Cognitive-inspired speaker affective state profiling

Norhaslinda Kamaruddin

2012

Kamaruddin, N. (2012). Cognitive-inspired speaker affective state profiling. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/51057>

<https://doi.org/10.32657/10356/51057>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

**COGNITIVE-INSPIRED SPEAKER AFFECTIVE
STATE PROFILING**

**NORHASLINDA KAMARUDDIN
SCHOOL OF COMPUTER ENGINEERING
2012**

COGNITIVE-INSPIRED SPEAKER AFFECTIVE STATE PROFILING

NORHASLINDA KAMARUDDIN

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2012

To my parents

With this thesis I fulfill my promise

ACKNOWLEDGEMENTS

PhD is an arduous and challenging journey. And yet, it offers exciting experience and exposes me to new knowledge that enriches my academic path as well as my life. With humble realization, it was hardly possible for me to thrive in my doctoral studies without the generous help, support and encouragement given by these personalities who played a major part in ensuring the completion of this PhD work.

First and foremost, I would like to express my deepest gratitude to my main supervisor, *Prof. Dr. Abdul Wahab Abdul Rahman* who goes beyond his call of duty in giving guidance, inspiring me to explore deeper understanding of knowledge work and think outside the box. His objective-oriented approach in supervising me and most importantly his unwavering belief that I can do it works fantastically well which allows me to soar beyond my limits.

Secondly, I would like to thank my second supervisor, *Assoc. Prof. Dr. Quek Hiok Chai* for his willingness to step in to be my supervisor when Prof Wahab took the Professorship appointment in IIUM. His constructive comments and firm guidance really helps in improving my writing quality and taught me to appreciate the good scientific work.

Then, I would like to extend my thanks to my husband, *Mohd Khairudin Mohd Nor* and my beautiful princesses, *Nurnabila Farhana* and *Nurathia Qistina* for bearing up with me through thick and thin and my reliable support system ever, my parents, *Zaleha Ahmad* and *Kamaruddin Muhammad*, for their unconditional love. Subsequently, to my families and friends, thank you for being able to lend your hands, ears and hearts in time of need. Last but not least, a special thank goes to my angel in disguise for always supporting me in every imaginable ways. Believe me, without all of you, it would not been possible. Thank you very much.

Norhaslinda Kamaruddin
Center for Computational Intelligence, NTU

ABSTRACTS

Human behavior is influenced by emotion and human expressed affective state through numerous channels using non-verbal communication; namely: facial expression, gestures, eye-gazing, body postures, as well as verbal communication. In verbal communication itself, there are lots of underlying information transmitted using acoustical features and the semantic meaning of the word/sentence used. Despite the evident complexity of such interaction, listener still can correctly perceive the propagated emotion conveyed by the interlocutor. This is due to the human cognitive functional ability to dissect and infer the information with high accuracy and then react accordingly with appropriate behavioral responses and feedbacks. Hence, this research work introduces novel technique in discriminating emotion to facilitate the understanding of speaker affective state, based on the hypothesis that emotion is propagated through speech and it can be quantified.

Speech emotion is a growing multi-disciplinary research field and is gaining greater momentum due to the increased need to improve on the quality of human computer interaction. Numerous researchers apply various feature extraction methods coupled with classifiers to produce acceptable accuracy performance. Nonetheless, the performance of such a system is bound to cultural influence which resulted in unpromising outcome once an unknown culture-influenced speech is introduced. Culture is always regarded as a trivial and inconsequential parameter that heeds minimal consideration in speech emotion recognition. Hence, in this work, the intricate relationship of cultural influence in term of intra-cultural and inter-cultural effects is studied in details. Two speech emotion datasets; of the NTU_American and NTU_Asian dataset representing the American and Asian culture influence to speech emotion respectively were collected and together with the standard Berlin speech emotion dataset were used to understand the speech emotion recognition system and the culture bias.

The work is then extended to investigate speaker affective state profiling using the Valence-Arousal (VA) analysis approach that enables visualization tool to be utilized for intra-cultural and inter-cultural assessments. The strength of this VA approach is that it is able to facilitate the observation of new finding as well as catering to dynamic data-driven affective space model generation that is able to empirically verify the psychologists' agreement of the affective space model. This proposed approach is developed to complement the discrete-class classification system that is rigid and lacking the explainable components. The result shows huge potential for future practical applications of such analysis system; which enables researchers, engineers, scientists, psychologists, medical practitioner as well as intelligent system developer to visualize emotions from a common view point.

