to formulate the optimized CLWF. For chromosome design, three binary digits represent 8 levels where GA will select.

A relatively limited population size (100) and high mutation rate (0.8) as well as the LOOCV design in the fitness function are for preventing over-fitting in GA. A schematic of the entire process of feature estimation is illustrated in Figure 8. In this figure, EEG signals from positions O2, F3, and Fp1 are used as

examples to illustrate the feature extraction process. The best combination of wavelet coefficients (those indicated in red in Figure 8 are selected based on GA from a huge number of combinations, and the statistical parameters are calculated for a later classification process.

2.3.2 Classification

Since each EEG record corresponds to picture stimulation from IAPS, it is easy to give labels to the classification tasks. The approach is to consider two testing scenarios for each individual including two valence levels (level 1: displeasure, level 3: pleasure) and three valence levels (level 1: displeasure, level 2: neutral, level 3: pleasure) for model validation by calculating the accuracy based on LOOCV using PNN. The same scenarios are tested also on two and three arousal levels.

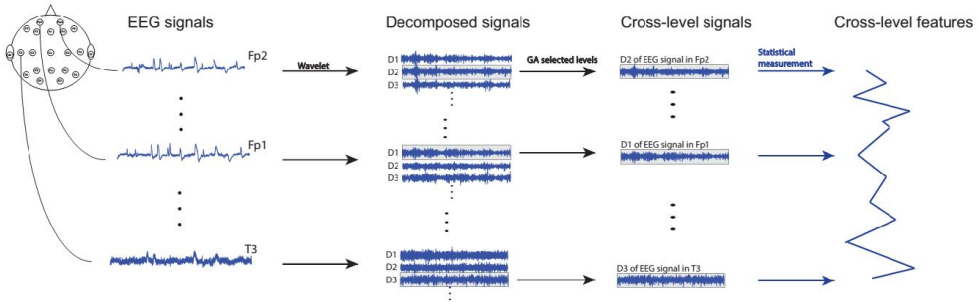

Figure 8. Schematic of cross-level wavelet feature estimation from EEG signals.

2.4 EXPERIMENTS

In order to develop and validate the methods, EEG signals are collected according to the following described procedures.

2.4.1 Valence elicitation

Three levels of valence and arousal have been defined based on arousal-valence space based on the emotion model proposed by Russell [32]. The most widely adopted emotion elicitation technique is to use pictures to evoke various emotion states based on valence-arousal space. I adopted a very popular picture-based, emotion-evoking database called the International Affective Picture System (IAPS) [70], which contains pictures labeled with values of valence and arousal and extensively adopted by many researchers [34], [71]–[75]. The IAPS labels respective pictures for males and females, so we selected different pictures for male and female subjects in order to achieve the same valence level elicitation. The positioning in terms of valence of the pictures

selected for the three levels of valence are illustrated in and against all pictures for males and females, respectively.

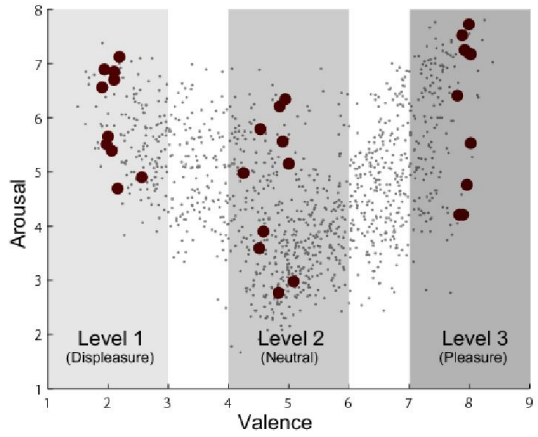

Figure 9. Indication of pictures selected from IAPS for valence stimulation (Males). The red dots represent the pictures selected from IAPS among all pictures, shown as gray dots.

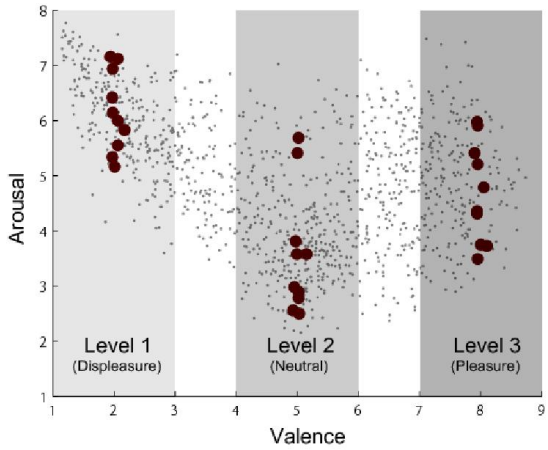

Figure 10. Indication of pictures selected from IAPS for valence stimulation (Females). The red dots represent the selected pictures among all pictures, shown as gray dots.

2.4.2 Experimental protocols

The experiments were designed to stimulate a certain level of valence from viewing multiple pictures from the IAPS database. The subjects were required to sit in a dark room and look at the pictures that appeared on a screen. The protocol of the stimulation procedure used with each subject is illustrated in Figure 11. I collected EEG signals from 50 healthy Japanese subjects (35 males and 15 females). The ages of the subjects ranged from the 20s to the 70s. The EEG signals were recorded by a *Nihon Kohden EEG-1200* using electrodes placed according to the international 10-20 system; the sampling rate of EEG signal acquisition was 1000 Hz. This experiment was conducted with the permission from Research Ethics and Safety committee of The University of Tokyo.

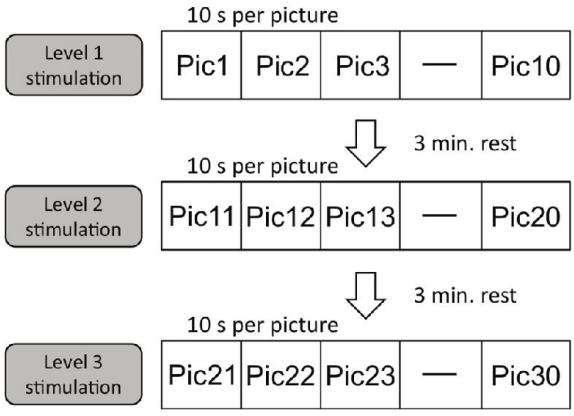

Figure 11. Emotional valence and arousal stimulation procedure.

2.4.3 EEG dataset

EEG signals are affected by noise such as pulses, line noise, and artifacts. The noise generated from eye blinks is the most difficult type to deal with by using signal processing techniques such as wavelet analysis. For this reason, the subjects were asked to refrain from blinking their eyes during the experiments. A filter was applied to delete the line noise at 50 Hz. On the basis of the correlation theory of brain activity asymmetry and emotions [41], [60], [76]–[78]. Moreover, previous research [51] shows that the emotions are not only correlated to the brain activity recorded in the frontal area but also to that recorded in other areas, and therefore, higher accuracy was obtained by using EEG electrodes along all the scalp. We used 16 channels (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, and T6) for analysis based on the International 10-20 system. My approach is to compare the frontal area, the posterior area, and the entire area of brain activity to gain a better understanding of how emotion states can be interpreted using EEG signals in different areas. The test cases using different combinations of EEG electrodes are illustrated in Figure 12.

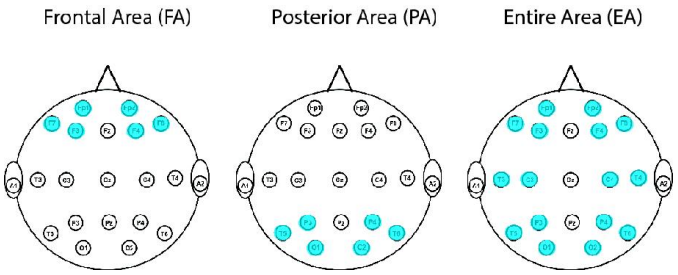

Figure 12. Test cases of brain activity in different areas with their notations.

2.5 RESULTS

2.5.1 Valence recognition results

My goal is to formulate robust features for valence detection using EEG signals, so the discernibility of different valence levels achieved by the proposed features was visualized using PCA. From the PCA results shown in Figure 13, there exists no distinct structures in feature space by using statistical measurements such as the mean, the kurtosis, and the skewness (Figure 13(a), Figure 13(b), and Figure 13(c)) based on raw EEG signals, however, a clearer structure can be observed using the std (standard deviation), although there are also overlaps between different valence levels (Figure 13(d)). The test results from the simplest two-level valence detection show that by using these four statistical measurements, 80% accuracy was obtained by using the std, while near or slightly better than chance accuracy (ranging from 50% to 60%) was obtained by using the other three statistical measurements.

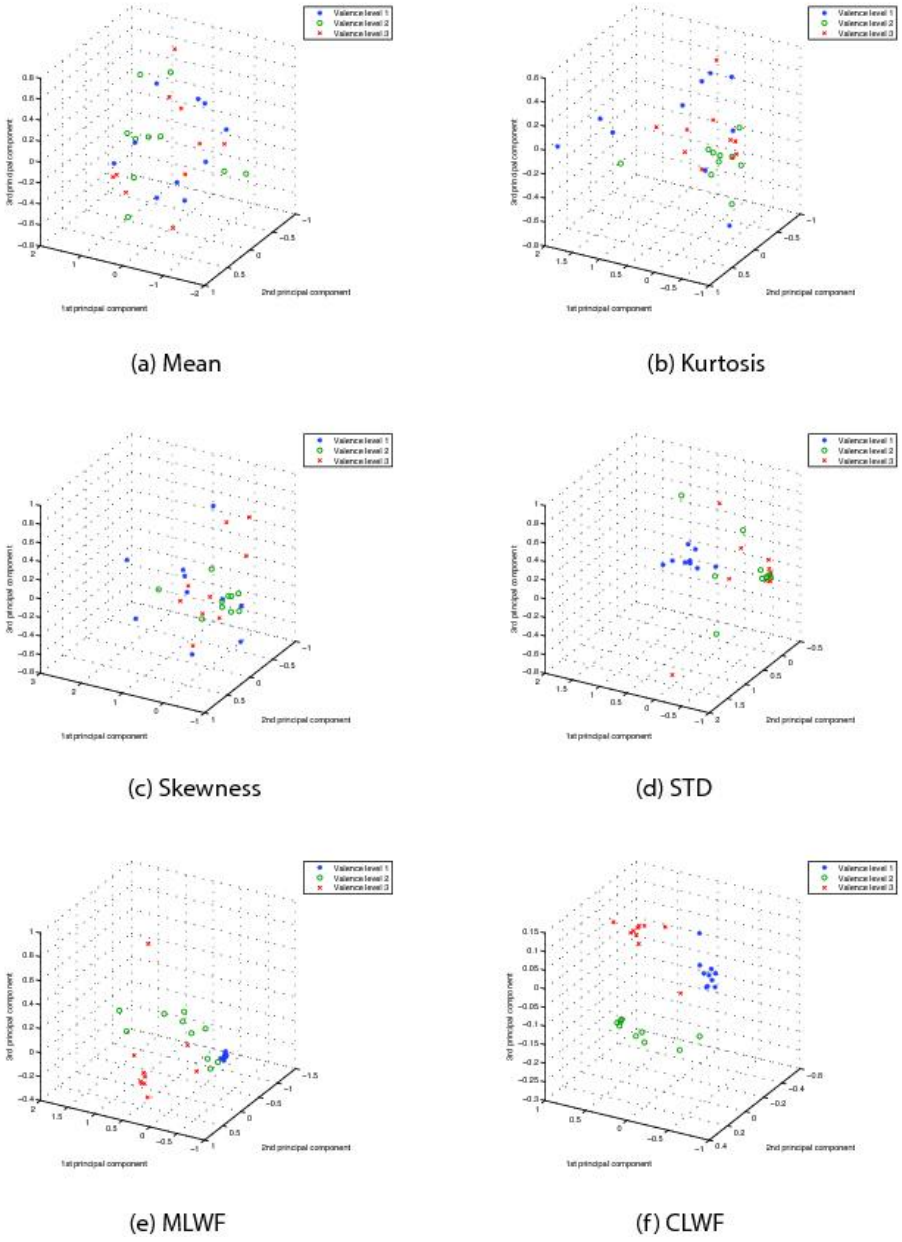

Figure 13. Visualization of feature space using PCA. (a)-(d) were calculated from raw EEG signals, (e) MLWF was the std calculated using the best performance level after wavelet decomposition, and (f) CLWF was the std calculated using the cross-level decomposed signals from discrete wavelet decomposition selected using GA.

Based on these facts, the std was selected for further analysis in order to develop our new strategy for feature extraction. After applying wavelet decomposition, we tested the performance using PNN by LOOCV and selected the optimum performance wavelet decomposed level. Then the features were visualized using PCA (Figure 13(e)), which indicate that the discernibility was further improved compared to the original features extracted based on raw EEG signals. Figure 13(f) illustrates the proposed cross-level wavelet features in this work. Clearer clusters have been obtained by plotting them using the first three principal components, which proves the robustness of the proposed strategy to extract features. Table 3 contains the results of the two and three classes of different valences using the frontal area, posterior area, and entire area of EEG signals.

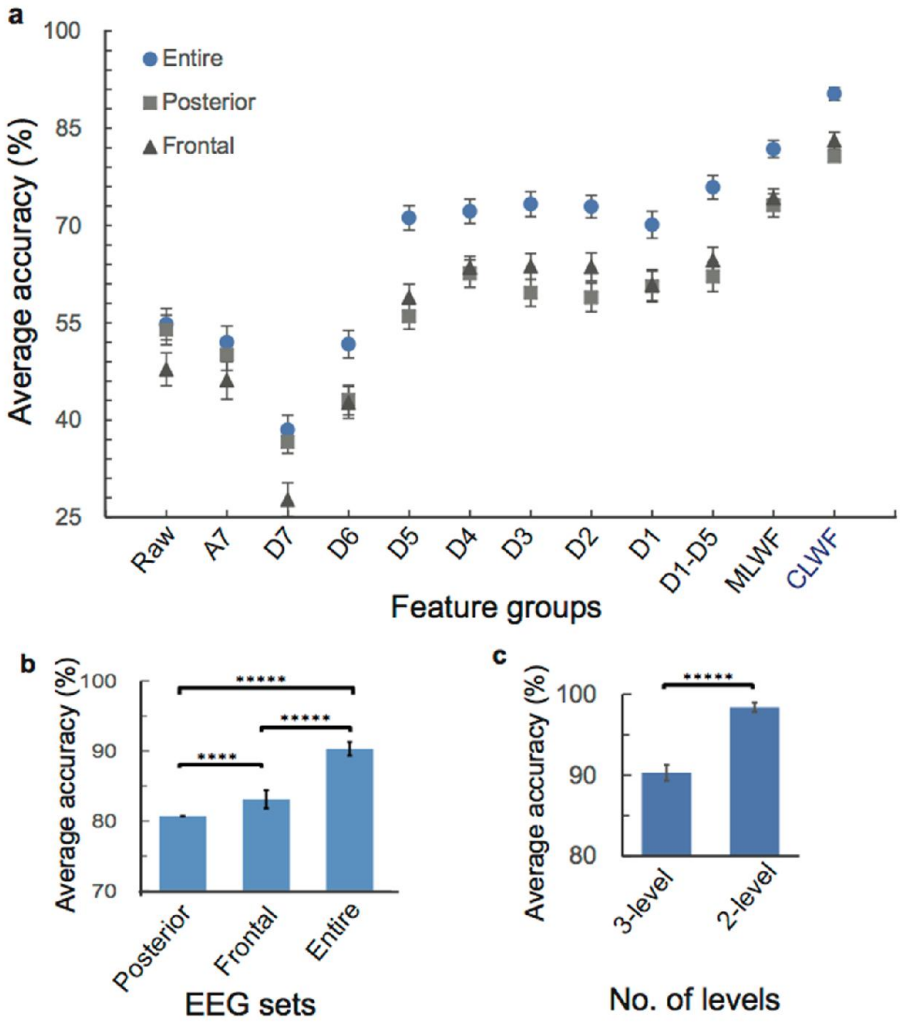

Figure 14. Average results for valence detection on 50 participants using different feature groups and EEG sets. a, Average accuracy for 3-level valence detection using 12 different feature groups (mean ± s.e.m). b, Comparisons of effectiveness for 3-level valence detection using EEG signals in different brain areas by proposed features (CLWF); ****p<0.00001, *****p<0.000001 by anaysis of variance (ANOVA) plus Tukey's Honestly Significant Difference (HSD) test. c, Comparison of 2-level (L1: 98.4% and L3: 97.8% respectively) and 3-level (L1: 94.0%, L2: 86.6%, and L3: 90.2% respectively) overall valence classification accuracy using proposed features (CLWF); *****p<0.000001 by paired t-test.

Figure 14a shows the average accuracy by adopting different groups of features for three valence levels classification using PNN (σ=0.15). To study the results illustrated by Figure 14a, we firstly apply analysis of variance (ANOVA) to check if the means representing average accuracy are unequal. It turns out that at least one mean is different by $p<0.000001$, and then we apply Tukey's Honestly Significant Difference (HSD) test to find which means are unequal. We found that every pair of means including proposed feature group are significantly different by $p<0.000001$. Some pairs of means are not significantly different by the condition of $p>0.05$, they are A7 and D6, Raw and D6, Raw and A7, pairs between D1, D2, D3, D4, and D5.

Table 3. Average results (mean ± std) for two-level and three-level valence detection under different stimulations (50 subjects).