TABLE OF CONTENTS

<i>ACKNOWLEDGEMENTS</i>	<i>ii</i>
<i>ABSTRACT</i>	<i>iii</i>
<i>TABLE OF CONTENTS</i>	<i>iv</i>
<i>LIST OF FIGURES</i>	<i>ix</i>
<i>LIST OF TABLES</i>	<i>xii</i>
CHAPTER 1: INTRODUCTION	1
1.1 Overview.....	1
1.2 Speech Emotion Recognition Challenges, Issues and Expectations.....	7
1.3 Research Motivations	11
1.4 Research Objectives	13
1.5 Phases of the Research Framework	14
1.8 Research Hypotheses	19
1.7 Research Contributions	19
1.8 Organization of the Thesis.....	24
CHAPTER 2: LITERATURE REVIEW	25
2.1 Psychologists' Theoretical Perspectives on Emotion.....	26
2.1.1 Relational Model of Emotion.....	27
2.1.2 Functionalist Model of Emotion.....	28
2.1.3 Social Constructivist Model of Emotion.....	30
2.1.4 Psychological Perspective of the Thesis.....	31
2.2 Archetype of Emotion.....	34
2.2.1 Discrete Basic Emotion.....	34
2.2.2 Multi-dimensional Emotion Model.....	37
2.3 Related Works in Speech Emotion Research.....	40
2.4 On-going Cross Cultural Speech Emotion Research.....	52

2.4.1	Cultural Effect of Emotion	52
2.4.2	Intra and Inter Cultural Assessment	54
2.4.3	3D Affective Space Model	55
2.5	Summary.....	60
CHAPTER 3: DATA CORPUSES, FEATURES EXTRACTIONS METHODS AND CLASSIFIERS.....		63
3.1	Speech Emotion Annotation and Data Corpus.....	67
3.1.1	The NTU_American Dataset.....	68
3.1.2	The NTU_Asian Dataset.....	74
3.1.3	The Berlin Database of Emotional Speech (Emo-db).....	78
3.2	Features Extraction Method.....	79
3.2.1	Mel Frequency Cepstral Coefficient (MFCC).....	81
3.2.2	Gaussian Mixture Model GMM).....	87
3.3	Classifiers	92
3.3.1	Multi-Layer Perceptron (MLP).....	94
3.3.2	Adaptive Network-based Fuzzy Inference System (ANFIS).....	96
3.3.3	Generic Self-organizing Fuzzy Neural Network (GenSoFNN)..	101
3.4	Summary	103
CHAPTER 4: SPEECH EMOTION VERIFICATION: A PRELIMANARY WORK.....		105
4.1	Feature Selection.....	106
4.1.1	MFCC Feature Selection using Automatic Relevance Determination (ARD).....	106
4.1.2	GMM Features Selection.....	112
4.2	Speech Emotion Verification.....	116
4.2.1	Experimental Set-up for Speech Emotion Verification.....	117
4.2.2	Classifiers Network Model Selection.....	119

4.2.3	Speech Emotion Verification Result and Discussion.....	127
4.2.4	Culture Influence in Speech Emotion Verification.....	130
4.3	Summary and Conclusion.....	133
CHAPTER 5: CULTURAL DEPENDENCY ANALYSIS FOR UNDERSTANDING SPEECH EMOTION.....		135
5.1	Experimental Set-up for Intra-cultural and Inter-cultural Assessment.....	136
5.1.1	Data Structuring for Different Types of Experiments	138
5.2	Speech Emotion Identification Result and Discussion.....	144
5.2.1	Speech Emotion Identification Result for Homogeneous Intra-cultural Assessment.....	145
5.2.2	Speech Emotion Identification Result for Heterogeneous Mixed-cultural Assessment.....	149
5.2.3	Speech Emotion Identification Result for Inter-cultural Assessment.....	155
5.2.4	Observation.....	161
5.3	Two-stage Culture-sensitive Speech Emotion Approach for Culture Detection.....	162
5.3.1	Culture Detection Experimental Result and Discussion.....	164
5.4	Summary and Conclusion.....	166
CHAPTER 6: AFFECTIVE SPACE MODEL APPROACH FOR SPEECH EMOTION UNDERSTANDING.....		170
6.1	VA Approach based on Affective Space Model Understanding.....	175
6.1.1	Generation of the Affective Space Model Origin and Quadrant based on Neutral Recalibration	183
6.1.2	Generation of the Affective Space Model Neutral-Emotion Boundary in a Specified Quadrant.....	187
6.1.3	Experimental Set-up and Data Structures for VA Approach.....	190

6.1.4	Speech Emotion Identification Result and Discussion using the Proposed VA Approach.....	191
6.1.5	Summary of VA Approach Applied on Intra-cultural and Inter-cultural Speech Emotion Recognition Task.....	203
CHAPTER 7: SUMMARY, CONCLUSION AND FUTURE WORK		206
7.1	Future Work and Potential Applications.....	210
7.2	Conclusions.....	211
REFERENCES		213
LISTS OF PUBLICATIONS		228
APPENDIX AI		232
APPENDIX AII		233
APPENDIX AIII		234

LIST OF FIGURES

Figure 1.1	A Brunswikian Lens of the Vocal Communication of Emotion	5
Figure 1.2	The Integrated Six Phases Framework of the Research	14
Figure 2.1	The Psychological Perspectives of Emotion	32
Figure 2.2	Plutchik's Emotions Color Wheel	36
Figure 2.3	Four Different Basic Emotion Quadrants of the Affective Space Model	38
Figure 2.4	Dimensional Labeling in FeelTrace Software	39
Figure 2.5	Different Approaches of Generating the Speech Emotional Corpora	40
Figure 2.6	Co-variance Plot for the Emotion Classes	56
Figure 2.7	Timeline of Selected Works in Speech Emotion Recognition	59
Figure 3.1	The General Architecture of the Proposed Speech Emotion Recognition System	64
Figure 3.2	The Summary of Data Collection Processes of NTU_American Dataset	69
Figure 3.3	The Detailed Ethnicity Fragmentation for Different Judges' Group in the Heterogeneous Human Perception Test	71
Figure 3.4	The Schematic Diagram of Cochlea	83
Figure 3.5	Anatomy of Human Ear	83
Figure 3.6	The Block Diagram of MFCC Features Extraction Method	85
Figure 3.7	Example of GMM Implementation	91
Figure 3.8	A Multi Layer Perceptron Network Architecture	94
Figure 3.9	The ANFIS Architecture	97
Figure 3.10	Block Diagram of the N-class mANFIS.	100
Figure 3.11	The GenSoFNN Architecture	102
Figure 4.1	Different Individual MFCC Feature Arrangements for NTU_American Anger Verification	110

Figure 4.2	ARD-MFCC and Standard-MFCC Arrangement Verification Performances	113
Figure 4.3	WEKA-MLP NTU_American Anger Verification Result	116
Figure 4.4	The MFCC Verification Training and Testing Class Data Organization	118
Figure 4.5	GMM-MLP Model Selection Local Validation Training Time based on NTU_American Anger Verification	122
Figure 4.6	Different Accuracy Performance Obtained by Different Radii Value for Different NTU_American Emotion Verification using MFCC Features	124
Figure 4.7	Different Accuracy Performance Obtained By Different Radii Value for NTU_American Emotion Verification Using MFCC Features	125
Figure 4.8	Emotion Verification Result Using Different Features Extraction	128
Figure 4.8	Different Plots for Anger and Neutral of NTU_American Dataset	132
Figure 5.1	Block Diagram of Discrete-Class Classification System Approach	137
Figure 5.2	Summary of Experimental Design for the Different Experimental Set-up	139
Figure 5.3	Block Diagram for the Different Experimental Set-up	142
Figure 5.4	Summary for Generalized Homogeneous Intra-cultural Experimental Result using Different Dataset and Classifier Pairing.	148
Figure 5.5	Generalized Heterogeneous (80/20) Mixed-Cultural Assessment Result for Different Dataset Combination	152
Figure 5.6	Result Comparison for Overall Accuracy Performance of 80/20 and 50/50 Rule Generalized Heterogeneous Mixed-Cultural Assessment	154
Figure 5.7	Radar Chart of the Performance Accuracy for Homogeneous Inter-Cultural Experiment with Different Training-Testing Dataset Pair	157
Figure 5.8	Radar Chart of the Performance Accuracy for Heterogeneous Mixed-Cultural Experiment with Different Training-Testing Dataset Pair	159

Figure 5.9	The Summary of Accuracy Performance for Different Cultural Dependency Assessment.	161
Figure 5.10	Block Diagram of the Proposed Two-Stage Culture-Sensitive Speech Emotion Approach	163
Figure 5.11	Experimental Results for Culture Detection for Different Heterogeneous Mixed Cultural	166
Figure 6.1	Block Diagram of Different Classification Approach	176
Figure 6.2	The Transformation Process of Speech Emotion Signal to VA Analysis	177
Figure 6.3	The Quadrant Labeling and Its Occupied Emotion	178
Figure 6.4	Block Diagram of the VA Approach	178
Figure 6.5	NTU_American Dataset Emotions and its Centroid using ANFIS Classifier	185
Figure 6.6	Generation of Affective Space Model Origin for Berlin Dataset Discriminated by ANFIS Classifier using Different Origin	186
Figure 6.7	Emotion Segregation Example using Soft-Boundary Approach	187
Figure 6.8	NTU_American Dataset with Neutral-Emotion Boundary Using ANFIS	190
Figure 6.9	The Different Emotion Centroid of Homogeneous Intra-cultural Experiment	193
Figure 6.10	The Different Emotion Centroid in Heterogeneous Intra-Cultural Experiment Using Different Classifiers	197
Figure 6.11	The Accuracy Performance for Homogeneous Intra-Cultural Experiment based on VA Analysis	199
Figure 6.12	The Accuracy Performance for Heterogeneous Intra-Cultural Experiment based on VA Approach	201
Figure 6.13	The Summary Accuracy Performance for VA Approach Experiment	204
Figure 7.1	Block Diagram of a Speech Emotion Generalized Application System	209