EEG area	Features	two-level accuracy (%)			three-level accuracy (%)			
		L1	L3	Overall	L1	L2	L3	Overall
Frontal area	Raw	80.4 ± 17.7	71.8 ± 22.3	76.1 ± 15.2	65.4 ± 26.2	38.4 ± 28.7	39.6 ± 25.7	47.8 ± 17.9
	A7	73.8 ± 22.0	62.6 ± 29.7	68.2 ± 21.0	58.0 ± 28.1	43.0 ± 30.1	37.4 ± 28.1	46.1 ± 20.4
	D7	60.4 ± 35.0	38.0 ± 28.0	49.2 ± 25.4	39.0 ± 32.8	22.8 ± 25.6	21.4 ± 21.9	27.7 ± 18.5
	D6	66.8 ± 24.9	64.0 ± 25.3	65.4 ± 19.6	53.2 ± 27.4	33.4 ± 26.8	41.4 ± 26.7	42.7 ± 17.2
	D5	81.4 ± 16.9	81.4 ± 17.4	81.4 ± 13.8	66.4 ± 23.9	46.4 ± 26.5	64.0 ± 23.5	58.9 ± 14.6
	D4	85.4 ± 15.9	82.0 ± 15.8	83.7 ± 12.4	73.8 ± 24.2	50.8 ± 25.2	66.0 ± 20.8	63.5 ± 12.5
	D3	87.8 ± 15.0	82.4 ± 15.6	85.1 ± 11.4	75.6 ± 24.2	49.6 ± 27.7	65.8 ± 23.9	63.7 ± 14.1
	D2	88.6 ± 13.6	79.8 ± 17.3	84.2 ± 11.6	76.8 ± 24.4	50.6 ± 25.0	63.4 ± 22.7	63.6 ± 15.3
	D1	85.8 ± 16.8	78.0 ± 17.5	81.9 ± 13.4	74.8 ± 22.8	45.8 ± 27.3	61.8 ± 23.4	60.8 ± 17.1
	D1 - D5	88.0 ± 15.9	83.6 ± 13.8	85.8 ± 11.1	76.6 ± 24.0	50.4 ± 24.1	66.8 ± 20.2	64.6 ± 14.5
	MLWF	94.2 ± 10.1	91.4 ± 11.3	92.8 ± 13.4	79.4 ± 20.0	69.2 ± 20.7	74.0 ± 22.1	74.2 ± 10.3
	CLWF	99.0 ± 3.0	95.6 ± 7.3	**97.3 ± 4.1**	90.4 ± 10.3	74.6 ± 20.9	84.4 ± 13.1	**83.1 ± 9.2**
Posterior area	Raw	73.8 ± 20.6	70.2 ± 19.8	72.0 ± 16.7	63.2 ± 25.2	47.0 ± 21.5	51.4 ± 21.9	53.9 ± 16.1
	A7	67.8 ± 22.2	64.4 ± 23.7	66.1 ± 19.5	59.0 ± 25.1	49.2 ± 21.7	41.8 ± 24.1	50.0 ± 16.8
	D7	55.2 ± 24.5	53.2 ± 23.3	54.2 ± 19.8	41.8 ± 22.7	34.0 ± 18.0	34.0 ± 20.2	36.6 ± 12.2
	D6	66.4 ± 21.1	62.0 ± 19.6	64.2 ± 17.4	52.2 ± 22.6	35.4 ± 20.1	41.6 ± 21.0	43.1 ± 16.3
	D5	78.8 ± 18.9	76.0 ± 17.0	77.4 ± 14.2	65.8 ± 24.0	47.0 ± 25.0	55.2 ± 20.7	56.0 ± 13.9
	D4	83.0 ± 17.3	79.8 ± 18.3	81.4 ± 15.1	70.4 ± 25.5	56.4 ± 25.4	61.0 ± 22.6	62.6 ± 15.0
	D3	80.6 ± 17.5	76.6 ± 21.9	78.6 ± 16.2	70.0 ± 22.6	50.4 ± 25.1	58.4 ± 25.5	59.6 ± 14.7
	D2	82.2 ± 16.8	76.8 ± 21.3	79.5 ± 14.8	69.6 ± 25.1	49.2 ± 23.3	57.8 ± 23.7	58.9 ± 15.5
	D1	82.2 ± 17.5	76.6 ± 21.1	79.4 ± 14.6	71.4 ± 21.1	52.4 ± 26.9	58.0 ± 26.8	60.6 ± 16.9
	D1 - D5	87.0 ± 14.9	80.4 ± 19.4	83.7 ± 13.5	72.0 ± 24.7	52.0 ± 24.7	62.2 ± 24.7	62.1 ± 16.1
	MLWF	89.4 ± 11.3	89.0 ± 10.5	89.2 ± 8.4	78.4 ± 17.1	67.8 ± 22.3	73.2 ± 20.3	73.1 ± 12.5
	CLWF	96.4 ± 6.3	93.8 ± 9.9	**95.1 ± 6.5**	85.2 ± 17.1	77.4 ± 17.6	79.4 ± 16.8	**80.7 ± 11.0**
Entire area	Raw	83.2 ± 15.9	76.4 ± 23.5	79.8 ± 15.0	71.2 ± 26.4	45.2 ± 25.8	48.0 ± 17.1	54.8 ± 17.1
	A7	77.8 ± 16.7	69.4 ± 28.1	73.6 ± 17.0	64.4 ± 25.6	48.4 ± 27.2	43.2 ± 25.6	52.0 ± 17.8
	D7	68.4 ± 27.0	55.2 ± 28.4	61.8 ± 19.3	51.6 ± 30.3	28.6 ± 26.0	35.2 ± 25.2	38.5 ± 16.0
	D6	73.2 ± 19.7	71.2 ± 21.8	72.2 ± 15.6	62.0 ± 24.4	44.0 ± 23.0	49.2 ± 21.8	51.7 ± 15.0
	D5	91.0 ± 9.5	91.8 ± 10.4	91.4 ± 8.1	79.0 ± 18.5	59.2 ± 24.3	75.4 ± 18.4	71.2 ± 13.3
	D4	91.0 ± 12.8	87.8 ± 12.8	89.4 ± 11.0	81.4 ± 20.9	62.6 ± 22.9	72.6 ± 17.9	72.2 ± 13.4
	D3	93.2 ± 9.8	88.8 ± 12.7	91.0 ± 9.3	83.6 ± 19.4	63.0 ± 23.6	73.2 ± 21.0	73.3 ± 13.8
	D2	93.6 ± 9.4	86.4 ± 14.1	90.0 ± 9.8	83.2 ± 16.2	63.4 ± 20.6	72.0 ± 20.1	72.9 ± 12.3
	D1	92.8 ± 9.5	84.6 ± 15.8	88.7 ± 10.1	81.6 ± 18.3	60.6 ± 22.0	70.8 ± 14.7	70.1 ± 14.7
	D1 - D5	94.4 ± 8.8	90.8 ± 12.4	92.6 ± 8.3	84.8 ± 18.1	67.0 ± 21.7	75.8 ± 20.5	75.9 ± 12.9
	MLWF	97.4 ± 5.3	94.8 ± 7.1	96.1 ± 5.2	85.4 ± 16.8	78.4 ± 14.8	81.6 ± 16.1	81.8 ± 9.4
	CLWF	98.4 ± 4.2	97.8 ± 4.6	**98.1 ± 3.3**	94.0 ± 7.8	86.6 ± 13.8	90.2 ± 8.9	**90.3 ± 6.9**

Different statistical parameters are tested and the most effective one is selected above, I further propose a strategy that dynamically selects statistical parameters for wavelet coefficients of EEG signals. An automatic statistical parameter selection method by prior identifying the signal data distribution by Lilliefors tests [79]. Statistical parameters such as the mean and the std are selected if the signal data obeying Gaussian distribution, and the skewness and the kurtosis are used elsewise. The accuracy can be improved from 90% to 93% for 3-level valence recognition. The results are shown in Figure 15.

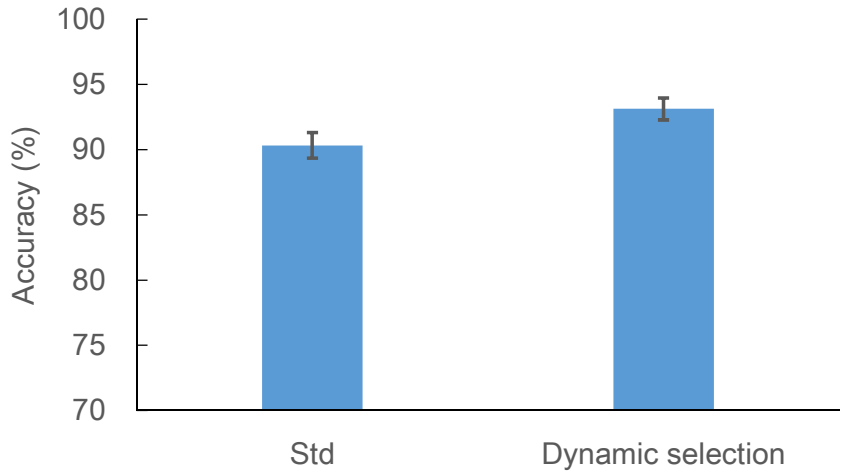

Figure 15. Results of comparing valence recognition performance with std using dynamic statistical parameter selection.

2.5.2 Comparisons

For comparison study, I compared with a different classifiers (SVM) as well as different features selection methods. The results illustrating the comparison with

SVM using the std are shown in Figure 16. The SVM performance has potential to be further improved by optimization of parameters, limited parameters are tested in this research to only show the robustness of proposed features.

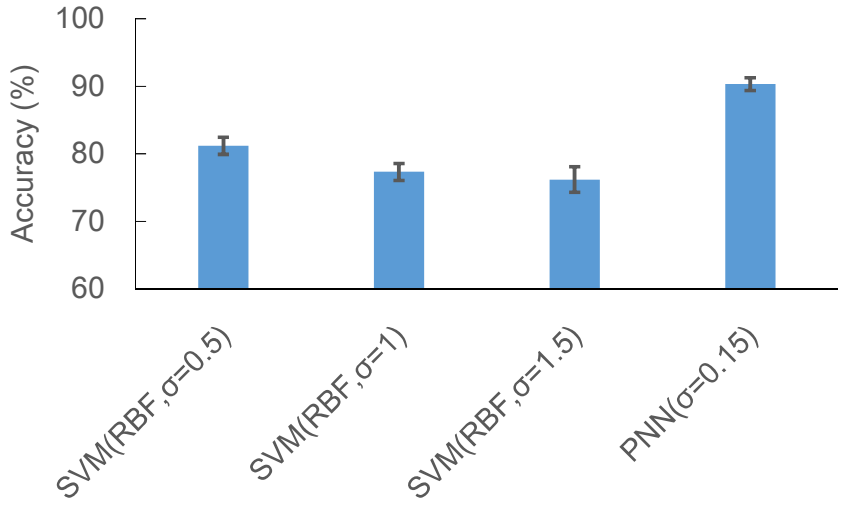

Figure 16. Comparison results of valence recognition with std using SVM and PNN.

The feature selection methods for comparisons are listed in Table 4 and the comparison results for 3-level valence recognition are illustrated in Figure 17.

Table 4. Feature selection methods for control study

Criterion	Full name	Authors (years)
MIFS	Mutual Information Feature Selection	Battiti (1994)
MRMR	Max-Relevance Min-Redundancy	Peng et al. (2005)
ICAP	Interaction Capping	Jakulin (2005)
RELIEF	Relief feature selection	Kira et al. (1992)

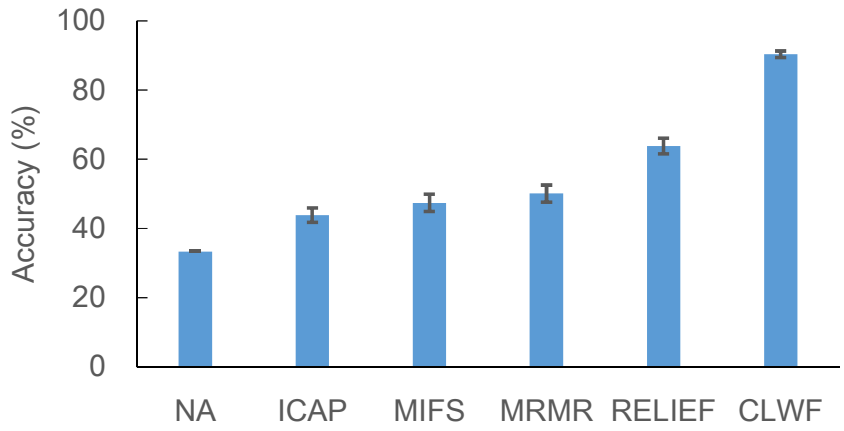

Figure 17. Comparisons of different feature selection methods.

2.5.3 Study on the effectiveness of EEG electrodes

Different sets of EEG electrodes are explored in this study to further understand the functional area for emotional activities, and reduce the number of EEG electrodes for the convenience of users and computational simplicity. Besides two comparison sets of frontal and posterior areas, the same GA approach is performed to find another optimized set of EEG electrodes placement for each subject. The input for GA is a matrix with n levels of wavelet decomposition for all EEG electrodes; the GA will select less than a certain number of electrodes for the reduced set. Multiple electrodes are needed to understand the difference of brain activities in different brain areas, I consider 6 electrodes are appropriate based on preliminary attempts and for comparisons with the other two sets using 6 electrodes mentioned in Figure 12.

Electrodes selected by GA for each subject are illustrated in a in which a blue dot indicates “a selected electrode” and the results using the reduced set of EEG electrodes for 3-level emotional valence detection are illustrated in Figure 18b for comparisons with other set of electrodes.

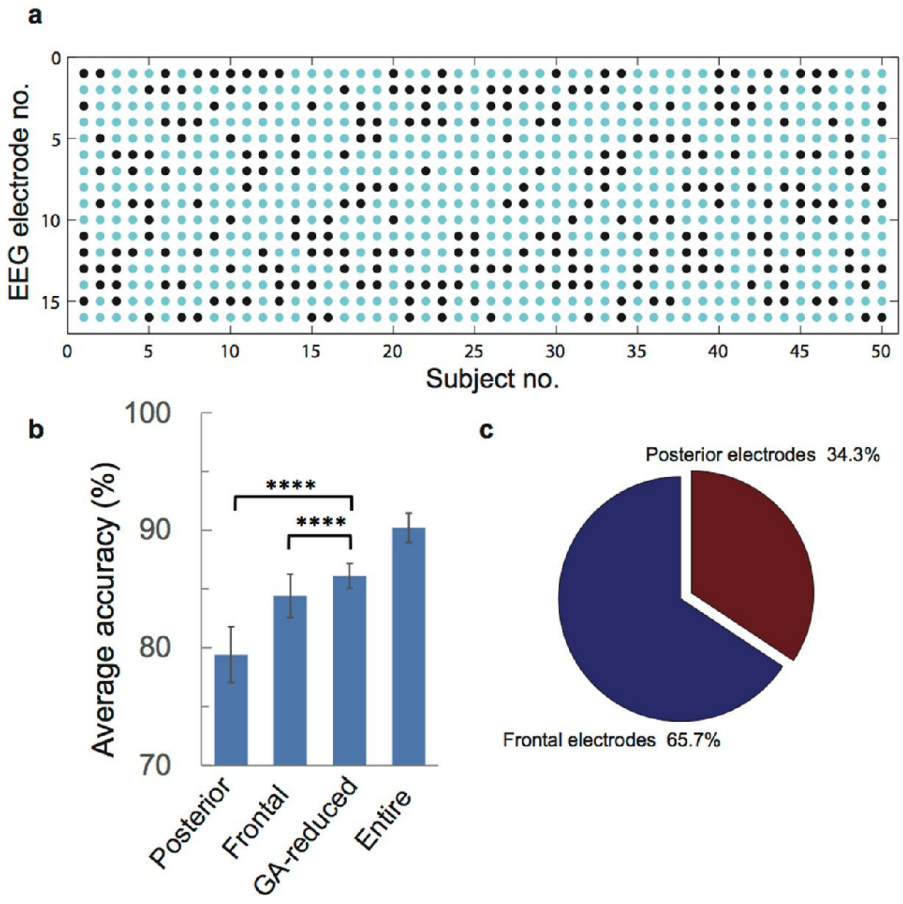

Figure 18. GA based EEG electrodes reduction. a, GA-reduced EEG electrodes distribution; no. 1-16: Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6. b, Comparison results of GA-reduced EEG sets; ****p<0.00001 by ANOVA plus Tukey's HSD throughout the figure. c, Selected electrodes' brain area summary.

2.6 DISCUSSION

Generally, the EEG signals from the frontal area are more effective for valence detection than the ones from the posterior area. However, just as in the research results demonstrated by Murugappan et al. [51], [68]. Our classification results also indicate that the accuracy can be improved by covering the entire area with more EEG electrodes. This finding is also supported by the emotion theory proposed by Heller, which argues that the frontal and parieto-temporal regions are involved in emotion [80]. The features extracted from raw EEG signals are not robust enough to get the best results, in contrast to the features extracted from decomposed signals by discrete wavelet decomposition. The overall results obtained using the 50 healthy volunteers who were presented with sets of elicitation pictures from IAPS indicate that the D2, D3, and D4 levels of the wavelet coefficients provide more information for the classifier compared to other levels, which represent the information in the frequency bands of gamma and lambda. Previously studies also support our findings by showing the effectiveness of the gamma band EEG for recognizing emotion activity. The gamma rhythm is widespread in areas associated with emotional processing [81], and the gamma band was further reported to have connections to emotions with a special emphasis on negative emotional processing [82], [83]. The gamma band power has been shown to decrease during periods of processing or imagining

negative emotional material [84]. Muller et al. also suggested that a significant valence due to hemisphere interaction emerged in the gamma band [85]. By contrast, very little research has been done on emotion analysis using the lambda band.

My proposed cross-level wavelet feature group (CLWF) that is searched from decomposed wavelet coefficients from each EEG electrode based on GA. It can search useful information related to different EEG bands and lead to a largely improvement of the classification performance to an accuracy of 98% with two-level and 90% with three-level valence detection. Figure 14c illustrates the comparison of 2-level and 3-level valence detection using proposed features with $p<0.000001$ by paired t-test. In contrast, the accuracy can be improved very much comparing to simply combining all the levels from D1-D5 shown in Figure 14a. The results demonstrate the importance and practicability of the cross-level strategy for extracting features in the time-frequency domain for valence level detection using EEG signals. A possible explanation for this is that the affective information that describes emotions is represented in different frequency ranges within different brain areas, although this issue is still unresolved because of the uncertainty of EEG signal interpretations. The results in this work demonstrate the importance and practicability of the cross-level strategy for extracting features in the time-frequency domain for valence level detection using EEG

signals, although the strategy does not provide a concrete interpretation of physiological meaning. This is because of the variation in the selected levels of wavelet coefficients for each EEG electrode that are used for feature extraction among different subjects. However, there are several points supporting the effectiveness of our proposal. Our analytical results show that the frontal area are more robust in recognizing emotions. In addition, It provide higher accuracy by using EEG signals collected from the entire area, which is supported by an important statement in Borod's emotion model [86], in which emotions are represented in cortico-limbic networks rather than in particular areas of the brain. Such a network produces spread, rather than focal cortical activity.

As we can see in Figure 18, a better accuracy (86.1%) can be achieved using the reduced set of EEG electrodes compared with frontal and posterior area EEG electrodes. On the other hand, as far as we know, the deep limbic system (DLS) plays a major role in a person's emotion states, and the prefrontal cortex (PFC) has functions to control emotion. However, there is no precise correspondence between these findings and the patterns in EEG signals reported in the literature. In our study, the reduced EEG set is selected based on our proposal of the wavelet coefficient selection from all EEG electrodes, the statistical summarization of the electrodes selected from 50 subjects shows that more frontal area electrodes (65.7%) are selected compared with posterior area

electrodes (34.3%), which is also consistent with the viewpoint that frontal area brain activities can better demonstrate emotion [80]. The analytical results also indicate the effectiveness of our proposed method.

2.7 APPLICATION ON EMOTIONAL AROUSAL DETECTION

The same method developed for emotional valence detection is applied on arousal detection. By using emotional valence and arousal, detailed emotion can be defined. Selected pictures for evoking arousal are illustrated in Figure 19 and Figure 20.

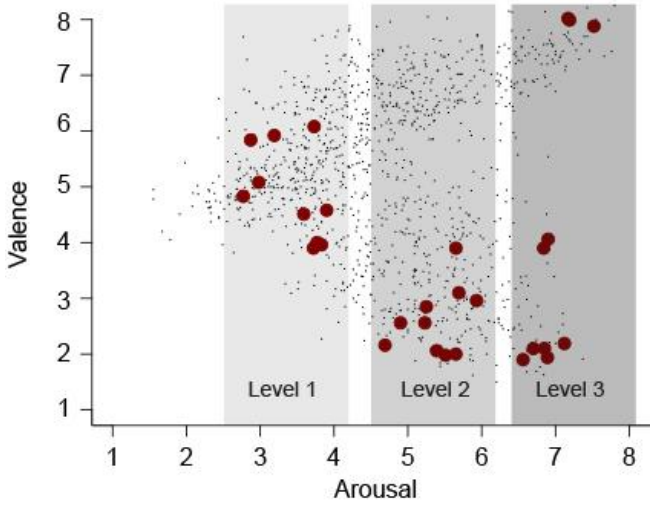

Figure 19. Indication of pictures selected from IAPS for arousal stimulation (Males). The red dots represent the pictures selected from IAPS among all pictures, shown as gray dots.

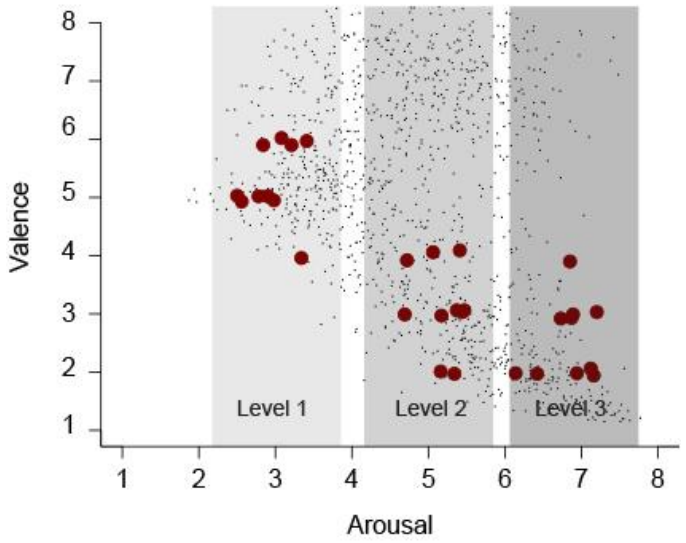

Figure 20. Indication of pictures selected from IAPS for arousal stimulation (Females). The red dots represent the selected pictures among all pictures, shown as gray dots.

The results for 2 and 3 level arousal recognition calculated using the same proposed method as for valence recognition are illustrated in Figure 21.

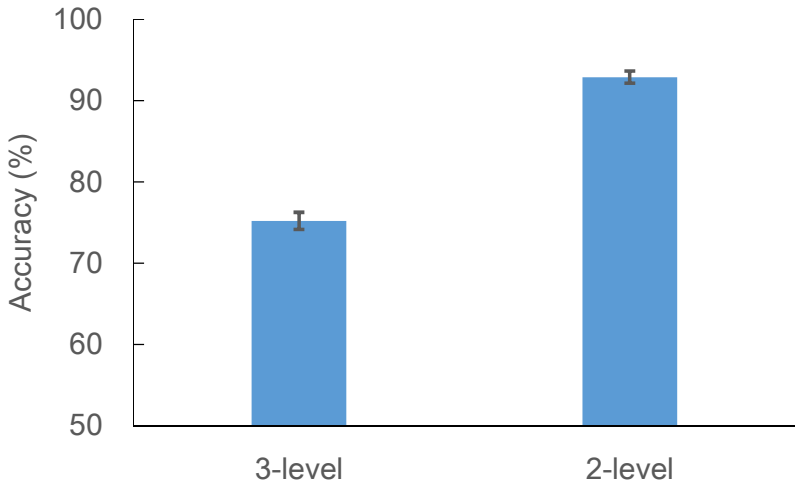

Figure 21. Results of arousal recognition.

2.8 SUMMARY

I proposed a new strategy to extract time-frequency domain features from EEG signals in cross levels of wavelet decomposition coefficients with different EEG electrodes for valence level detection. The proposed features (CLWF) substantially increase the accuracy compared to conventional features extracted directly from EEG signals or from transformed signals in the time-frequency domain. 93% accuracy for three-level valence detection and 99% accuracy for two-level valence detection are achieved by using the proposed features.

The results show the importance of taking into consideration information in different frequency bands with different EEG electrodes, as it results in higher accuracy than that achieved when only considering a single level or combined levels of wavelet decomposed signals from different EEG electrodes.

The results achieved in this research would be interests of practitioners in a number of related fields such as health informatics and BCI. This work gives hints of many paths for future model development.

3 EMOTIONAL SPEECH DATABASE CONSTRUCTION

In this chapter, a Japanese emotional database is constructed that contains speech and physiological signals that can be used to develop algorithms for emotion recognition using audio, physiological signals, or several combined signals.

3.1 RESEARCH BACKGROUND AND REMAINING ISSUES

According to different research purposes, lots of emotional databases have been established [87]–[93]. The most important discussion issues for building an emotional database are demonstrated as follows.

Firstly, it usually has two different patterns of the database design, which are real-world emotions or acted emotions. They are indeed different concerning many aspects. William and Stevens states that acted emotions tend to be more exaggerated than real ones.

Secondly, how the speech signals are simulated? The most databases of emotional speech are not naturally recorded in daily life and not obtained during a conversation. Thus, how to get speech signals under nature emotions is very important for an emotional database. As for the database design, one strategy is to use emotional speech from experienced actors act as if they are in a specific emotional state. Another strategy is to induce an actor or ordinary person to a

certain emotion state using stimuli such as pictures, videos, computer games, etc. The latter one is known as more similar as real-world emotions.

Thirdly, whether the database emphasizes the speaking contents? Researchers sometimes regulate the speaking contents in order to avoid the perceived emotion influenced by its lexical content. Some corpus is designed to include same number of utterances for each emotion. Both strategies must predefine the contents that examinee speaks, which are the situations will not happen in a real world.

Currently most emotional speech databases are not accessible to the public. Mainstream databases are summarized in Table 5.

Table 5. Summary of existing emotion databases (DB).

DB name	Language	Data	Source	Labels
Berlin emotional database	German	800 utterances	Professional actors	Angry, joy, fear, sadness, disgust, boredom, neutral
Danish emotional database	Danish	4 participants read 2 words, 9 sentences, and 2 passages in 5 emotions	Nonprofessional actors	Angry, joy, sadness, surprise, neutral
KES	Korean	5400 utterances	Nonprofessional actors	Neutral, joy, sadness, anger
INTERFACE	English, Slovenian, Spanish, French	186 utterances in English, 190 utterances in Slovenian, 184 utterances in Spanish, 175 utterances in French	Actors	Anger, disgust, fear, joy, surprise, sadness, slow neutral, fast neutral
TALKAPILLAR	French	78 utterances	Actor	Angry, joy, fear, sadness, disgust, boredom, neutral
Hao Hu et al.	Chinese	1600 utterances	Nonprofessional actors	Anger, fear, joy, sadness, neutral
Amir et al.	Hebrew	60 Hebrew and 1 Russian actors	Nonprofessional actors	Anger, fear, joy, disgust, sadness, neutral
MPEG-4	English	2440 utterances	Movies	Joy, anger, fear, disgust, sadness, surprise, neutral

3.2 PROCEDURE OF EXPERIMENTS

The experiments consisted of two parts, which were an online survey and onsite experiments. The Internet survey was designed to collect materials representing participants' real emotional experiences. After the materials for emotion elicitation were collected, the onsite experiments were arranged to collect the speech and physiological signals.

3.2.1 Online survey

Basic information such as that on gender and age ranges was collected from an online survey. Simple questions to collect information on participants' real

emotional experiences were asked in forms such as “Please explain one or two memories that aroused your deepest emotions of happiness”.

3.2.2 Onsite experiment

There is a photograph of the onsite environment setting for an experiment in Figure 22. An assistant introduced the experimental protocols, how the sensors were worn, and checked the sensor signals for participants, while a coordinator helped to elicit their emotions.

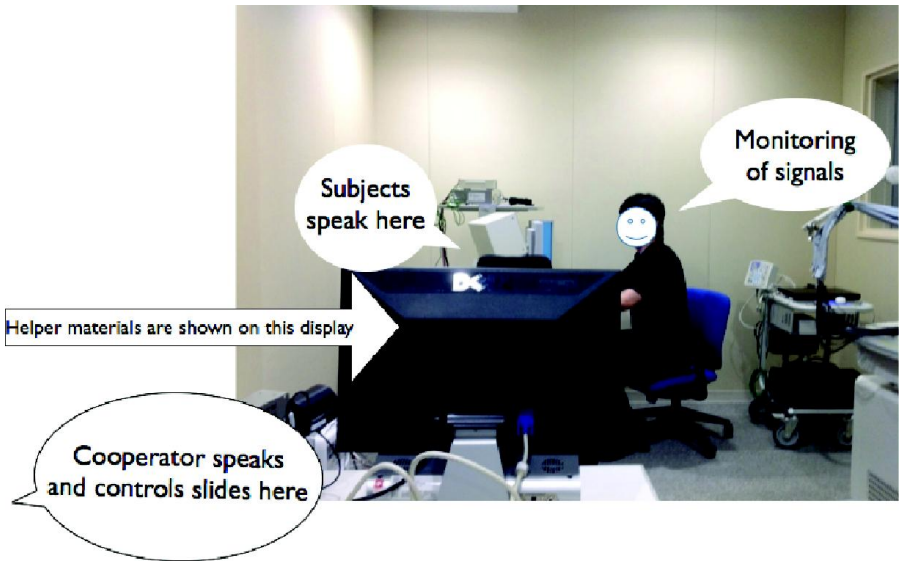

Figure 22. Environment setting for experiment from viewpoint of coordinator.

The participants recalled their emotional experiences and described them during the experiment, and the coordinator asked questions and made small talk about the same emotions with prior knowledge from the survey that they had

previously completed. The procedure for eliciting the six emotions is outlined in Figure 23.

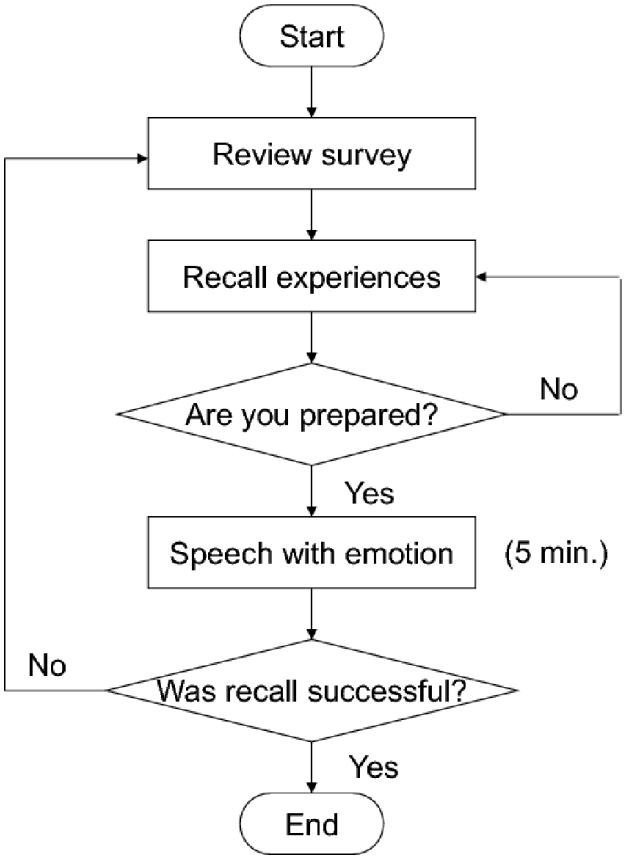

Figure 23. Procedure for eliciting emotions by recalling experiences.

Eight signals were collected in the experiments including those from electroencephalography (EEG), speech, electrocardiography (ECG), electromyography (EMG) skin temperature, respiration, blood volume pulse, and skin conductance. Where the sensors were worn and what signals were collected

are illustrated in the photographs in Figure 24. Finally, the participants completed a five-point Likert scale for self-assessment.

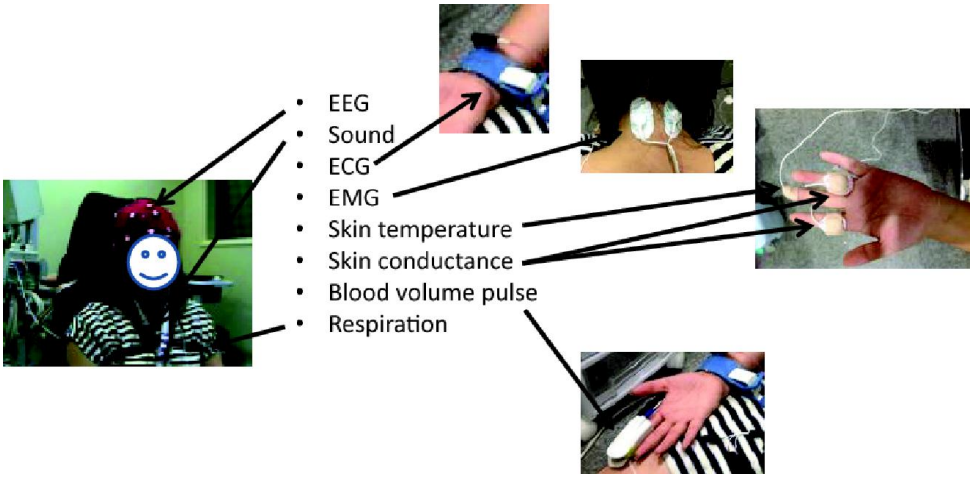

Figure 24. Placement of sensors and signals that were collected.

3.3 DESCRIPTION OF SIGNALS

How signals were collected and explained in the following. Examples are also given. The EEG and ECG signals are collected using Nihon Kohden EEG-1200. Physiological signals such as respiration, EMG, blood volume pulse, skin temperature signals are collected by Bioplux (Figure 25).

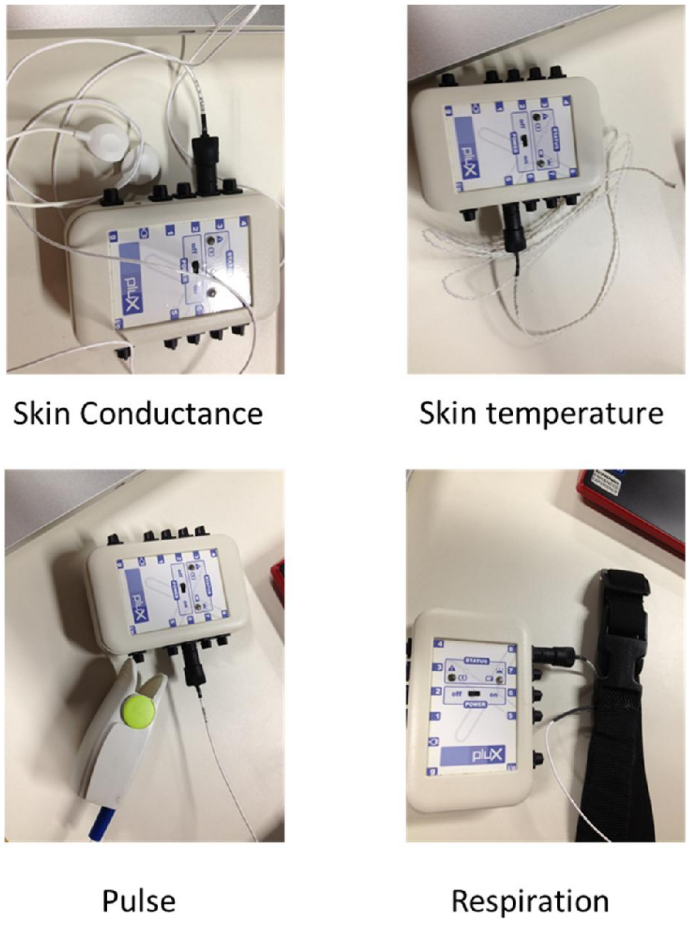

Figure 25. Image of bioplux.

Detailed information of collected signals is demonstrated as follows.

Speech signals. Speech signals are very popular in emotion research since they are easy to obtain in daily life for making applications. The speech signals were recorded during participants' recalling real-world emotional experiences. I collected long fragments of speech signals for the six emotions (Figure 26) in this database.

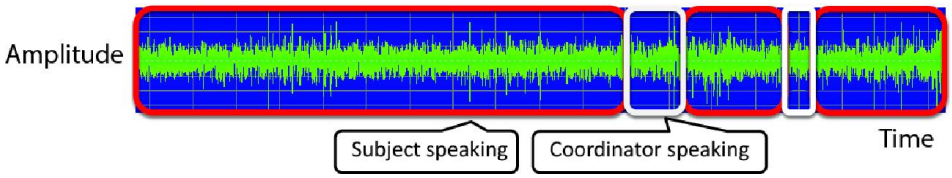

Figure 26. Example of collected speech signals.

EEG signals. EEG measures voltage fluctuations resulting from ionic current flows within the neurons of the brain. Much research [94] has revealed that there is a relationship between EEG signals and different kinds of emotions and it is advantageous to use these as it is difficult for people to manipulate EEG signals. Figure 27 illustrates the positions at which the EEG signals were collected according to 10-20 International system and provides examples of collected signals. The reference electrode was A1.

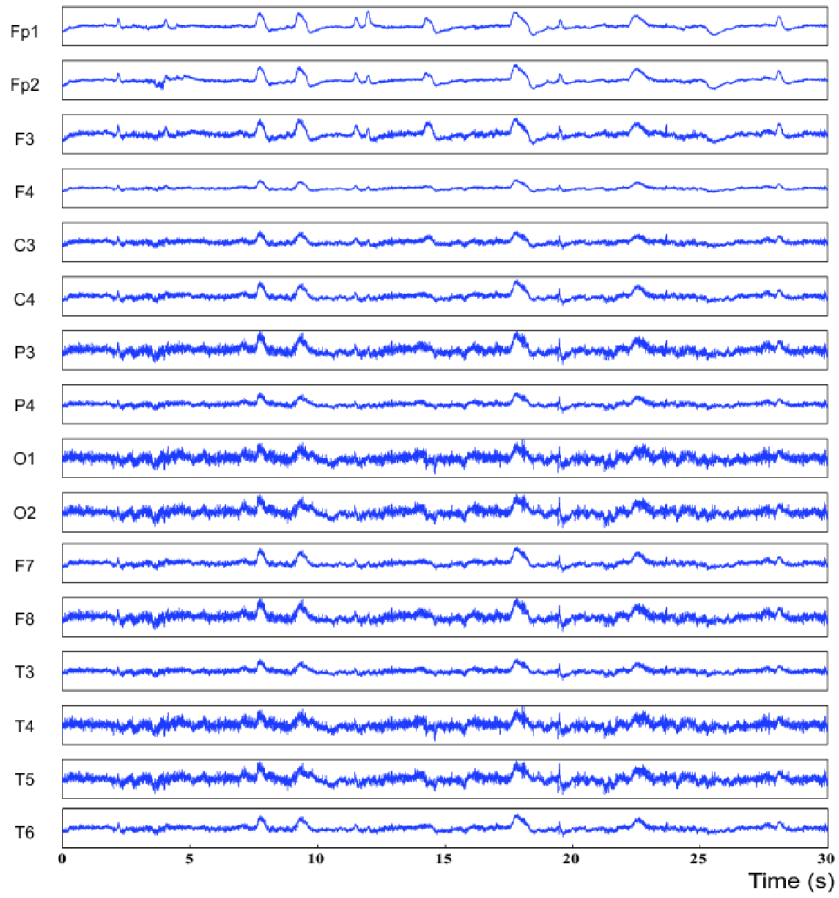

Figure 27. Example of collected EEG signals.

ECG signals. ECG is used to measure the electrical activities of the heart. QRS positions and other features have been reported to have a correlation to emotions [95]. ECG was also used with other signals such as speech for improving emotion recognition performance. Figure 28 has an example ECG with QRS positions.