LIST OF TABLES

Table 2.1	Example of Stimulus Events and Possible Reaction and Effect	37
Table 2.2	Summary of Selected Works in Speech Emotion Recognition System	49
Table 2.3	Summary of Selected Computerized Culturally-influenced Speech Emotion Recognition System	60
Table 3.1	The Confusion Matrix for Heterogeneous Human Perception Performance based on the NTU_American Dataset	73
Table 3.2	The Confusion Matrix for Homogeneous Human Perception Performance based on the NTU_Asian Dataset	76
Table 3.3	The Confusion Matrix for Homogeneous Human Perception Performance based on the Berlin Dataset	78
Table 4.1	Performance Accuracy for Different Window Size for GMM Features Extraction	114
Table 4.2	MFCC-MLP Model Selection Local Validation Training Time	121
Table 4.3	Convergence Computation Time Performance using MFCC Features	129
Table 4.4	Convergence Computation Time Performance using GMM Features	129
Table 4.5	Speech Emotion Verification Result for NTU_American and Berlin Dataset	132
Table 5.1	Accuracy Performance for Localized Homogeneous Intra-cultural Experiment (%)	145
Table 5.2	Accuracy Performance for Generalized Homogeneous Intra-cultural Experiment (%)	146
Table 5.3	Accuracy Performance for Localized (In-sample) Heterogeneous (50/50) Mixed-Cultural Experiment (%)	150
Table 5.4	Accuracy Performance for Generalized (Out-sample) Heterogeneous (50/50) Mixed-Cultural Experiment (%)	150
Table 6.1	Summary for Different Quadrant Bias Based on Neutral Centroid	194

CHAPTER 1

INTRODUCTION

1.1 Overview.

The essence of human social communication is deep-rooted in affection and perception. Human are able to express their affective states and infer their emotional information that is being conveyed during interaction. Variation of emotional states can be used as an essential evaluation criteria to access the efficacy of cognition processing particularly for action tendency [Frij87] and decision making [Dama94]. Emotion is important in displaying and regulating patterns of behaviour and maintaining the homeostatic balance in personal well-being. Emotion also plays a significant role in influencing motivation and focus of attention. Thus recognizing emotion will be the first step in “*humanizing*” the interactive system.

Correlation between action tendency and emotion is explained by the psychologists who support emotion appraisal theory [Laza91, Frij87, Arno60]. Emotions or affective states determine the impulse to act in certain ways that is appropriate to the socially relevant information. The actions are generally understood as a reaction or initial feedback to an emotional response rather than involuntary. The tendency and readiness to act involves both bodily arousal and psychological preparation which corresponds to different types of emotion perceived. Essentially, emotional appraisal induced either positive or negative outcome that caused specific reactions in different people. For instance, an angry driver may experience the desire to lash out at some target upon encountering some obscene gesture by another driver while driving. On the other hand, if such scenario is perceived as childish, that particular individual may brush off the incident as trivial and keep on driving calmly. However, an action (in this example,

hitting out at some target) may not necessarily have been committed or realized because human can control the urge to act based on his/her better judgement. This self-control ability is due to the decision-making capability in human cognitive processing system in which emotion plays a significant role in influencing the decision making process.

Decision-making is one of the crucial cognitive activities that include processes of selecting a preferred course of action as a reaction from several sets of alternatives [Naqv06, Bech04]. Although logical decision-making process seems to exclude emotion, human tend to incorporate a certain degree of emotional influence to help them decide. This finding was also reported by Damasio [Dama94]. Decision-making ability can be easily impaired for people who had suffered brain injuries which had affected their emotional ability. The patients' capability to make correct decision diminished due to their inability to access the emotion information although they are able to think logically and intellectually. As an extension to the work by Damasio [Dama94], Bechara and Damasio coupled the Iowa Gambling Task with the skin-conductance response (SCR) to see the impact of elicited bodily states while choosing between advantageous and disadvantageous decks [Bech05]. Emotion arousal can be physiologically indexed by measuring the degree of the skin's electrical conductance based on the variation of perspiration level. The experimental result shows that people with some degree of brain injury always repeat the selection of disadvantageous deck insistently. The patients seemed to be insensitive to the negative consequences of their choice as though they are unable to learn from the previous mistakes. Normal people however would sample equally between the advantageous and disadvantageous decks at the initial stage. After 40 and 50 deck selections, they would learn the benefit of choosing the advantageous deck and keep doing so. This shows that emotion plays an important role in influencing rational decision-making process.