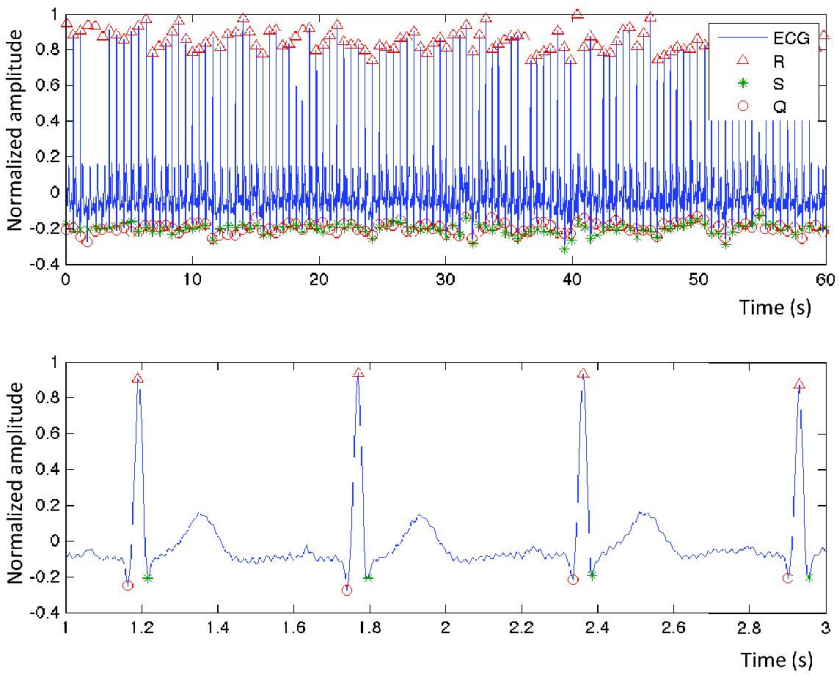

Figure 28. ECG signal with QRS positions.

Respiration signals. Respiration signals record the activity of the lungs. Different respiration patterns also provide emotion information. It's usually used together with other physiological signals for emotion recognition [96]. A respiration signal was recorded with a belt-type sensor, as shown in Figure 29.

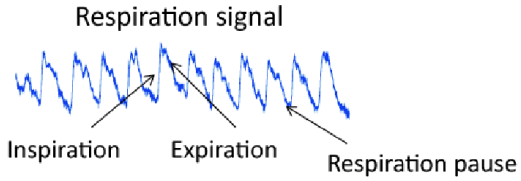

Figure 29. Respiration signal with explanations.

Electromyography (EMG) signals. EMG is a technique for evaluating and recording the electrical activity produced by skeletal muscles. Research has indicated the frequency of muscle tension, action potential amplitude, and the duration of action potential have a relationship with emotions [97]. A sample of collected EMG signal is shown in Figure 30.

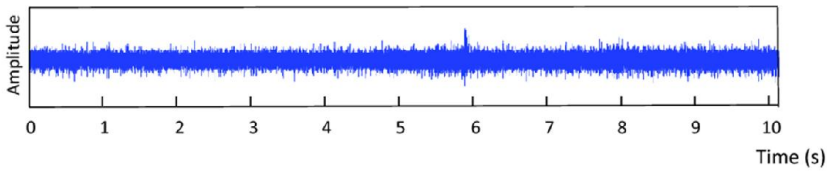

Figure 30. EMG signal illustration.

Skin temperature signals. The literature has indicated that skin temperature is dependent on the emotional state [98]. We measured skin temperature at the finger tips.

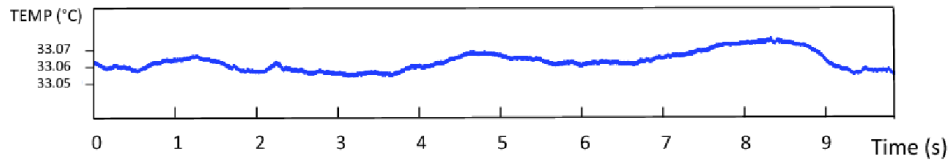

Figure 31. Skin temperature signal.

Blood volume pulse (BVP) signals. Photoplethysmography (PPG) is used to bounce infrared light against the skin surface and measure the amount of reflected light. The literature has indicated that high values for blood volume pulse represent anger and stress, while low values represent happiness and

relaxation [96]. The signal from a blood volume pulse from a finger tip is given in Figure 32.

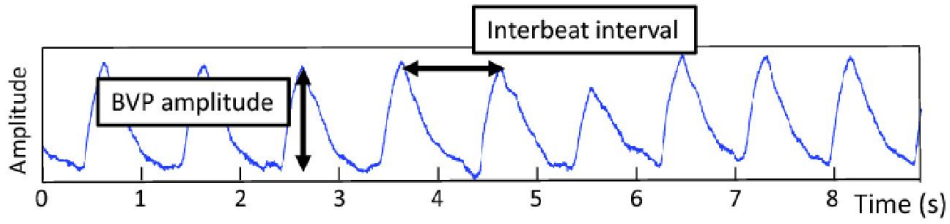

Figure 32. Signal from blood volume pulse.

3.4 DESCRIPTION OF DATA

Speech and physiological signals from fifty healthy Japanese participants were successfully collected. EEG and ECG signals were recorded by a *Nihon kohden EEG-1200* using electrodes placed according to the international 10-20 system; other physiological signals were recorded using Bioplux. This experiment was conducted with the permission from Research Ethics and Safety committee of The University of Tokyo.

Figure 33 has pie charts of the age and gender distributions of participants in the experiments. Most of the participants were in their 20s and 30s and 70% were male and 30% were female.

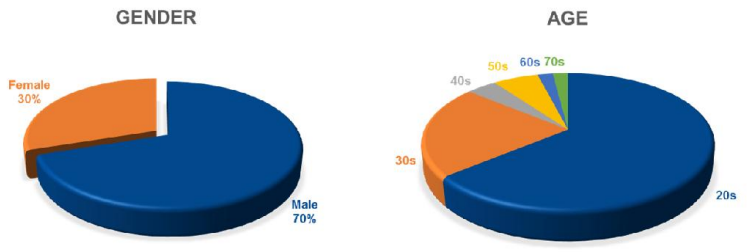

Figure 33. Data distribution related to gender and age.

A self-assessment survey was administered immediately after each experiment. The question of "Did you successfully arouse the emotion of happiness?" was asked after each emotion stimulation. Then, five levels of confidence could be selected as answers, where level 1 (L1) represented the lowest confidence level of a participant's assessment and level 5 (L5) represented the highest confidence level. Figure 34, Figure 35, and Figure 36 plot the answers from the participants.

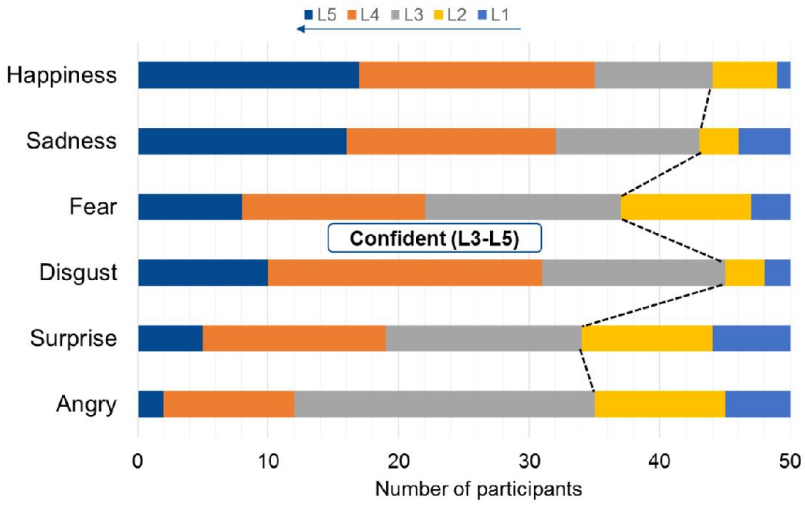

Figure 34. Summary of self-assessments.

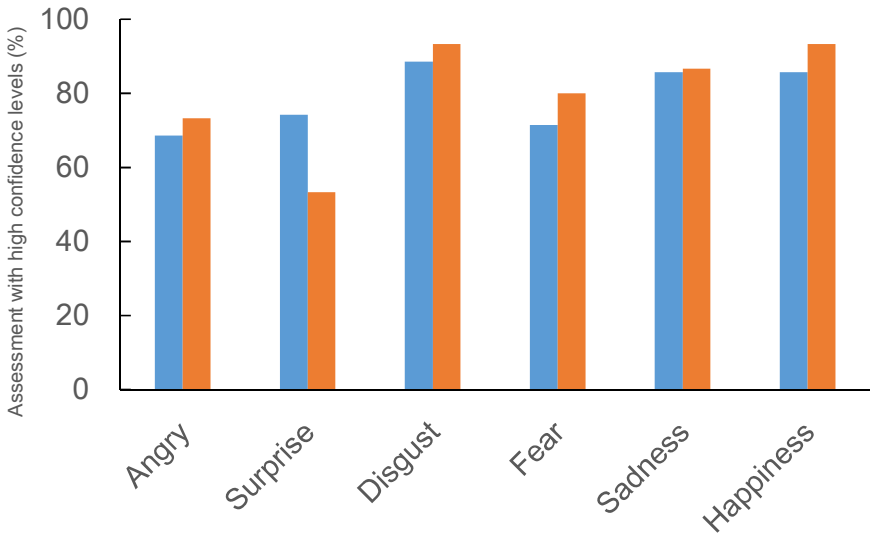

Figure 35. Summary of self-assessment results according to gender. Data (y-axis) indicate percentage of confidence levels no less than three.

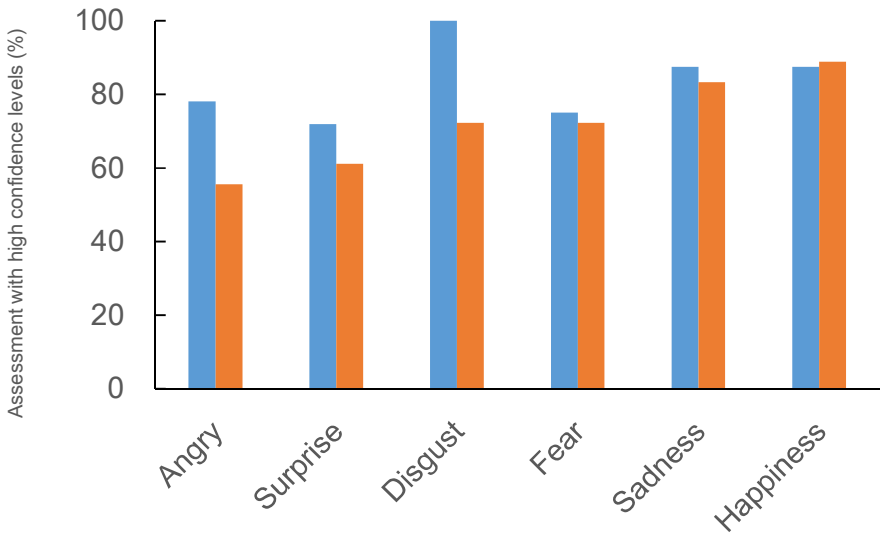

Figure 36. Summary of self-assessment results according to two age groups of less and greater than 30. Data (y-axis) indicate percentage of confidence levels no less than three.

I could easily see from the self-assessment result of Figure 34 that the majority of participants (80%) were confident emotions were elicited with a confidence level of not less than three. Japanese participants were more confident about their emotional arousal of happiness and sadness out of all six emotions, and very confident of their emotional arousal of fear and disgust. They found it relatively more difficult to reach the emotion of anger than the other emotions, and surprise was difficult to arouse during the experiments. In Figure 35, female participants were generally more sensitive to emotions such as happiness, fear, disgust, and anger, while male participants were more confident of surprising experiences. Male participants were less confident about fear experiences than female participants as expected since males are usually stronger and more difficult to scare. However, it seems that male participants were more confident of surprising experiences. This seems difficult to understand at first glance, but it might be due to careful preparations of surprising scenarios arranged by their families. A fact supporting this hypothesis was that most surprising situations were arranged by females based on our survey. Another phenomenon was that more female participants were confident of their anger experiences than male participants. This could have been caused by fewer ways for them to release their emotion of anger compared to male participants. We divided the participants into two groups of different ages with ages younger and older than 30. We found that

younger people had more confidence about the majority of emotions according to statistical analysis (Figure 36), especially stronger emotions such as disgust, anger, and surprise. However, older participants retained more happy experiences and had lower confidence with strong negative emotions and surprising experiences.

I introduced a Japanese database with signals involving six basic human emotions elicited by real experiences to develop an algorithm for health-care oriented applications. Speech signals and a variety of physiological signals were included in the database. Since other people's assessments create errors in real emotion targets, I only carried out self-assessments and analyzed the results.

Self-assessments are very reliable for labeling data on real emotions. In the following sections, other than subjective methods of evaluation such as self-assessments, and further objective methods of selecting high quality speech signals will be proposed to obtain a more reliable database, other physiological signals are also included for further researches.

3.5 SPEECH DATA SELECTION

In order to include high quality speech data in the database, additional data selection procedure is added besides self-assessment. By reviewing emotion theories, no agreement of schematic for emotions was reached. However,

emotional experiences happen at the same time with other bodily changes, which means some clues for emotional activities can be found by bodily changes. The latest emotion theories claim brain is the center for emotional experiences and physiological signals are also highly evolved with emotional experiences. With the development of emotion theories, more researches consider the emotional experiences are involving cognitive evaluation. So there exist conflicts of the relationship between physiological responses and emotional experiences since different emotions might be experienced by different cognitive evaluations with the same physiological responses. In the sense, it's more reliable to identify whether certain kind of emotion is indeed experienced by using brain activities. There are multiple technologies developed for studying human brain activities including EEG, functional magnetic resonance imaging (fMRI) and near-infrared spectroscopy (NIRS). The pros and cons are demonstrated in Table 6.

Table 6. The advantages and disadvantages of different brain activity sensing technologies.

Technique	Advantages	Disadvantages
EEG	• Non-invasive • Inexpensive • High sensitivity	• Graphs brainwaves – no images • Poor resolution • Low signal-to-noise ratio
fMRI	• Non-invasive • Excellent resolution • Shows brain activity • Multidirectional	• Lengthy procedure • Not for those with mental implants • Large noise
NIRS	• Non-invasive	• Technology still in development

fMRI provides more information of brain activities, but it's not suitable for the situation of speech signal collection due to its large noise. And NIRS is a relatively new technology that is still under development. Moreover, it usually detects only the frontal brain activities. The EEG signals are adopted for assessing the emotional status of the participants.

3.5.1 Strategy for speech data selection

In the database, the labels of emotional data with real experience recalling are discrete. For evoking different emotional states of participants, IAPS database is adopted for EEG signals collecting as described in Chapter 2. It's agreed that discrete emotions can be represented using dimensional model with axis of valence and arousal. In the following, literatures are reviewed for defining

different discrete emotions on valence and arousal space [99]–[104]. There exists common agreement on the positions of happiness, anger, and sadness. However, researchers have different opinions on the position of surprise. Lots of researchers consider it has positive arousal. In this research, it is assumed that surprise is in the position of positive valence and negative arousal. Arousing emotions are often described as 'hot' or 'warm', higher arousal makes people want to talk and communicate more; and not arousing ones as 'cold' or 'cool', lower arousal makes people silence. When surprising events happen, people's reactions are often silence and holding the breath. The defined positions are shown in Table 7.

Table 7. Representation of emotions in dimensional space.

Emotion	Position (Valence)	Position (Arousal)
Happiness	Positive	Positive
Anger	Negative	Positive
Sadness	Negative	Negative
Surprise	Positive	Negative

The procedure for EEG assessment is illustrated in Figure 37.

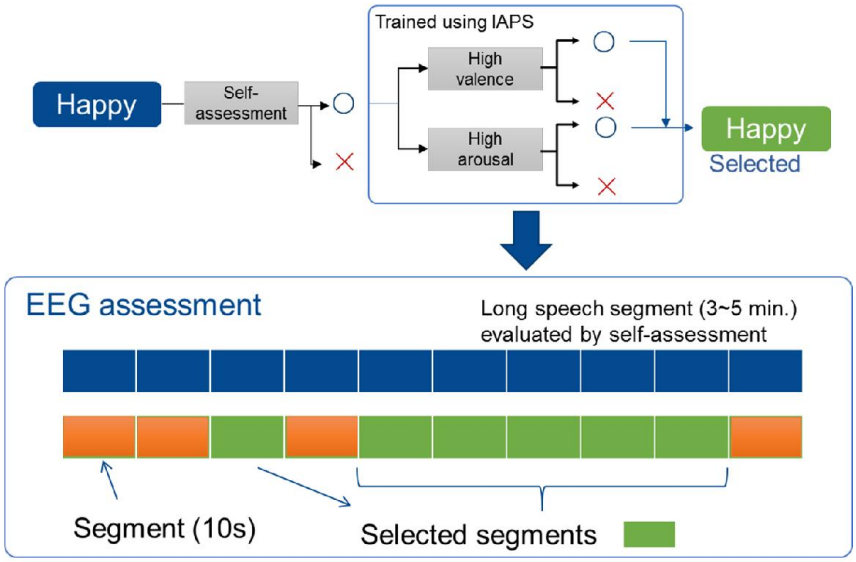

Figure 37. Illustration of data selection.

3.5.2 Results

The selected number of samples is illustrated in the following Figure 38.

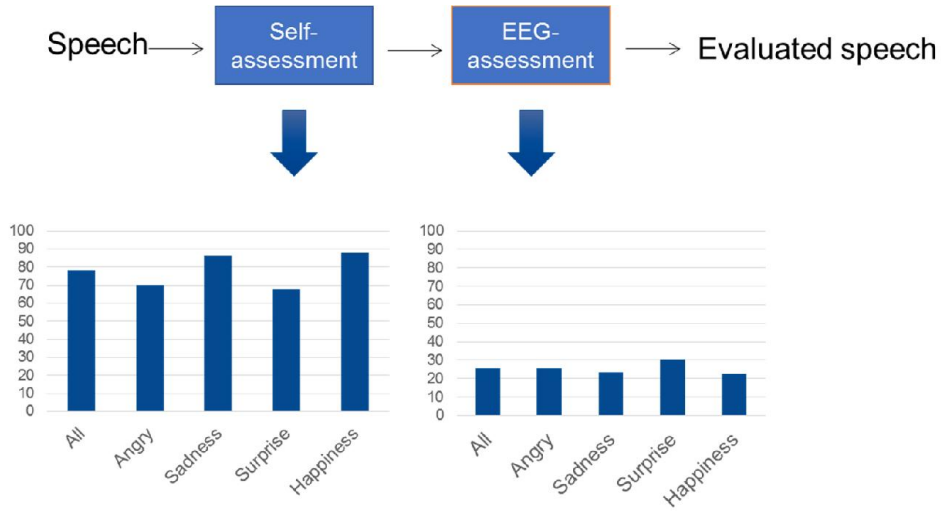

Figure 38. Illustration of data selection results.

3.6 SUMMARY

An emotional speech database is constructed. Variety of physiological signals is also included for further research. The data is validated not only by self-assessment but also by assessing by brain activities. In some other similar studies, another way for data validation called other persons' assessment is adopted. The reason why I replace other persons' assessment by EEG assessment is because of the difficulty to decide proper candidates for other persons' assessment. Other persons' assessment itself is a subjective procedure; the group of people selected for other persons' assessment will largely influence the evaluation results. Moreover, it's more difficult to evaluate the real-world emotions, which are related to the human health.

4 PURELY SEGMENT-LEVEL SPEECH EMOTION RECOGNITION

Recognizing human emotion from speech introduces promising applications of emotion analysis such as healthcare system, commercial conversations, virtual humans, emotion-based indexing, and in information retrieval. However, it's a challenging field and current status of emotion recognition performance still needs to be improved.

4.1 RESEARCH BACKGROUND

Recently, increasing attention has been drawn to identify emotions by using speech signals. There are many reasons for the popularity of speech signals for emotion recognition. It's the most natural and important way for human communications. As shown above, many researchers' purpose is for providing better healthcare system. In addition, with the tremendous research on speech recognition since the late fifties, emotion seems to be a huge gap between human and machines [105]. So computer scientists also have great interests in emotion recognition. Motivations vary among different researchers, but they are fundamentally same in terms of technologies [89], [106]–[115]. The information is summarized concerning features, and classification procedures.

1. Features

A large amount of features for characterizing emotional content of speech was tried by different researchers. The features can be grouped as follows.

Time-depended speech features. Well known features are pitch-related and energy-related features. Commonly used global features are the mean, the median, the standard deviation, the maximum, the minimum, the range, the mean of first difference, the linear regression coefficients, etc.

Spectral-based speech features. Emotional content of an utterance may have an impact on the shape of the spectral energy. These kind of features can be extracted in multiple ways such as linear predictor coefficients (LPC) [112].

Cepstral-based speech features. The bandwidth of filter is following the modified Mel-frequency scale instead of linear scale since the human perception of pitch doesn't follow it. Common used features are linear predictor ceptral coefficients (LPCC) [116] and Mel-frequency cepstrum coefficients (MFCC) [111].

There exist other approaches to extract features for characterizing emotional contents by detecting whether a voice is breathy by signal level (amplitude) and durations of a certain signal level.

2. Classification procedures

The whole classification procedure includes feature extraction and classification. After extracting the features, the approaches for detecting the emotional contents from speech utterances are introduced in details. It's common that researchers prefer extracting features from small frames divided from long speech signals because speech signals are not stationary. After extracting features from each frame, there are usually two ways to use them. One is to use all features directly and form a vector as local features; another is to calculate statistics of all speech features from all frames that are called as global features. Other than extracting features from same length frames, some researchers tried to extract features based on phonemes and voiced speech segments using previous demonstrated strategies. Since lots of the proposed work using global features shows higher performance, increasing number of researchers adopt global features in their research. Global features overlook the dynamic nature of a multivariate time-series (features) extracted from small frames of an utterance. Recent researches indicate that improvements can be made by adding segment level features to the common utterance level features using super utterance level features. The procedures are demonstrated and illustrated as follows.

The idea is from doubting whether the mainstream global features are proper for extracting enough information to detect emotional contents from speech signals. This approach adds segment level features into the feature vector of

global features. There are multiple segmentation strategies for extracting and utilizing segment level features. Absolute time intervals (ATI) segmentation of speech signals is the most straightforward way. Another approach is called relative time intervals (RTI), segment features are extracted from fixed relative positions from an utterance. A variant is called ATIR, which combines absolute time intervals and relative positions. Illustrates of these approaches are shown in Figure 39.

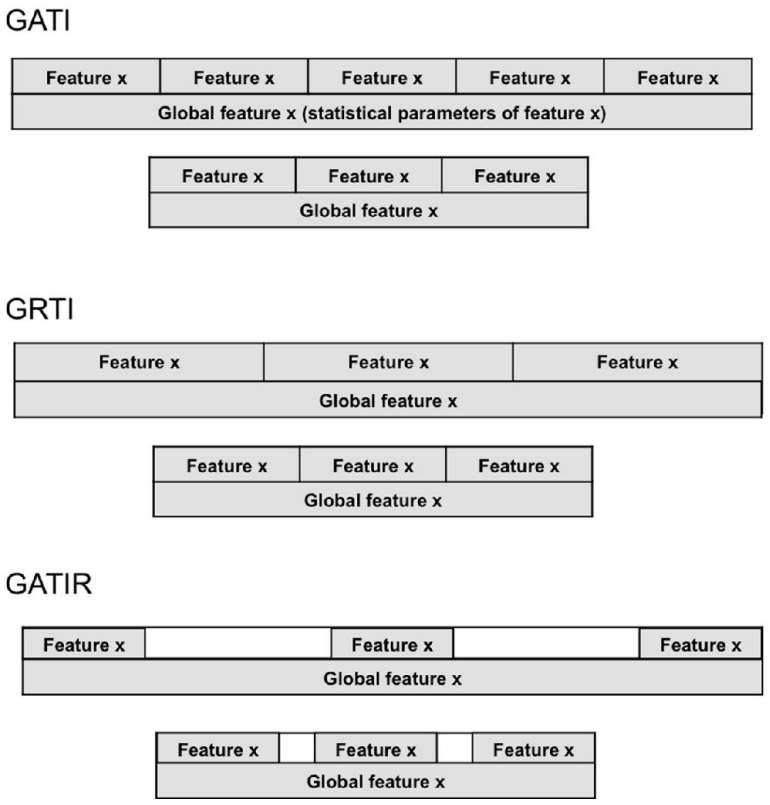

Figure 39. Feature extraction schemes illustrations of a short and a long utterances.

Many classifiers have been tested for speech emotion recognition such as support vector machine (SVM), artificial neural networks, hidden Markov model (HMM), Gaussian mixture models (GMM), k-nearest neighbors' algorithm, and so forth. Classifiers such as HMM are widely used in automatic speech recognition, therefore many works targeting automatic emotion recognition also adopt HMM. However, there are many issues for applying it on emotion recognition. The differences between automatic speech recognition and emotion recognition are obvious; HMM states in automatic speech recognition are aligned with acoustic features corresponding to phonemes or syllables. The accuracy adopting HMM on emotion recognition has no obvious advantage compared to other classification technologies such as SVM and neural networks. Take artificial neural network for example, they are proved to be more robust in modeling non-liner patterns, and the classification accuracy is better than HMM. SVM is also a very powerful classifier that have been applied extensively in many applications, it uses certain kind of kernel functions to map the original features to a high-dimensional space nonlinearly and then classify them using a linear classifier.

To summarize, speech emotion recognition is a very difficult task based on the following facts. Aside from the non-technical problems such as cultural difference among speakers, the disputes of emotion definition, there are many

technical issues to be solved. Firstly, it's not clear how to use speech features for classifying emotions. Speaking contents, speaking styles, and also speaking speed largely influence the acoustic features. For instance, common features such as pitch and energy contours are affected by the speaking speed. Secondly, multiple emotions can be represented in the same utterance. In this case, each portion can convey different emotional information so that it is not clear what kind of emotion should be detect from an automatic emotion recognizer.

4.2 DEFINING TERMINOLOGY OF EMOTIONS

An essential issue for conducting research to realize automatic emotion recognition from certain signal or multiple signals is to determine a set of the important emotions to be classified. By reviewing previous researches in the field of ecologists and psychologists, we know that multiple categories of emotions exist though there is no universally agreed theoretical definition. In order to fulfill different parts of this research, I utilize both discrete and dimensional models. Emotional categories are adopted in this thesis are from anger, disgust, happiness, sadness, fear, and surprise.

Discrete emotion model enables us to define a set of separated emotions that human beings are able to feel. Several researchers propose typical emotion states that are about 300 categories. It's not easy to directly classify such huge number of emotions; especially the technologies are still far from maturated. Additionally,

for developing emotion recognition methods, the defined emotion categories have to have the property of differentiable. Based on previous researchers and experiments, it is shown that many emotions cannot be separated easily even by human. It's a difficult task to confirm whether discrete emotional experiences exist since individual differences occur in emotional experiences. Moreover, each individual has his or her own definition of different emotions. By summarizing multiple theories, emotions can be considered uniquely by individuals such as happiness, angry, surprise, fear, disgust, and sadness. Ekman argues that cross-cultural basic emotions exist in the judgment of facial expression and these emotions are argued to be different from each other, characteristics can be described which are useful in distinguishing them.

Dimensional model of emotion draws more researchers' attention recently. Ekman, who is one of the founders of discrete emotion model, also consider it is enough to describe different emotions with a pleasant-unpleasant and active-passive scale. It shows the possibility to define some region for different emotion categories. The individual differences can also be considered and summarized in order to represent these "basic" emotions on dimensional model with a spatial region.

To summarize, we selected 4 human basic emotions including happiness, angry, surprise and sadness, which can be represented in dimensional model using valence and arousal axis and validated by using the proposed EEG assessment.

4.3 UTTERANCE LEVEL OR SEGMENT LEVEL?

The utterances (phrases, short sentences, etc) referred to in current research papers is often considered the fundamental unit and is recognized based on the global utterance-wise statistics of derived segment low-level descriptors (LLD), so the segment features are transformed into a single feature vector for each emotional utterance [12], [106], [109], [110], [117]. Most of these works rely on this assumption although lots of evidence indicates that the human brain (neurons networks) sometimes processes information and reacts within a second less than the utterance time. In recent research, an increasing number of scientists and psychologists have been arguing that emotion activity changes occur within a very short period of time are very important. Several studies have emphasized the importance of the temporal dynamics of emotions [118], [119]. Furthermore, one study that illustrates that emotions are inherently dynamic [120], the paper contains an illustration showing that within 2.6 seconds a person went through several emotional activities, such as surprise, fear, aggressive stance, and relaxation. In addition, a proposed method demonstrates that the emotion affect occurs within hundreds of milliseconds [121].

Motivated by these findings, we focus on a novel scheme to improve speech emotion recognition using segment level features instead of the strategy that includes ensemble learning from different classifiers using the same utterance-wise features [90], [108], [115]. Many researchers have recently been focusing on an issue that questions whether or not the utterance level is the right choice for modeling emotions [114]. They are concerned with this because of the difficulties with utterance-wise statistics in avoiding influence from spoken content, which requires accurate partitioning of an utterance for segmentation. Moreover, valuable but neglected information could be utilized in the segment-level feature extraction approach rather than calculating only the utterance-wise statistics. This hypothesis is also supported by many researches [111], [113], [114], based on the fact that improvements can be made by adding segment-level features to the common utterance-level features. The different schemes for obtaining these segments that have been discussed by these researchers are the global time interval (GTI), absolute time intervals (ATI), relative time intervals (RTI), and these schemes were used for constructing super-vectors, including the fusion of all segment features plus the global features (GRTI), and the fusion of the segment features from the absolute time intervals at relative positions plus the global features (GATIR). In addition to these super-vector features with a single classifier such as support vector

machine (SVM), some researchers have also used classifier ensembles that combine several base classification schemes into a larger meta classifier for utilizing both the utterance level and segment level information. Another approach for the decision model uses two classifiers for both the utterance level and segment level information, and then a decision fusion is used based on the results from the two classifiers [122].

I took into consideration a purely segment level strategy for speech emotion recognition and abandoned the utterance-wise features in order to reduce the noises from spoken content and utilize the neglected information in calculation of the utterance-wise statistics in this study. An issue raised when using segment-level speech emotion recognition is that it increases the difficulties for training to a large extent because a single utterance is divided into a number of segments. The aim of this paper is to properly design an approach for utterance level emotion recognition that is based on aggregating the segment level labels without introducing a huge computational complexity.

4.4 SEGMENT LEVEL BASED SPEECH EMOTION RECOGNITION

4.4.1 Experimental data for evoking emotions

A well-annotated database is needed to construct a robust model for recognizing emotions using speech signals. The experiment emphasizes natural

speaking. The participants are prevented from being aware that they are in an experimental environment during the experiments, which is much more realistic than experiments that were conducted with scripted acting speech [123]. Natural speech is difficult to analyze, but more suitable than scripted acting speech for validating the robustness of an emotion analysis method.

4.4.1.1 Experimental protocols

The experimental setup is composed of one instructor, one coordinator, and two participants. The coordinator cooperates with the participants in order to help better stimulate their emotions. The coordinator pretends to be one of the participants in the experiment to avoid being an extra obstruction for the real participants. The stimulation process unfolds through conversations with the aid of videos. The steps are demonstrated as follows.

- The instructor setups the experiment's environment, such as a projector for the videos and microphones for collecting the speech signals, and gives instructions to the participants.
- The instructor also explains the steps to the participants, including the coordinator, for freely providing their impressions related to the videos.
- For building an easy speaking atmosphere, self-introductions are made.

- After watching each emotion evoking video, which lasts several minutes, the speech signals are recorded from the impressions.

After the experiment, the emotion corresponding to each utterance from the recorded speech signals is not only self-assessed by the participants but also by 10 other people.

4.4.1.2 Data information

Ninety-six people participated in the experiments, which included 53 males and 43 females ranging from their early teens to 40s. We provided the sample selections to obtain reliable data in two steps. First, only the samples with the same label (pleasure or displeasure) based on the self-assessment and others-assessment were taken into consideration. Second, for maintaining a balance of the sample numbers for each label, we selected 300 utterances with higher rankings using the others-assessment, which consisted of 150 utterances as pleasure data and 150 utterances as displeasure data from 50 participants. Ten specialists put a label each for every utterance in the others-assessment, and the rank for each utterance was calculated based on the ratio of the numbers from the specialists who gave labels that were consistent with the label set from the self-assessment.

4.4.2 Methodology

The proposed methodology for emotion recognition is based on purely segment-level speech frames, and the important issues for consideration here are the increased number of samples that raise the computational burden in terms of both the memory capacity and execution speed, and the decline in the generalization ability of the classifier. In this work, I address the quantitative analysis of various analytical schemes related to segment level speech emotion recognition, and I propose an automatic approach for decreasing the number of samples in order to reduce the computational complexity and improve the classifier generalization ability.

4.4.2.1 Feature extraction

We focused on a set of 162 acoustic features from speech signals, including 50 Mel-Frequency Spectral Coefficients (MFCC), 50 Linear Predictive Coefficients (LPC), 10 statistical features (mode, median, mean, range, interquartile range, standard deviation, variation, absolute deviation, skewness, and Kurtosis) calculated from each of the five levels of detailed wavelet coefficients by using the Discrete Wavelet Decomposition (DWT), pitch, energy, zero-crossing rate (ZCR), the first seven formants, centroid, and 95%-roll-off-point from FFT-spectrum.

4.4.2.2 Segmentation approach

1. Existing segmentation schemes

Several segmentation strategies were proposed in a previous study [114]. In this research, I paid particular attention to the strategies on the segment level, and compared them with the utterance level approach. The proposed segmentation schemes in the previous study are demonstrated as follows and illustrated in Figure 40.

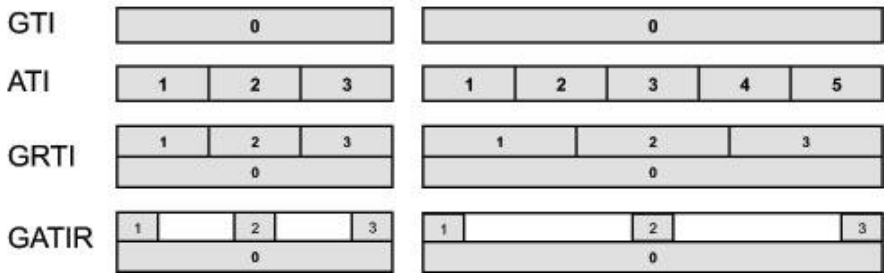

Figure 40. Illustrations of segmentation schemes.

GTI segmentation (Utterance-level segmentation). The speech signals are segmented by pauses during the speech without word or syllable boundary detection.

ATI segmentation. Different from utterance-level segmentation, speech utterances are segmented at the same fixed time interval.

RTI segmentation. Speech utterances are segmented at the fixed relative positions.

ATIR segmentation. ATIR segmentation combines the ideas of ATI and RTI segmentation. Fixed length segments are constructed at fixed relative positions, and this overcomes the drawback of different segment lengths and numbers obtained from different utterance lengths.

2. Proposed segmentation approach

The following two difficulties must be overcome in order to carry out segment level based speech emotion recognition. The first one is the computation burden imposed on the decision model is caused by a large increase in the number of segments. The second one, which is a more difficult issue, concerns defining the labels of the segments representing the classifiers. The proposed approach is illustrated in Figure 41.

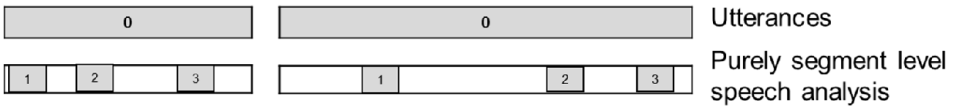

Figure 41. Illustration of proposed segmentation approach.

I propose the novel segmentation strategies, entropy based ATIR (eATIR), mutual information based ATIR (miATIR), and correlation coefficients based ATIR (crATIR) inspired from the ATIR segmentation method due to the

advantages of getting a smaller and fixed number of segments from an utterance. Moreover, it's assumed that some segments do not share the same label as the utterance. More precisely labels can be defined when taking into consideration a much smaller number of selected segments and these segments can better represent the utterance label than simply assuming the labels to all the segments obtained using the ATI segmentation method are the same as the utterance level label.

A classifier is trained by using the learning from the information contained in the input feature vectors to build a model. In the real world, the final uncertainty will not be ideally zero after training because of insufficient input information. In addition, the classifier might be "confused" due to ambiguities in the input information. The most likely solution is to increase the number of training samples, but this is not desirable in our case since the large increase in the number of training samples by splitting the utterance into segments has already been a great computational burden. However, a more efficient way is to find more informative segments by minimizing the amount of mutual information between the two feature vectors. In this study, fixed length segments are constructed at selected positions based on the designed indexes. So, not all the parts of the utterance are used in the analysis. Framing is used after dividing the

utterance into several sections to deal with the utterance signal. A 10-ms window with no overlap is used for calculating the ranking of the fixed length segment.

I use several measuring indexes on the extracted features to get the most representative ones among all the segments, and the average of the index values within each segment is used for the selection. The indexes are introduced in the following paragraphs.

Entropy index. Entropy is a measure of the information content, which is introduced as "a measure of how much 'choice' is involved in the selection of an event" [124]. In order to model an emotion with segments that are represented by an unknown probability distribution, we have to determine how to obtain the best approximation for creating the model. One approach is to have the distribution with the maximum entropy ensure that the approximation satisfies and subjects to any constraints on the unknown distribution [125]. In our study, the approximate distribution of acoustic features within a frame is defined as f, which assigns a probability $f(x)$ to each feature x from feature vector X. The entropy of f is defined as

$$H(f) = -\sum_{x \in X} f(x) \ln f(x) \tag{10}$$

where ln is the natural logarithm. A number of top ranked segments is selected after calculating the entropy of the segment features.

Mutual information index. The mutual information measures the "lumpiness" of the joint distribution, which is defined in terms of the entropy as follows for two feature vectors X and Y.

$$I(f_X; f_Y) = H(f_X) + H(f_Y) - H(f_X, f_Y) \tag{11}$$

where f_X and f_Y are the approximate distribution of X and Y, and the joint entropy can be defined as

$$H(f_X, f_Y) = -\sum_{x \in X} \sum_{y \in Y} f(x, y) \ln f(x, y) \tag{12}$$

Thus, several of the most informative segments can be selected by minimizing the redundant information.

Correlation coefficient index. The correlation coefficient [126], which is also known as the Pearson product-moment correlation coefficient, is a measure of the linear dependence between two feature vectors. It is defined as

$$\gamma = \frac{\sum_{x \in X, y \in Y} (x - \bar{X})(y - \bar{Y})}{\sqrt{\sum_{x \in X} (x - \bar{X})} \sqrt{\sum_{y \in Y} (y - \bar{Y})}} \tag{13}$$

$$\bar{X} = \frac{1}{n} \sum_{x \in X} x \tag{14}$$

$$\bar{Y} = \frac{1}{n} \sum_{y \in Y} y \tag{15}$$

Where n is the number of features. This index shares the same concept with the mutual information for reducing the redundancy.

The concept of the proposed segmentation methods is illustrated in Figure 42.