Emotion propagation can be executed in many ways as human converse to one another. The verbal and non-verbal information are freely transmitted through multiple sources of communicative channels that are strongly connected and systematically synchronized. Speech acoustical features such as intonation, speech rate and pitch, as well as facial expression, body language, eye gazing and gestures are some appropriate examples of such non-verbal communicative channels. Subsequently, the linguistic and semantic content of the speech are also used to complement the rich volume of information that is being transferred. Such combinations help to ensure the correctness of the conveyed message to be encoded by the listener. Nonetheless, in most cases, emotion can still be transmitted using single channel although it may not be as effective as multimodal communication. For instance, an infant can detect the mother's feeling just by listening to her voice even when the child's vision capability is not fully developed. One can also recognize the affection portrayed by pantomime actors who use only facial expressions, body language and gestures in their performance. In addition, a traveler with limited native language proficiency can understand the direction pointed to by the local person using gestures even when the traveler cannot comprehend the semantic contents of the speech. Hence, it can be summarized that affective state can be disseminated by various channels towards the convergence and optimization of the interactive process. However the focus of this research work is on understanding and analyzing speech emotion based on the affective state modeling and profiling using the speech cognition process.

Unlike emotion detection from facial expression that has been extensively studied dating back to Darwin [Darw72], speech emotion research is relatively new. Over the decades, speech emotion research has attracted a growing number of researchers to study and understand the multi-disciplinary nature of speech affection and perception. The

findings have been discussed in detail to some extent that each domain gives its own contributions. Neuroscientists provides analysis and understanding of ways the brain perceive the input stimuli by processing activities and applying the knowledge stored in the brain to be reflected in various actions and feedbacks. Psychologists examine emotion from a psychological perspective by treating them as mental processes and behaviours. Linguistic field contributed to the body of research by viewing speech as an input that provides emotional knowledge through semantic and syntactic components of the speech. Researchers in the area of digital signal processing provide an understanding of acoustic characteristic for features to be classified. Subsequently, computational intelligence researchers provide tools to empirically verify the knowledge gained from the various fields through mathematical solution using machine learning and data mining. These combinations of inter-disciplinary research efforts have shown tremendous potential in understanding and analyzing the speech emotion for practical implementation and applications. It is hoped that by incorporating speech emotion recognition ability in interactive system, the performance of human machine interaction systems can be further improved to be more human-like interaction and hence easily adopted by the end users.

Human speech carries information about both the linguistic content as well as the current emotional or attitudinal state of the speaker. Scherer explicitly explained the speech emotion inference processes using the modified Brunswik's functional model of perception [Sche78, Sche03]. Figure 1.1 represents the concept of speech emotion perception using the respective model. The process begins with encoding the expression of the interlocutor emotional states through certain voice and speech characteristics that can be objectively measured from the speech signals. This is because the emotional arousal from speech is always accompanied by physiological changes that affect respiration, phonation and articulation in such a way that produced emotion-specific

patterns of acoustic parameters [Sche86]. The respiratory system, larynx, vocal tract, muscular system, heart rate as well as blood pressure changes the pattern of acoustical features with respect to the different emotional states. The information of the affective states are transmitted to the listener and perceived via the auditory perceptual system using the following processes. Firstly, the listener's cochlea accepts the signal which activates the basilar membrane vibrations. These motions excite the inner hair cells to extract important features for neural decoding. Selected features are then fired through the auditory neuron to the auditory cortex in the brain. Auditory cortex processed the features by decoding the information to infer the interlocutor affective state so that appropriate response and feedback can be formulated and subsequently elicited as a reaction.

Figure 1.1: A Brunswikian Lens of the Vocal Communication of Emotion [Sche78]

Based on this concept of the modified Brunswikian Lens illustrated in Figure 1.1, the general approach in developing the speech emotion recognition system is proposed. The proposed system consists of three main modules; namely: data collection, features extraction and classification. These three modules are important to ensure the success of

the proposed system. The aim of building such a system is to distinguish the emotion of the unseen speech signals (test data) based on the training information gathered. Detailed explanations of the proposed speech emotion recognition system will be discussed in the thesis.