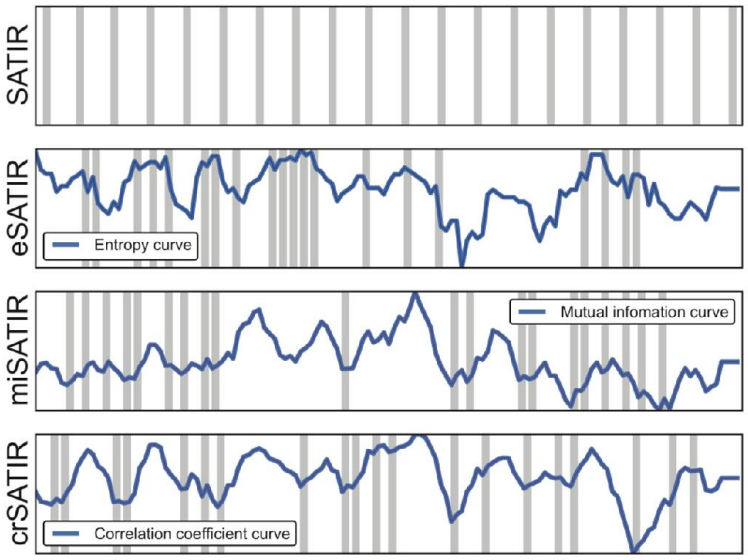

Figure 42. Fixed length segment positions illustration using proposed segmentation approaches (20-segment selecting situation is shown and the positions are represented using grey lines). 'S' is included in the abbreviation to represent the purely segment-level concept.

4.4.2.3 Decision model

The decision for the utterance is based on the prediction of its segments from a classifier. I simply use the efficiency approach called the majority vote, which determines the label of the utterance from the label in majority, in order to pay more attention to examining the effectiveness of the proposed segment-level approaches for speech emotion recognition. The decision model is shown in

Figure 43. Our decision model is based on a classifier called the probabilistic neural network (PNN) [63]. The PNN operations are designed into a multi-layered feed-forward network with four layers. It has many advantages compared to other kinds of artificial neural networks and nonlinear learning algorithms, including a very fast learning speed and less parameters.

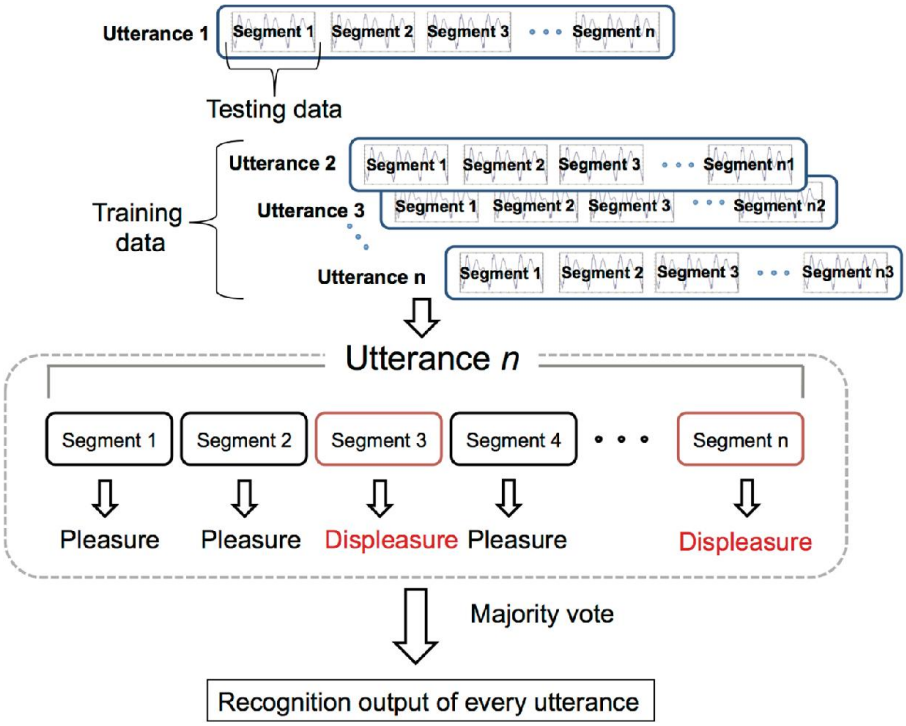

Figure 43. Illustration of segment-level classification concept for decision model.

4.4.3 Results

A 10-fold cross validation is used to evaluate and test our proposed approaches as well as make comparisons with previous researches because it is used in many

other emotion recognition researches for validating general models [113]. We reviewed all the most recent research on the aspect of classifiers for making a solid illustration and found that the support vector machine (SVM) is one of the most robust and popular classifiers in the field of affective researches, and it beats out many other kinds of classifiers in terms of the recognition accuracy [10]. Thus, our evaluation results based on PNN are compared with those based on SVM. Figure 44 clarifies the results with our proposed 162 acoustic features applied to the existing schemes, such as GTI and segment-level features that include schemes such as GRTI and GATIR. 500-ms segments are constructed at fixed relative positions for GATIR [114].

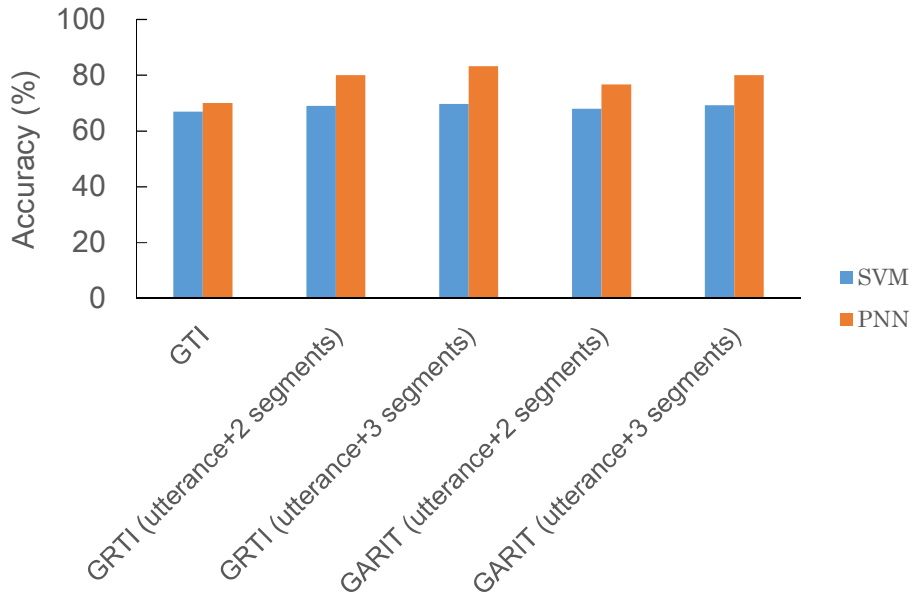

Figure 44. Comparison of emotion recognition accuracy between existing segmentation schemes using segment features with global features.

As shown in Figure 44, both classifiers succeeded in classifying emotions, while PNN performs a little better. These results are also consistent with other similar researches, which show a better potential for use as segment features for emotion recognition extracted by using RTI segmentation. However, this idea doesn't share the common goal of our research, which aims at reducing the computational complexity and better labeling the segments by selecting only partial information from within an utterance. Hereafter, we begin our pure segment features with a name beginning with S in order to discriminate those from global features or both proposed in previous researches. SATI (ATI) is used with different time intervals to conduct a first scenario that is different from that in utterance-level analysis. Figure 45 illustrates the obtained utterance-level emotion recognition results aggregated by a majority vote from the segment-level results with four different segment lengths. These results indicate that a higher level of accuracy can be achieved by using our proposed decision model with purely segment-level features than using segment features together with global features. In addition, this indicates that using a larger number of segments for training does not guarantee a higher level of accuracy and a longer segment will reduce the accuracy when using purely segment-level features. I, therefore, choose 50 ms as the segment length for our analysis.

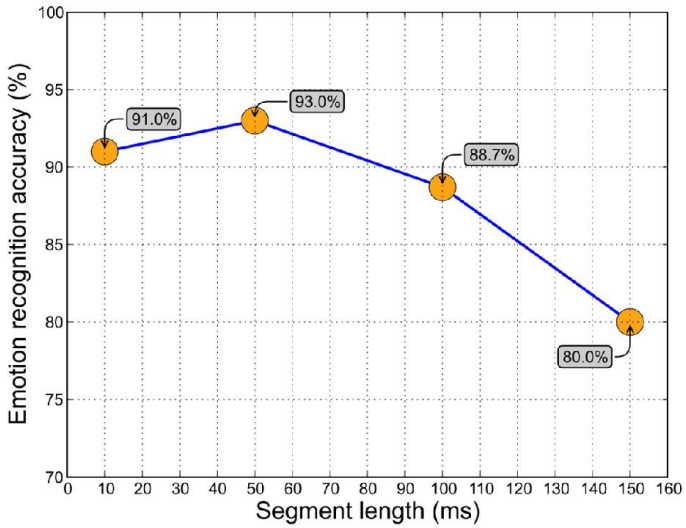

Figure 45. Speech emotion recognition accuracy based on segment-level analysis using absolute interval segment features (SATI).

I compare our proposed segment-level feature extraction strategies, which generate feature groups called eSATIR, miSATIR, and crSATIR in the sentences and paragraphs that follow. I also compare them with a segment-level feature extraction approach named SATIR directly inspired from previous research. I have to define the number of segments we want to generate from each utterance before applying the proposed strategies. Since a majority vote is used in the decision model, 10 segments are considered as reasonable for use as the smallest number for voting. It is also taken note of the fact that some utterance lengths are as short as one second, and thus, 20 segments is considered to be the largest

number for a majority vote. Figure 46 provides a comparison of the emotion recognition accuracy between different segmentation schemes using purely segmented features. According to these comparison results, high performance is achieved using both 10 and 20 segments for voting. The results also show that miSATIR and crSATIR can greatly increase the utterance-level speech emotion recognition accuracy, where crSATIR leads to the best results.

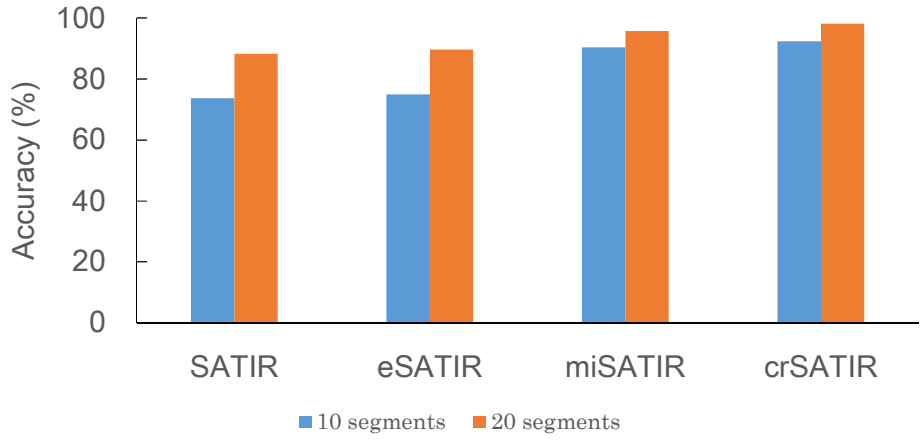

Figure 46. Comparison of emotion recognition accuracy between different segmentation schemes using purely segmented features and PNN (σ =0.2).

4.4.4 Discussion

Previous research has reported on the strategies for improving the speech emotion recognition accuracy by utilizing segment-level features together with global features extracted from utterances. The effectiveness of these strategies was proved in many reports [113], [114]. This research further develops a new

approach that totally abandons the global features from the utterances. The analytical results in 4.4.3 indicate the robustness of this advancement, which leads to a higher level of recognition accuracy by only using segment-level features in the proposed decision model. I tested different segment lengths for SATI in order to discover the proper one for extracting features, which also changes the number of obtained segments. According to the results illustrated in Figure 45, the highest number of training samples did not lead to the best level of accuracy, which might be explained by the redundant information existing among all the segments. The accuracy also decreases when the length of segments is longer. The reasons for this are that the number of segments for the decision model decreases, and the features are more likely to be similar to the global ones (GTI) from the utterance. We finally choose 50 ms as the segment length not only based on the better level of accuracy, but also based on consideration of the emotional variability and content. A shorter duration such as may lack the emotional content necessary for learning a decision model. While a longer length of segment may influenced by the time vary characteristic of speech signals or contain redundant information for representing the utterance label, the analytical results also show that the performance of the decision model is adversely affected.

In order to select propose segments in the large increased number of training samples by splitting the utterances and minimizing the redundancy, I proposed several approaches for segmentation in order to select the appropriate number of segments within an utterance. SATIR is simply extracted from the selected segments at fixed positions according to the length of the utterance, and the recognition results can be seen as a reference. Two kinds of features are introduced from the segmentation approaches that are based on information theory including eSATIR and miSATIR, but the improvement when using miSATIR is much larger than that for eSATIR, where eSATIR can only slightly increase the level of accuracy compared to SATIR. Since eSATIR doesn't consider the relationship between the utterance and its segments, the segments generated by only considering the maximum entropy have a higher chance to obtain confusing information such as noise or non-representative information for the utterance label and this might be the cause for the difference. However, the miSATIR and crSATIR approaches generate segments with less redundant information for the decision model, which contributes to a better understanding of the utterance label, and a better comprehensibility of the learned model.

As can be seen in Figure 46, the results of crSATIR outperform those of miSATIR. This phenomenon seems hard to explain at first glance. In lots of situations, they share the same purposes and the mutual information makes it

more powerful for detecting the complex non-linear relationships between two feature vectors, and this helps in problem solving for a lot of difficult issues such as in network analysis [126], [127]. However, the situation for feature extraction is different from that in our case, with the aim of better extracting information for representing an utterance label. The mutual information, which is the strategy adopted in our research in order to reduce the redundant information when extracting features from segments, is able to detect these non-linear relationships and avoid them to a large extent. However, the correlation coefficients strategy investigates whether a linear relationship between feature vectors exists and minimizes it. In our case, two feature vectors with a small or no linear relationship might have a strong non-linear relationship, and this dynamic relationship can be considered useful information for better representing the utterance labels, which might benefit the decision model learning so that using crSATIR leads to an improved level of accuracy compared with using miSATIR.

I used a 162-dimension feature set for a complete analysis, but a remaining point is that we are not including a feature selection procedure before the segmentation. The full set of extracted features is chosen instead and let the segmentation algorithm decide on the more appropriate segments for representing the utterance labels, and this is appropriate with respect to feature dimensions with a large number of sample. However, the interaction between the

feature selection and segmentation approaches and its meaning will be discussed as a future issue.

The analytical results show the effectiveness of 50ms in segment-level emotion recognition. A research demonstrates that emotion effect happens in hundreds of milliseconds [121]. A shorter time interval for studying emotions have not been studied so far in neuroscience. From the view of typical frame size for extracting acoustic features, 50ms is an appropriate frame size due to it contains several fundamental periods of the audio signals. If it is too short such as less than 10ms, some features cannot be correctly estimated such as pitch. And if the frame duration is too big, the time-varying characteristics of the speech signals may influence the features. However, More discussions of proper frame length need to be discussed.

The majority voting based on 20 segments outperforms that based on 10 segments using proposed segment selection approaches. As for entropy based segment selection method, a consecutive part of speech signal may contain larger entropy as shown in Figure 42. If less number of segments are used for voting, it highly possible that all segments are selected from that consecutive part so that it overlooks the overall emotional information of an utterance. In the other two methods including mutual information and correlation coefficient based selection, they can be less influence by this matter by selecting a segments with have less

dependency. This explains that the improvements are larger for entropy based selection when using 20 segments. However, the improvements will become less if the number of segment increases. In majority of the testing, 20 segments can overcome the demonstrated problem and give better fault-tolerance. Considering the segment length and numbers as well as utterance length, 20 segment voting is adopted for avoiding situations that the total segment length is larger than utterance length.

4.5 REAL-WORLD SPEECH EMOTION RECOGNITION BY PROPOSED METHODS

The proposed method is further evaluated using the emotion speech with 4 labels elicited using real experiences. The results are calculated using 10-fold cross validation. The selected database contains 1491 utterances (29820 segments).

4.5.1 Results of four-emotion recognition

The segment-level emotion recognition accuracy is illustrated in Figure 47 and the emotion recognition accuracy of utterances using majority voting is shown in Figure 48.

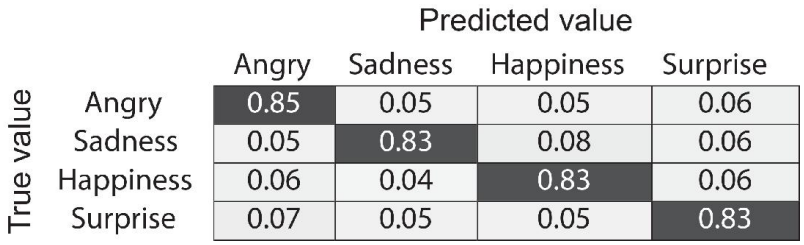

Figure 47. Confusion matrix of four-emotion recognition in segment level (10-fold cross validation).

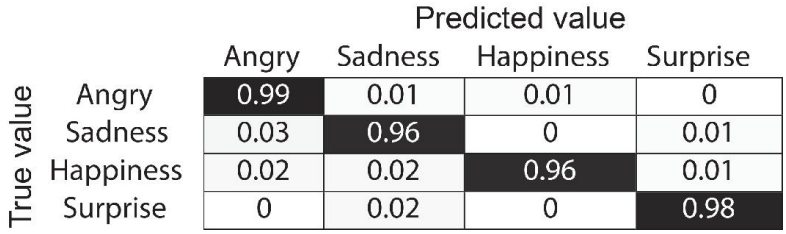

Figure 48. Confusion matrix of four-emotion recognition in utterance level (10-fold cross validation).

4.5.2 Comparison results

As for comparisons, accuracies with optimized conventional method (utterance-level features) are calculated. The utterance-level features are selected using RELIEF according to the performance among feature selection methods that are shown in Table 4. The results are shown in Figure 49.

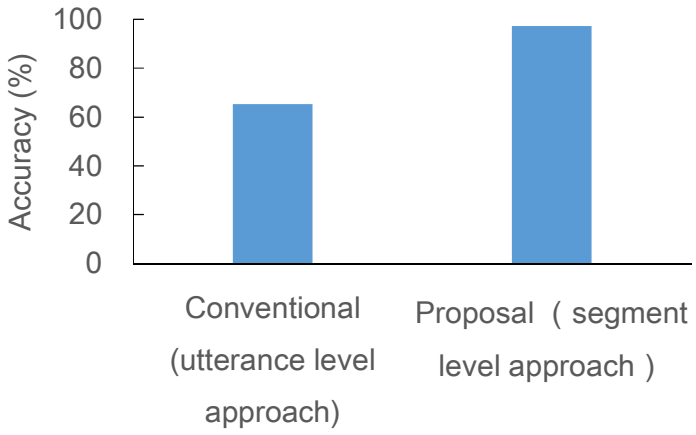

Figure 49. Comparison of conventional and proposed methods.

4.5.3 Results of database evaluation

In order to evaluate the quality of selected database, DB quality is defined as testing accuracy of the following scenario. 18000 samples are selected randomly in each database shown in Figure 50 for training, and 2000 samples are selected randomly in the original database. There is no overlap samples between training and testing databases. Emotion recognition performances in segment level using selected and unselected database (control) are shown in Figure 51.

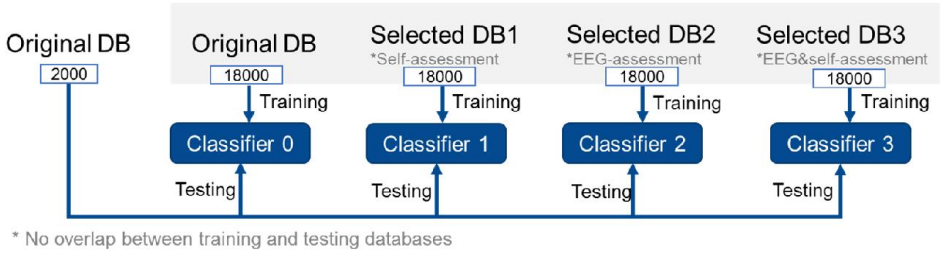

Figure 50. Illustration of original and selected speech databases.