Extracting emotional information from the speech signals can be universally carried out based on the assumption that people with wide varying cultural background are able to recognize emotion conveyed regardless of the gender, linguistic ability or geographical boundary. It was reported that listeners from different cultures can discriminate emotion from speech with recognition accuracy scores that is much higher than chance probability in a forced-choice format [Elfe02, Sche01a]. These finding indicates that there are common tendencies in the acoustical features that correlates basic emotion across different cultures. On the contrary, other researchers show a clear distinction on how emotion is conveyed and perceived by each culture. This information is exclusively shared among people of the same culture and translated to higher degree of accuracy for generalized homogeneous intra-cultural speech emotion identification [Dewa08, Jusl03, Bezo83]. For instance, Dewaele found that different cultures expressed their love differently from one culture to another [Dewa08]. Hence, these mixed observations can be summarised as emotion is universal to a certain extent that it allows people across culture to correctly discriminate emotion. In addition, people within the same culture shares a unique implicit emotional information that can facilitate them to better identify speech emotion. Such scenario only can be achieved if they are given similar culturally-influenced speech emotion in their interaction. Thus, understanding cultural effect on speech is important to unravel this intricate relationship between culture and emotion perception so that the effect of culture can be minimized as an exiguous

parameter that may hinder good performance for a global speech emotion recognition system.

Subsequently, speech emotion profiling is proposed in this research as an alternative approach to tackle two issues; namely: 1) to discriminate emotion and 2) to assess the inter-cultural and intra-cultural relevancy to the speech emotion. However, the latter issue has not been well explored to date. It is imperative for the speech emotion community to acknowledge the effect of culture in speech emotion since culture is known to have significant impact on human social interaction in both the conveyance and perception of the emotion. Hence, it is envisaged that the findings obtained would help boost the understanding of speech emotion and enhance the accuracy of such a system for practical implementation with human-like interaction.

1.2 Speech Emotion Recognition Challenges, Issues and Expectations.

The importance of integrated emotion recognition in an interactive system has grown with the increased demands for human involvement especially in service, education and gaming industry. In fact, the degree of user satisfaction is directly proportional to the amount of interaction that the users have with the system. If the users are satisfied, they will frequently access the system. On the other hand, the unsatisfied users will reduce the frequency of accessing such systems unless otherwise necessary. Hence, developing an interactive system today would require more effective ways for human to engagingly communicate and interact with machines. Most of these efforts focus on enhancing naturalness of human-machine interaction so that users can gain the full benefit from the efficient way by which knowledge is communicated.

Traditionally, disembodied artificial human voice was employed as the main response mechanism of current interactive system. ‘*Command-like*’ nature of the interaction between application users giving instruction for the system to execute commands does not create a palpable bonding between human and the respective system. Although the communication is fast, simple, straightforward and able to get things done, user starts to demand more natural and human-like interaction with machine as the complexity and dependency towards applications increase. With the current advancement in speech synthesis technology, a highly intelligible and human-like speech can now be produced. Major progress in signal processing technology also contributed to the speech emotion field by allowing the scientists and engineers to analyze a massive amount of speech emotion data with greater accuracy. Despite tremendous effort and improvement over the years, current technology are still far from reaching a comprehensive explanation of human spoken language behavior and are unable to create a speech technology that can perform better or on par with human ability. Such speech production system cannot accurately imitate the human speech mechanism due to the limitation in natural emotional responses. This lack of intuitive approach provides an unnatural way to communicate effectively, thus resulting in irritation, false illusion and discouragement that further lead to the users’ frustration and dissatisfaction. Hence it is imperative that such speech synthesis system should be able to closely mimic the human social interaction since it is known that human-to-human communication is the most effective.

The two main factors that need to be addressed before speech emotion can be fully utilized are as follows:

- 1) The developed interactive system must accommodate the speech emotion in an appropriate manner to improve the naturalness and efficiency of human machine interfaces, and

- 2) The reliability of speech emotion recognition system can be improved by mimicking human cognitive approach.

In addition, computation time and complexity factor, limitation of appropriate training data, environment-influenced constraints such as language and culture as well as user expectation are also key determinants that influences the performance of speech emotion recognition system.

Accuracy is the focus of most researchers in speech emotion research while minimal effort is spent on addressing the importance of computation time and complexity. However, practicality of the speech emotion recognition system relies heavily in these two factors. Computation time is defined as the length of time required to perform the training and testing operation for such system to map the non-linear pattern of speech emotion. Complexity refers to the inherent difficulty of the different emotions segregation task in speech so that appropriate action can be articulated as a response. Both factors are intricately correlated in such a way that the increase in complexity may cause the expansion of computation time and vice versa. In order to cater multitudinous of the users' needs, accuracy needs to be considered together with computation time and complexity factor. For instance, real-time applications need fast computation time with acceptable degree of accuracy. Call-in center, intelligent tutoring system and virtual guide are some example of applications that may gain full benefit with this characteristic. Users would assume that the system is able to detect the underlying affective state of the speaker almost automatically with a fairly reliable and accurate result. Nonetheless, some applications are able to tolerate longer response time with very high accuracy; namely: a lie-detector system, psychology analytical tool to assist patient with social impairment and disability as well as believable speech synthesizer for entertainment and gaming

purposes. Hence, a compromise to balance these two trades off is needed to ensure the developed system meet the users needs and expectations.

The limited availability of appropriate training data with suitable emotion annotation labels posed as a major challenge to an effective development of the speech emotion recognition system. Different perception of the speech signals due to different environmental background and exposure such as culture and language complicates the generalization of speech emotion understanding. Hence, it is not surprising that most researchers in speech emotion tend to ignore the effect of culture and language in speech emotion. This notion is based on the hypothesis that emotion is universal that human from different background, language barrier as well as geography boundary can understand the emotion conveyed by the interlocutor. Contradictory to such assumption, it was noted that intra-cultural speech emotion recognition system performance was reported higher compared to inter-cultural speech emotion system performance [Jusl03]. Thus, it is important that the training speech emotion data reflects the universality of the emotion and at the same time are able to provide distinct characteristic of one culture variation to another. This is to ensure that the proposed system is robust and well generalized for implementation.

Subsequently, user expectation for the speech emotion recognition system determines the acceptance of such a system. User expects to have reasonable degree of accuracy within tolerable computation time regardless of the type of applications used. The user also expects the system to be fully dependant on user's perception of the system's intelligence and responsiveness. Repeatability and consistency of the system plays an important role in ensuring user confidence and at the same time would allow the user to adapt to the technology. Thus, there is a high likelihood that the system may be rejected if the system does not match the users' expectation.

There are many factors contributing to the success of a speech emotion recognition system. Due to the complexity of the task, the culture issue will be studied and analyzed in detail. Intra-cultural and inter-cultural assessment and analysis of the speech emotion will be comprehensively discussed and presented in the thesis.

1.3 Research Motivation.

Emotion is much related to the wellness of brain functionality. In fact, many of brain degenerative diseases and brain disorder; such as Parkinson, Autism and Attention Deficit Hyperactivity Disorder (ADHD), show a decline in patients' emotion conveyance and perception ability [DaFo09, Pell08, Bege06, Weis06, Love97]. Speech emotion research can provide early detection of such symptoms by measuring the emotional characteristic of the patients' speech and communication. To date, researchers prefer to use Electroencephalography (EEG) device to measure the brain signals activity due to its high temporal resolution and limited spatial resolution. In addition, functional Magnetic Resonance Imaging (fMRI) is also preferred by the researchers to study brain activation during task such as motor response inhibition. However, such approaches are more expensive compared to a speech emotion recognition system. This is because in speech emotion recognition system only microphones are required to capture the input data. The patients may also feel overwhelmed or nervous using either the EEG or the fMRI devices due to its procedures. In addition, the EEG and fMRI devices require experts to operate the machine and translate the data. The procedures for fMRI would require patients to be injected with chemical substances which can be harmful to some in the long run, while the EEG patients would need to have multiple probes on the scalp which can be cumbersome. On the other hand speech emotion data collection can be easily achieved by capturing the patients speech through natural interaction. Hence, it would be beneficial if

speech emotion recognition can be used as the front-end diagnostic tools prior to other classical approaches. Such procedures can minimize analysis cost, time and effort. The speech emotion recognition system can also be used to detect any abnormality such as disruptive and antisocial behaviors through natural speech communication. This is to help psychologists and psychiatrists diagnose their patients in a more systematic and natural manner [Jones10, Sche89]. Aggression, depression and poor regulation of impulses are some examples of such behaviors.

Other motivation of this research is the ability to enhance and facilitate human computer interaction (HCI) in a more natural way. Such HCI system must be able to provide adaptation and high degree of responsiveness to different user behavior. In any verbal communication, it is important to acknowledge the underlying emotion of the user so that appropriate behavioral responses and feedbacks can be provided. In many of today's system applications there will be plenty of opportunity to further enhance the end products such as in the following examples:

- i. ***Virtual guide*** [Hein08] - to assist tourist while exploring the places of interest by providing appropriate landmarks and directions and keeping he/she calm when lost.
- ii. ***Intelligent tutoring system*** [Forb11, Eske09, Duch09, Litm06] – by controlling the content and the speed of the tutorial delivered. In addition, based on the student emotion, the system can encourage bored student by capturing his/her attention to learn new knowledge or skill.
- iii. ***Call centers application*** [Morr07, Lee05] – to identify irritated caller so that a particular caller can be directly addressed via a human operator instead of computer-assisted operator. This is to maintain customer satisfaction and loyalty.

- iv. ***Smart homes environment application*** [Kost08, Moll06] - to facilitate changes in the house ambient to suit the mood of the occupier by adapting the room's brightness.
- v. ***Robotic pet*** [Vogt07, Brea03, Mozz00] - as an entertainment that is able to recognize the owner's mood so that the pet can react accordingly; for instance by cheering together with the owner when he/she is in a jubilant mood.