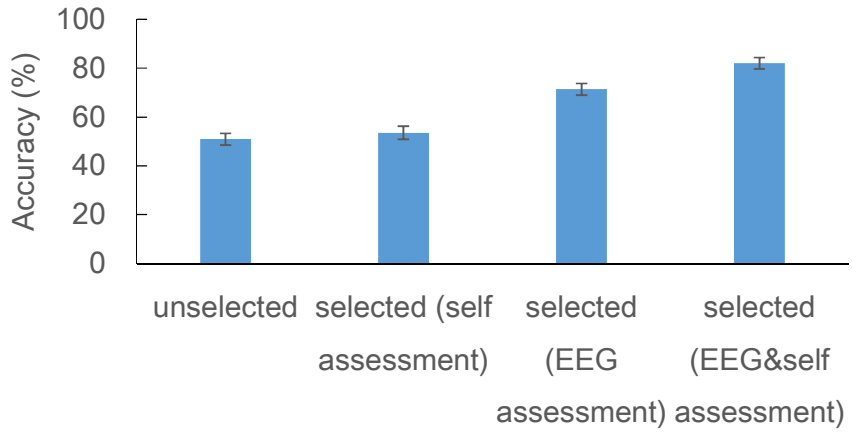

Figure 51. Emotion recognition performance using different databases.

4.5.4 Further validation of four-emotion recognition

Additional validation has been carried out on a more severe scenario. Since in the real case, the testing samples cannot be selected in advance, the testing samples (2000 samples) are selected randomly in the original database, and training database are selected randomly in the selected database (18000 database). There is no overlap between training and testing databases. The detailed results of the validation using PNN is shown in Figure 52.

		Predicted value			
		Angry	Sadness	Happiness	Surprise
True value	Angry	0.85	0.03	0.13	0
	Sadness	0.12	0.76	0.09	0.03
	Happiness	0.02	0.02	0.95	0
	Surprise	0.03	0.14	0.11	0.72

Figure 52. Confusion matrix of four-emotion recognition.

4.5.5 Discussions

The proposed purely segment-level method is further validated using emotional speech data elicited by real experiences, which is more similar to real-world emotions that are not exaggerate. The segment-level accuracy is more than 80% and utterance-level accuracy is more than 90%, while the classification of 4 emotions failed using global features extracted from utterances. The accuracy in utterance-level is more than 80% on the unselected testing database using a different training database.

The performance using selected and unselected databases indicates that the selected data can be better modeled compared to the unselected original database, which shows the effectiveness of the proposed data selection method using EEG signals. Self-assessments can also improve the performance of cross-validation, the quality of selected emotional speech database is further ensured together with self-assessments.

4.6 APPLICATION PERSPECTIVE: EMOTION STRENGTH ANALYSIS

A very interesting potential application area is emotion strength analysis by using segment-level speech emotion recognition. Majority voting is used for the utterance labels prediction with the assumption that the label of the most predicted segments represents the utterance label. For better understanding the segment labels, we further look into the ratio of the most predicted label that can

represent the strength of the utterance emotion. SATI is used because we want to examine all the segments in terms of the emotions.

4.6.1 Experimental data

I used the International Affective Picture System (IAPS) [70] for evoking emotions with different strengths. The IAPS is an emotion stimulation system built from the results of many emotion experiments. The picture system is composed of about 1000 pictures labeled with a standard scale of valence (pleasure-displeasure) and arousal (exciting-sleepy). Therefore, it meets our requirement for stimulating emotions with different strengths. shows the four kinds of emotions we defined using the IAPS.

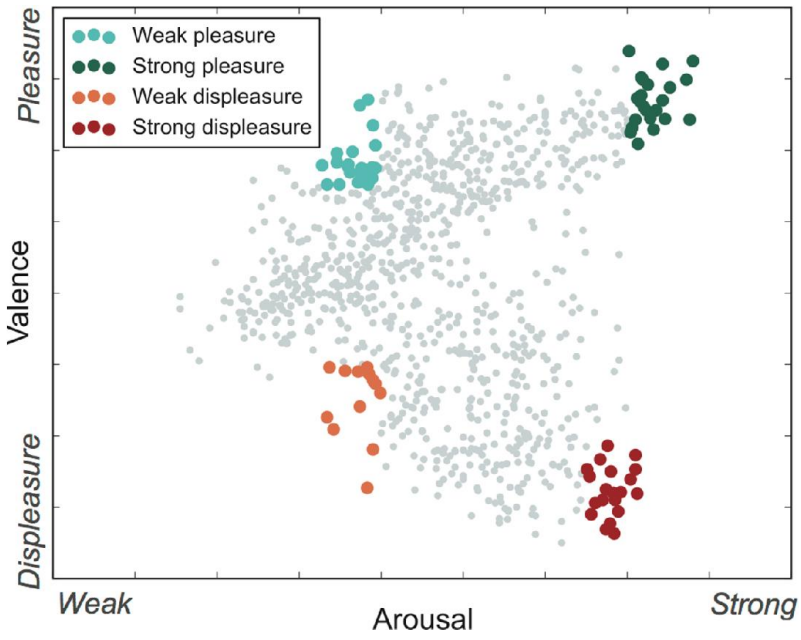

Figure 53. Defined emotion strength based on IAPS.

The experimental approach was made up of four parts according to the pleasure and displeasure emotion stimulation, which includes the defined emotion strength (weak and strong). The detailed experimental protocol is shown in . The pictures selected from the IAPS during the stimulation period were projected on a screen to evoke emotions and the speech signals were then collected when the participants were reading designed scripts with their evoked emotion after viewing each picture, and they were requested to close their eyes to relax during the control time. Seven Japanese males took part in the experiment. Data was collected using the previously described procedures for estimating the emotional strength, which contains 312 samples including 156 pleasure (78 strong, 78 weak) and 156 displeasure (78 strong, 78 weak) data.

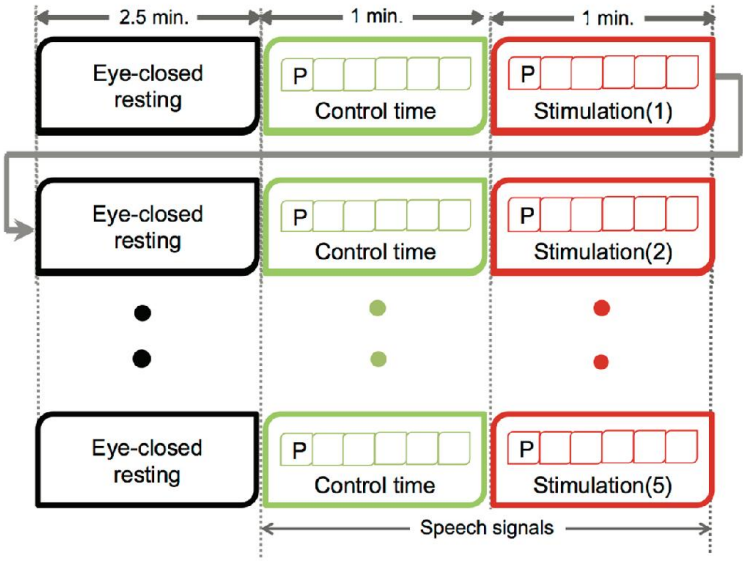

Figure 54. Experimental protocols.

4.6.2 Results

I statistically analyzed the components of all the samples and then visualized the information using a bar chart with a standard deviation to illustrate the correlations between stimulations (Figure 55) and the output of the emotion components represented by segment-level predictions within an utterance using the proposed segment-level speech emotion recognition method.

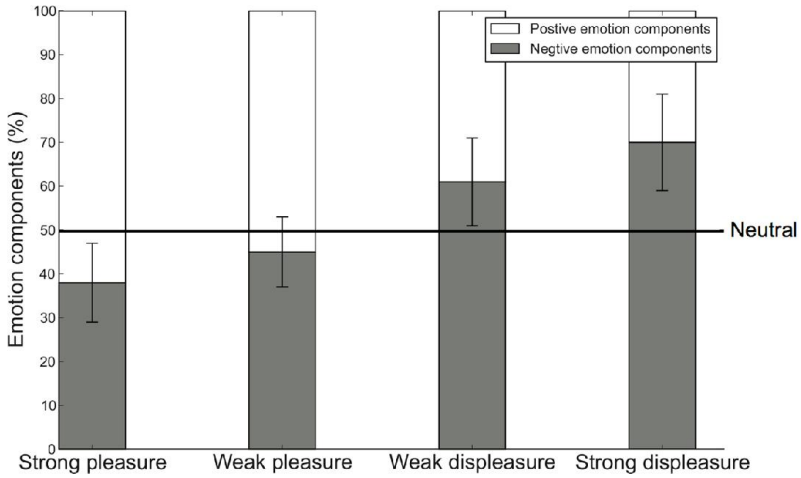

Figure 55. Statistical analysis for emotion components using segment-level speech emotion analysis for all speech samples.

4.6.3 Discussions

The emotion recognition of utterances is one of the more attractive topics in speech analysis for HCI. However, emotion strength analysis has been a very essential but difficult research area. We discussed the potential for using segment-level frames for the emotion strength analysis within utterances. As

shown in Figure 55, the proposed method can indeed reflect the strengths of emotions in utterance clusters for a number of spoken phrases or short sentences over a short period of time. However, difficulties still exist in applying it to a single utterance because of the variances in the emotional components regarding the utterances. Although further validation is necessary for collecting more solid findings in terms of the emotion strength analysis of utterances, segment level speech emotion analysis creates a new focus for better recognizing human emotion strength using machines.

4.7 SUMMARY

A purely segment level speech emotion recognition method is proposed in the first time. In order to make the proposed method more efficient and accurate, advanced relative segmentation method is firstly introduced by using mutual information (miSATRI) and correlation coefficients (crSARTI). Fixed length segment selection at relative positions is proposed, which is essential for realizing the purely segment level approach. The proposed method can greatly increase the emotion recognition accuracy of four-emotion recognition to more than 80% using a validation database with speech signals from 50 participates. The proposed method also showed the effectiveness of determining the emotion strength of utterances over a period of time.

5 CONCLUSIONS

1. EEG based emotion recognition method is proposed with high accuracy

According to variety of emotion theories and researchers, brain activities are considered as the most reliable signals for emotion recognition. EEG signals are extensively studied for demonstrating brain activities. However, emotion recognition accuracy by using only EEG signals are still not high enough as a reference for the purpose of data selection for building high quality database. For increasing the EEG-based emotion recognition accuracy, cross-level wavelet features are proposed. This approach increases emotional valence and arousal recognition accuracy to more than 90%, whereas previous research accuracy is from 40% to 70%. The high accuracy enables EEG signals to be used as a reference signal for selecting high quality data in the database. Moreover, it has more candidate applications. Additional emotion information can be extracted using current brain computer interfaces.

2. A natural emotional speech database is presented

Recent popularity of researches on human machine interfaces promotes many researchers to design emotional databases for algorithm evaluations. Current public emotional databases mostly include acting emotions for this demonstrated purpose. Increasing researches indicate that acting emotions have less or no influence of human health, and real-world emotions are usually not as exaggerate

as acting emotions. For the purpose of evaluating the developed emotion recognition approach, a database with natural speech and real-world emotions is necessary. Thus, a new database with natural speech under emotions evoked by real experiences is presented. EEG signals are recorded as a reference for ensuring the quality of collected emotional data.

3. Purely segment level emotion recognition using speech is achieved

For achieving a higher accuracy of speech emotion recognition, the classification scheme is evaluated. Purely segment level classification is proposed in this thesis, although it has been a long time for treating utterances as classification targets. Current features extracted from speech signals have been proved effective in speech recognition and further applied in most of emotion recognition researches. However, popular features for speech recognition and emotion recognition largely influenced by each individual's speaking content, speaking speed, speaking style, etc. Moreover, the emotional activity in brain changes in much shorter time than an utterance; voice is directly influenced by these changes. Utterance features are not efficient for capturing such information. Thus, a new emotion recognition approach using segments are proposed in order to overcome the described issues. In the proposed approach, utterance level features are totally abandoned, large improvements of classification accuracy compared to current emotion recognition approaches using speech, more than 80%

accuracy is achieved for recognizing human emotions from speech during real emotional experience recalling.

REFERENCES

[1] L. D. Kubzansky, N. Park, C. Peterson, P. Vokonas, and D. Sparrow, "Healthy psychological functioning and incident coronary heart disease: the importance of self-regulation.," *Arch. Gen. Psychiatry*, vol. 68, no. 4, pp. 400–408, 2011.

[2] T. W. Smith, K. Glazer, J. M. Ruiz, and L. C. Gallo, "Hostility, anger, aggressiveness, and coronary heart disease: an interpersonal perspective on personality, emotion, and health.," *J. Pers.*, vol. 72, no. 6, pp. 1217–1270, 2004.

[3] B. C. Sirois and M. M. Burg, "Negative emotion and coronary heart disease," *Behav. Modif.*, vol. 27, no. 1, pp. 83–102, 2003.

[4] G. H. Elder and E. C. Clipp, "Combat experience and emotional health: impairment and resilience in later life.," *J. Pers.*, vol. 57, no. 2, pp. 311–341, 1989.

[5] S. D. Rauch, "Clinical Hints and Precipitating Factors in Patients Suffering from Meniere's Disease," *Otolaryngologic Clinics of North America*, vol. 43, no. 5. pp. 1011–1017, 2010.

[6] L. S. Richman, L. Kubzansky, J. Maselko, I. Kawachi, P. Choo, and M. Bauer, "Positive emotion and health: going beyond the negative.," *Health Psychol.*, vol. 24, no. 4, pp. 422–429, 2005.

[7] E. Diener and M. E. P. Seligman, "Very happy people.," *Psychol. Sci. a J. Am. Psychol. Soc. / APS*, vol. 13, no. 1, pp. 81–84, 2002.

[8] D. M. Clark and J. D. Teasdale, “Diurnal variation in clinical depression and accessibility of memories of positive and negative experiences.,” *J. Abnorm. Psychol.*, vol. 91, no. 2, pp. 87–95, 1982.

[9] R. W. Picard, *Affective Computing*. MIT Press, 2000.

[10] D. Morrison, R. Wang, and L. C. De Silva, “Ensemble methods for spoken emotion recognition in call-centres,” *Speech Commun.*, vol. 49, no. 2, pp. 98–112, 2007.

[11] J. Gratch and S. Marsella, “Evaluating a Computational Model of Emotion,” *Autonomous Agents and Multi-Agent Systems*, vol. 11, no. 1. pp. 23–43, 2005.

[12] M. K. Shan, F. F. Kuo, M. F. Chiang, and S. Y. Lee, “Emotion-based music recommendation by affinity discovery from film music,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7666–7674, 2009.

[13] W. Swartout, J. Gratch, R. W. W. Hill, E. Hovy, S. Marsella, J. Rickel, and D. Traum, “Toward virtual humans,” *AI Mag.*, vol. 27, no. 2, p. 96, 2006.

[14] B. N. Colby, A. Ortony, G. L. Clore, and A. Collins, “The Cognitive Structure of Emotions.,” *Contemporary Sociology*, vol. 18, no. 6. p. 957, 1989.

[15] K. Oatley and P. N. Johnson-Laird, “Cognitive approaches to emotions,” *Trends in Cognitive Sciences*, vol. 18, no. 3. pp. 134–140, 2014.

[16] J. A. Ruth, F. F. Brunel, and C. C. Otnes, “Linking Thoughts to Feelings: Investigating Cognitive Appraisals and Consumption Emotions in a

Mixed-Emotions Context," *Journal of the Academy of Marketing Science*, vol. 30, no. 1. pp. 44–58, 2002.

[17] D. L. Hull, "Deconstructing Darwin: Evolutionary Theory in Context," *Journal of the History of Biology*, vol. 38, no. 1. pp. 137–152, 2005.

[18] J. D. Newman and J. C. Harris, "The scientific contributions of Paul D. MacLean (1913-2007).," *J. Nerv. Ment. Dis.*, vol. 197, no. 1, pp. 3–5, 2009.

[19] J. M. Barbalet, "William James' Theory of Emotions: Filling in the Picture," *J. Theory Soc. Behav.*, vol. 29, pp. 251–266 Click Here http //www epnet com/ehost/indi, 1999.

[20] P. J. Lang, "The varieties of emotional experience: a meditation on James-Lange theory.," *Psychol. Rev.*, vol. 101, no. 2, pp. 211–221, 1994.

[21] B. H. Friedman, "Feelings and the body: The Jamesian perspective on autonomic specificity of emotion," *Biological Psychology*, vol. 84, no. 3. pp. 383–393, 2010.

[22] S. Schachter and J. E. Singer, "Cognitive, social, and physiological determinants of emotional state.," *Psychol. Rev.*, vol. 69, pp. 379–399, 1962.

[23] R. Reisenzein, "The Schachter theory of emotion: two decades later.," *Psychol. Bull.*, vol. 94, no. 2, pp. 239–264, 1983.

[24] L. F. Barrett, "Solving the emotion paradox: categorization and the experience of emotion.," *Pers. Soc. Psychol. Rev.*, vol. 10, no. 1, pp. 20–46, 2006.

[25] D. DeSteno, R. E. Petty, D. D. Rucker, D. T. Wegener, and J. Braverman, "Discrete emotions and persuasion: the role of emotion-induced expectancies.," *J. Pers. Soc. Psychol.*, vol. 86, no. 1, pp. 43–56, 2004.

[26] C. Darwin, *The origin of species by means of natural selection; or, The preservation of favored races in the struggle for life*, vol. 78. 1872, pp. 1–248.

[27] P. Ekman, "Are there basic emotions?," *Psychol. Rev.*, vol. 99, no. 3, pp. 550–553, 1992.

[28] P. Ekman, "Basic emotions," *Cognition*, 1999. .

[29] J. Prinz, "Which emotions are basic," *Emot. Evol. Ration.*, pp. 1–19, 2004.

[30] A. Ortony and T. J. Turner, "What's basic about basic emotions?," *Psychol. Rev.*, vol. 97, no. 3, pp. 315–331, 1990.

[31] C. E. Izard, "Basic emotions, relations among emotions, and emotion-cognition relations.," *Psychol. Rev.*, vol. 99, no. 3, pp. 561–565, 1992.

[32] J. A. Russell, "A circumplex model of affect.," *Journal of Personality and Social Psychology*, vol. 39, no. 6. pp. 1161–1178, 1980.

[33] P. A. Lewis, H. D. Critchley, P. Rotshtein, and R. J. Dolan, "Neural correlates of processing valence and arousal in affective words.," *Cereb. Cortex*, vol. 17, no. 3, pp. 742–748, 2007.

[34] M. M. A. Nielen, D. J. Heslenfeld, K. Heinen, J. W. Van Strien, M. P. Witter, C. Jonker, and D. J. Veltman, "Distinct brain systems underlie the

processing of valence and arousal of affective pictures," *Brain Cogn.*, vol. 71, no. 3, pp. 387–396, 2009.

[35] B. R. Sheth and T. Pham, "How emotional arousal and valence influence access to awareness," *Vision Res.*, vol. 48, no. 23–24, pp. 2415–2424, 2008.

[36] S. Anders, M. Lotze, M. Erb, W. Grodd, and N. Birbaumer, "Brain activity underlying emotional valence and arousal: a response-related fMRI study.," 2004.

[37] L. N. Jefferies, D. Smilek, E. Eich, and J. T. Enns, "Emotional valence and arousal interact in attentional control.," *Psychol. Sci. a J. Am. Psychol. Soc. / APS*, vol. 19, no. 3, pp. 290–295, 2008.

[38] E. Buch, C. Weber, L. G. Cohen, C. Braun, M. A. Dimyan, T. Ard, J. Mellinger, A. Caria, S. Soekadar, A. Fourkas, and N. Birbaumer, "Think to move: a neuromagnetic brain-computer interface (BCI) system for chronic stroke.," *Stroke.*, vol. 39, no. 3, pp. 910–917, 2008.

[39] F. Lotte, J. Faller, and C. Guger, "Combining BCI with Virtual Reality: Towards New Applications and Improved BCI," in *Proceedings of the 6th International Conference on Foundations of Digital Games*, 2013, pp. 1–24.

[40] S. Marcel and J. D. R. Millán, "Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation.," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 743–752, 2007.

[41] R. Adolphs, D. Tranel, and A. R. Damasio, “Dissociable neural systems for recognizing emotions,” *Brain Cogn.*, vol. 52, no. 1, pp. 61–69, 2003.

[42] N. Lee, A. J. Broderick, and L. Chamberlain, “What is ‘neuromarketing’? A discussion and agenda for future research,” *Int. J. Psychophysiol.*, vol. 63, no. 2, pp. 199–204, 2007.

[43] S. A. Hosseini and M. B. Naghibi-Sistani, “Emotion recognition method using entropy analysis of EEG signals,” *I.J. Image, Graph. Signal Process.*, vol. 5, no. August, pp. 30–36, 2011.

[44] P. C. Petrantonakis and L. J. Hadjileontiadis, “EEG-based emotion recognition using hybrid filtering and higher order crossings,” *2009 3rd Int. Conf. Affect. Comput. Intell. Interact. Work.*, 2009.

[45] A. B. Geva and D. H. Kerem, “Forecasting generalized epileptic seizures from the EEG signal by wavelet analysis and dynamic unsupervised fuzzy clustering.,” *IEEE Trans. Biomed. Eng.*, vol. 45, no. 10, pp. 1205–1216, 1998.

[46] J. Liu and J. Xu, “Compared study of the analyzing methods for EEG data,” *Comput. Sci. Inf. Technol. (ICCSIT), 2010 3rd IEEE Int. Conf.*, vol. 9, 2010.

[47] W. Y. Hsu, C. C. Lin, M. S. Ju, and Y. N. Sun, “Wavelet-based fractal features with active segment selection: Application to single-trial EEG data,” *J. Neurosci. Methods*, vol. 163, no. 1, pp. 145–160, 2007.

[48] Q. Xu, H. Zhou, Y. Wang, and J. Huang, "Fuzzy support vector machine for classification of EEG signals using wavelet-based features," *Med. Eng. Phys.*, vol. 31, no. 7, pp. 858–865, 2009.

[49] E. E. Hall, P. Ekkekakis, and S. J. Petruzzello, "Predicting affective responses to exercise using resting EEG frontal asymmetry: Does intensity matter?," *Biol. Psychol.*, vol. 83, no. 3, pp. 201–206, 2010.

[50] A. Subasi, "EEG signal classification using wavelet feature extraction and a mixture of expert model," *Expert Systems with Applications*, vol. 32, no. 4. pp. 1084–1093, 2007.

[51] M. Murugappan, "Classification of human emotion from EEG using discrete wavelet transform," *Journal of Biomedical Science and Engineering*, vol. 03, no. 04. pp. 390–396, 2010.

[52] K. Schaaff and T. Schultz, "Towards emotion recognition from electroencephalographic signals," *2009 3rd Int. Conf. Affect. Comput. Intell. Interact. Work.*, 2009.

[53] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 81–97, 2010.

[54] L. I. Aftanas, N. V. Reva, A. A. Varlamov, S. V. Pavlov, and V. P. Makhnev, "Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: Temporal and topographic characteristics," *Neurosci. Behav. Physiol.*, vol. 34, pp. 859–867, 2004.

[55] S. G. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, 1989.

[56] J. Sherwood and R. Derakhshani, “On classifiability of wavelet features for EEG-based brain-computer interfaces,” *2009 Int. Jt. Conf. Neural Networks*, 2009.

[57] P. Zarjam, M. Mesbah, and B. Boashash, “Detection of newborn EEG seizure using optimal features based on discrete wavelet transform,” *2003 IEEE Int. Conf. Acoust. Speech, Signal Process. 2003. Proceedings. (ICASSP ’03).*, vol. 2, 2003.

[58] E. L. van den Broek and J. H. D. M. Westerink, “Considerations for emotion-aware consumer products.,” *Appl. Ergon.*, vol. 40, no. 6, pp. 1055–64, Nov. 2009.

[59] J. H. Holland, “Adaptation in Natural and Artificial Systems,” *Ann Arbor MI Univ. Michigan Press*, vol. Ann Arbor, pp. 1–200?, 1975.

[60] N. T. Alves, S. S. Fukusima, and A. Aznar-Casanova, “Models of brain asymmetry in emotional processing,” *Psychology and Neuroscience*, vol. 1, no. 1. pp. 63–66, 2008.

[61] C. Chu, A. L. Hsu, K. H. Chou, P. Bandettini, and C. Lin, “Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images,” *Neuroimage*, vol. 60, no. 1, pp. 59–70, 2012.

[62] J. Shlens, “A Tutorial on Principal Component Analysis,” *Measurement*, vol. 51, p. 52, 2005.

[63] D. F. Specht, “Probabilistic neural networks,” *Neural networks*, vol. 3, no. 1, pp. 109–118, 1990.

[64] S. L. Wang, X. Li, S. Zhang, J. Gui, and D. S. Huang, “Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction,” *Comput. Biol. Med.*, vol. 40, no. 2, pp. 179–189, 2010.

[65] A. Sivakumar and K. Kannan, “A novel feature selection technique for number classification problem using PNN-A plausible scheme for boiler flue gas analysis,” *Sensors Actuators, B Chem.*, vol. 139, no. 2, pp. 280–286, 2009.

[66] R. M. Balabin, R. Z. Safieva, and E. I. Lomakina, “Near-infrared (NIR) spectroscopy for motor oil classification: From discriminant analysis to support vector machines,” *Microchem. J.*, vol. 98, no. 1, pp. 121–128, 2011.

[67] K. Takahashi, “Remarks on emotion recognition from multi-modal bio-potential signals,” *2004 IEEE Int. Conf. Ind. Technol. 2004. IEEE ICIT ’04.*, vol. 3, 2004.

[68] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, I. Zunaidi, and D. Hazry, “EEG feature extraction for classifying emotions using FCM and FKM,” *Int. J. Comput. Commun.*, vol. 1, no. 2, pp. 21–25, 2007.

[69] M. Arvaneh, C. Guan, K. K. Ang, and C. Quek, "Optimizing the channel selection and classification accuracy in EEG-based BCI.," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 6, pp. 1865–1873, 2011.

[70] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, *International affective picture system (IAPS): Affective ratings of pictures and instruction manual.* University of Florida, 2008, p. Technical Report A–8.

[71] M. M. Bradley and P. J. Lang, "The International Affective Picture System (IAPS) in the study of emotion and attention.," in *Handbook of emotion elicitation and assessement*, J. A. Coan and J. J. B. Allen, Eds. Oxford University Press, 2007, pp. 29–46.

[72] B. Verschuere, G. Crombez, and E. Koster, "The International Affective Picture System: A Flemish validation study.," *Psychol. Belg.*, vol. 41, no. 4, pp. 205–17, 2001.

[73] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the International Affective Picture System.," *Behav. Res. Methods*, vol. 37, no. 4, pp. 626–630, 2005.

[74] M. Dufey, A. M. Fernández, and R. Mayol, "Adding support to cross-cultural emotional assessment : Validation of the International Affective Picture System in a Chilean sample," *Univ. Psychol.*, vol. 10, no. 2, pp. 521–534, 2011.

[75] G. Valenza, A. Lanata, and E. P. Scilingo, "The Role of Nonlinear Dynamics in Affective Valence and Arousal Recognition," *IEEE Transactions on Affective Computing*, vol. 3, no. 2. pp. 237–249, 2012.

[76] R. J. Davidson, “Affective neuroscience and psychophysiology: toward a synthesis.,” *Psychophysiology*, vol. 40, no. 5, pp. 655–665, 2003.

[77] E. T. Rolls, “Précis of The brain and emotion.,” *Behav. Brain Sci.*, vol. 23, no. 2, pp. 177–191; discussion 192–233, 2000.

[78] P. Gable and E. Harmon-Jones, “Relative left frontal activation to appetitive stimuli: considering the role of individual differences,” *Psychophysiology*, vol. 45, no. 2, pp. 275–278, 2008.

[79] H. W. Lilliefors, “On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown,” *J. Am. Stat. Assoc.*, vol. 62, no. 318, pp. 399–402, 1967.

[80] W. Heller, “Neuropsychological mechanisms of individual differences in emotion, personality, and arousal.,” *Neuropsychology*, vol. 7, no. 4. pp. 476–489, 1993.

[81] D. R. Collins, J. G. Pelletier, and D. Paré, “Slow and fast (gamma) neuronal oscillations in the perirhinal cortex and lateral amygdala.,” *J. Neurophysiol.*, vol. 85, no. 4, pp. 1661–1672, 2001.

[82] Q. Luo, T. Holroyd, M. Jones, T. Hendler, and J. Blair, “Neural dynamics for facial threat processing as revealed by gamma band synchronization using MEG,” *Neuroimage*, vol. 34, no. 2, pp. 839–847, 2007.

[83] A. Matsumoto, Y. Ichikawa, N. Kanayama, H. Ohira, and T. Iidaka, “Gamma band activity and its synchronization reflect the dysfunctional emotional processing in alexithymic persons.,” *Psychophysiology*, vol. 43, no. 6, pp. 533–540, 2006.

[84] L. Sebastiani, A. Simoni, A. Gemignani, B. Ghelarducci, and E. L. Santarcangelo, "Human hypnosis: Autonomic and electroencephalographic correlates of a guided multimodal cognitive-emotional imagery," *Neurosci. Lett.*, vol. 338, no. 1, pp. 41–44, 2003.

[85] M. M. Muller, A. Keil, T. Gruber, and T. Elbert, "Processing of affective pictures modulates right-hemispheric gamma band EEG activity," *Clin. Neurophysiol.*, vol. 110, no. 11, pp. 1913–1920, 1999.

[86] J. C. Borod, "Interhemispheric and intrahemispheric control of emotion: a focus on unilateral brain damage.," *J. Consult. Clin. Psychol.*, vol. 60, no. 3, pp. 339–348, 1992.

[87] F. Burkhardt, a Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," *Ninth Eur. Conf. Speech Commun. Technol.*, vol. 2005, pp. 3–6, 2005.

[88] N. Amir, S. Ron, and N. Laor, "Analysis of an emotional speech corpus in Hebrew based on objective criteria," in *ISCA workshop on speech and emotion*, 2000, pp. 29–33.

[89] D. Schwarz, B. Grégory, V. Bruno, and B. Sam, "Real-time corpus-based concatenative synthesis with catart," *Proc. Int. Conf. ...*, no. September, pp. 1–7, 2006.

[90] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker Independent Speech Emotion Recognition by Ensemble Classification," *2005 IEEE Int. Conf. Multimed. Expo*, 2005.

[91] H. Hu, M.-X. Xu, and W. Wu, "GMM Supervector Based SVM with Spectral Features for Speech Emotion Recognition," *2007 IEEE Int. Conf. Acoust. Speech Signal Process. - ICASSP '07*, vol. 4, 2007.

[92] V. Hozjan, K. Zdravko, M. Asuncion, B. Antonio, and N. Albino, "Interface Databases: Design and Collection of a Multilingual Emotional Speech Database," in *LREC*, 2002, pp. 2024–2028.

[93] I. S. Engberg, A. V Hansen, O. Andersen, and P. Dalsgaard, "Design, Recording and Verification of a Danish Emotional Speech Database," in *Proceedings of Eurospeech 1997*, 1997, vol. 4, pp. 1695–1698.

[94] L. A. Schmidt and L. J. Trainor, "Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions," *Cognition & Emotion*, vol. 15, no. 4. pp. 487–500, 2001.

[95] N. Maglaveras, T. Stamkopoulos, K. Diamantaras, C. Pappas, and M. Strintzis, "ECG pattern recognition and classification using non-linear transformations and neural networks: A review," in *International Journal of Medical Informatics*, 1998, vol. 52, no. 1–3, pp. 191–208.

[96] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System," in *Affective dialogue systems*, vol. i, Springer Berlin Heidelberg, 2004, pp. 36–48.

[97] B. Cheng and G. Liu, "Emotion recognition from surface EMG signal using wavelet transform and neural network," *The 2nd International Conference on Bioinformatics and Biomedical Engineering*, 2008. .

[98] R. A. McFarland, “Relationship of skin temperature changes to the emotions accompanying music.,” *Biofeedback Self. Regul.*, vol. 10, no. 3, pp. 255–267, 1985.

[99] D. O. Bos, “EEG-based Emotion Recognition - The Influence of Visual and Auditory Stimuli,” *Emotion*, vol. 57, pp. 1798–806, 2006.

[100] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, “Multimodal emotion recognition from expressive faces, body gestures and speech,” in *IFIP International Federation for Information Processing*, 2007, vol. 247, pp. 375–388.

[101] A. J. Gerber, J. Posner, D. Gorman, T. Colibazzi, S. Yu, Z. Wang, A. Kangarlu, H. Zhu, J. Russell, and B. S. Peterson, “An affective circumplex model of neural systems subserving valence, arousal, and cognitive overlay during the appraisal of emotional faces,” *Neuropsychologia*, vol. 46, pp. 2129–2139, 2008.

[102] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, “A dimensional approach to emotion recognition of speech from movies,” *2009 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2009.

[103] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, “Physiological signals based human emotion Recognition: a review,” *2011 IEEE 7th Int. Colloq. Signal Process. its Appl.*, pp. 410–415, 2011.

[104] K. Jonghwa and E. Andre, “Emotion recognition based on physiological changes in music listening,” *Pattern Anal. Mach. Intell. IEEE Trans.*, vol. 30, pp. 2067–2083, 2008.

[105] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 10, 2001.

[106] O. Pierre-Yves, "The production and recognition of emotions in speech: Features and algorithms," *Int. J. Hum. Comput. Stud.*, vol. 59, no. 1–2, pp. 157–183, 2003.

[107] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *2003 Int. Conf. Multimed. Expo. ICME '03. Proc. (Cat. No.03TH8698)*, vol. 1, 2003.

[108] C. M. L. C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, 2005.

[109] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Improved Emotion Recognition With a Novel Speaker-Independent Feature," *IEEE/ASME Trans. Mechatronics*, vol. 14, no. 3, 2009.

[110] D. Ververidis and C. Kotropoulos, "Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm," *2005 IEEE Int. Conf. Multimed. Expo*, pp. 1500–1503, 2005.

[111] T. Vogt and E. Andre, "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition," *2005 IEEE Int. Conf. Multimed. Expo*, 2005.

[112] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Speech Emotion Recognition Using Eigen-FFT in Clean and Noisy Environments," *RO-MAN 2007 - 16th IEEE Int. Symp. Robot Hum. Interact. Commun.*, 2007.

[113] J. H. Yeh, T. L. Pao, C. Y. Lin, Y. W. Tsai, and Y. Te Chen, "Segment-based emotion recognition from continuous Mandarin Chinese speech," in *Computers in Human Behavior*, 2011, vol. 27, no. 5, pp. 1545–1552.

[114] B. Schuller and G. Rigoll, "Timing Levels in Segment-Based Speech Emotion Recognition," in *Proc. INTERSPEECH 2006*, 2006.

[115] D. Morrison and L. C. De Silva, "Voting ensembles for spoken affect classification," *J. Netw. Comput. Appl.*, vol. 30, no. 4, pp. 1356–1365, 2007.

[116] Y. Yujin, Z. Peihua, and Z. Qun, "Research of speaker recognition based on combination of LPCC and MFCC," *Intell. Comput. Intell. Syst. (ICIS), 2010 IEEE Int. Conf.*, vol. 3, 2010.

[117] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion Recognition in Speech Using Neural Networks," *Neural Comput. Appl.*, vol. 9, no. 4, pp. 290–296, 2000.

[118] S. H. Hemenover, "Individual differences in rate of affect change: studies in affective chronometry.," *J. Pers. Soc. Psychol.*, vol. 85, no. 1, pp. 121–131, 2003.

[119] L. G. Eaton and D. C. Funder, “Emotional experience in daily life: valence, variability, and rate of change.,” *Emotion*, vol. 1, no. 4, pp. 413–421, 2001.

[120] S. C. Marsella and J. Gratch, “EMA: A process model of appraisal dynamics,” *Cogn. Syst. Res.*, vol. 10, no. 1, pp. 70–90, 2009.

[121] J. Ye, Y. Li, L. Wei, Y. Tang, and J. Wang, “The Race Effect on the Emotion-Induced Gamma Oscillation in The EEG,” *2009 2nd Int. Conf. Biomed. Eng. Informatics*, pp. 1–4, 2009.

[122] M. T. Shami and M. S. Kamel, “Segment-based approach to the recognition of emotions in speech,” *2005 IEEE Int. Conf. Multimed. Expo*, 2005.

[123] M. Shuzo, T. Yamamoto, M. Shimura, F. Monma, S. Mitsuyoshi, and I. Yamada, “Construction of Natural Voice Database for Analysis of Emotion and Feeling,” *J. Inf. Process.*, vol. 53, no. 3, pp. 1185–1194, 2011.

[124] C. Shannon, “A mathematical theory of communication,” *ACM SIGMOBILE ...*, vol. 27, no. July 1948, pp. 379–423, 1948.

[125] E. Jaynes, “Information Theory and Statistical Mechanics,” *Physical Review*, vol. 106, no. 4. pp. 620–630, 1957.

[126] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, “The mutual information: detecting and evaluating dependencies between variables.,” *Bioinformatics*, vol. 18 Suppl 2, pp. S231–S240, 2002.

[127] R. Battiti, “Using mutual information for selecting features in supervised neural net learning.,” *IEEE Trans. Neural Netw.*, vol. 5, no. 4, pp. 537–550, 1994.

ACKNOWLEDGEMENTS

The content of the thesis is the summary of the research I conducted at Yamada-Warisawa Laboratory in the University of Tokyo from 2011 to 2014. Professor Ichiro Yamada, who is my supervisor, provided me lots of extremely valuable guidance and support on my PhD study and research. His meticulous scholarship impresses me and I have much to learn from it. Associate professor Shin'ich Warisawa helped me a lot in finalizing this research and gave many advices for writing thesis and making presentation materials.

Professor Naoki Mukawa, Professor Hideyoshi Yanagisawa, Professor Masamichi Shimosaka provides many helpfulness comments and suggestions for finalizing this thesis.

Professor Kazutaka Ueda coached me for better understanding EEG related knowledge. He also provided many useful advices for my experiment design and checked the experimental procedure.

Previous assistant professor Guillaume Lopez provided lots of practical help in discussing the research topic and proceeding my experiments. Dr. Shunji Mitsuyoshi, who is an expert in emotion recognition field, shared lots of his valuable time for having meeting in order to discuss emotion recognition related

issues. I am also very impressed by his energetic altitude towards research and life.

Professor Pierre Maret invited me to his laboratory in France as a visiting student and discussed many detailed aspects for emotion recognition using speech. Professor Fabrice Muhlenbach, who is an expert in data mining, gave lots of advices about emotion recognition techniques.

All the students in the laboratory provide me lots of support both in research and daily life. In addition, I received lots of help from all the participants to complete the experiments for collecting variety of signals.

Last but not the least, I would like to mention that I received great support from all my family, I cannot accomplish my PhD study and research without their understanding and support.

I would like to express my sincere gratitude to the people I mentioned above and all other people who helped this research and thesis.