1.4 Research Objectives.

Based on the issues discussed in the previous section, the field of speech emotion offers vast room for research due to its complexity. The scope of this work aims to understand speech emotion in more details hence by developing a tool for speech emotion recognition and analysis. The objectives of the research are to offer alternative solution to speech emotion recognition and analysis by:

1. Mimicking the human brain approach to speech emotion processing by extracting features and processing the information to develop the speech emotion recognition system with reasonable accuracy.
2. Developing affective state profiling tools to enable one to visualize and analyze the profile patterns based on psychologists' understanding of affective space model.
3. Studying the similarity and dissimilarity measurements between intra-cultural and inter-cultural assessment to enhance the understanding of culture influences in speech emotion recognition accuracy.

1.5 Phases of the Research Framework.

This research work is conducted based on six phases framework that serves as a basis to ensure the research objectives are addressed. It is integrated from four different correlations; namely: psychological, neural, signal processing and cognitive research area. The framework presents the abstract, logical structure of meaning that enables the findings to be linked to the body of knowledge. Figure 1.2 illustrates the block diagram of this integrated framework.

Figure 1.2: The Integrated Six Phases Framework of the Research.

The followings are the detailed descriptions of each phase:

Phase 1: Problem analysis

This phase is derived from the fundamental understanding of the psychologists' correlation. The different psychological perspectives towards emotion provide alternatives in defining emotion that shapes the direction and scope of this research. Three prominent theories of emotion are reviewed; namely: relational model [Ekma07, Sche01b, Laza94], functionalist model [Camp94, Frij88, Schl54] and social constructivist model [Gord89, Lewi83, Aver80]. Relational model of emotion focuses on the appraisal process of emotion. Based on this model, emotion can be elicited assuming appropriate stimuli are presented and can be measured using physiological changes of the voice. Hence, this understanding justifies the usage of speech as the input. Functionalist model refers to the dynamicity of the emotion. Psychologists adopting the functionalist model believed that emotions could be represented using affective space model. This affective space model is generated from the emotion primitives' values of valence, activation and dominance that are descriptively relevant to describe the emotional characteristics. Based on this notion, it is possible to visualize speech emotion using this space model, thus providing a better understanding and analysis of speech emotion. Subsequently, social constructivist model promotes culture as an integral parameter in emotion conveyance and perception. This notion is used in the thesis to further examine the culture influence dependency on speech emotion recognition system. In addition, two different archetypal of emotion are considered. Emotion can be viewed as discrete class [Cowi03, Ekma71c, Ekma71b, Ekma71a, Ekma69] and as multi-dimensional affective space model [Woll09, Plut02, Sche01a, Buck99, John92]. Both views are used in designing the speech emotion verification and identification experiment in the later phase. To strengthen the

fundamental understanding, taxonomy of other researchers' finding and on-going cultural-influenced speech emotion research are also reviewed.

Phase 2: System design

Design issues of speech emotion recognition system are addressed in this phase. Three different data collection methods of natural speech, elicited speech and acted speech are discussed [Call08, Verv06]. Based on this understanding, three cultural-influenced speech data corpora using acted speech were employed; namely Berlin Emo-db [Burk05], NTU_American and NTU_Asian dataset. The Berlin Emo-db was collected by the Institute of Communication Science of the Technical University of Berlin and is considered a standard speech emotion dataset. It represents the European culture influence as the speakers used the German language. The other two datasets are constructed using video clips from the Internet to represent the American and Asian cultures respectively in inter-cultural and intra-cultural assessments. Details of the speech emotion corpora will be discussed in Chapter 3. In addition, Real time Speech Driving (RtSD) dataset was collected to measure the correlation between emotion and driver behavior state. This is because emotion directly influences judgment by serving as experiential and bodily information regarding how one feels about the object that affect judgment [Bell09, Clor07]. Appendix AIII discussed the experimental results showing possibilities of using the RtSD dataset to detect different driver behavior states.

Features extraction methods of Mel Frequency Cepstral Coefficient (MFCC) [Skow04, Slan98, Youn95, Davi80] and Gaussian Mixture Model (GMM) [Reyn00, Rip196, Reyn95b, Reyn95a] are selected to transform the speech emotion signal to appropriate form for further analysis. MFCC is employed due to its capability to emulate the functionality of the human cochlea in extracting important features for neural

decoding in simulated auditory cortex. On the other hand, GMM is used to provide a smooth approximation to the underlying sample distribution of the speech based on the statistical nature of speech signals. Once the speech data are in appropriate form, classifiers are used to segregate the different emotions. In this thesis, a neural network classifier of Multi Layer Perceptron (MLP) [Bish95, Rume86, Rose62] and fuzzy neural network classifiers of Adaptive Network Fuzzy Inference System (ANFIS) [Jang93] and Generic Self-organizing Fuzzy Neural Network (GenSoFNN) [Tung02a] are adopted. This is to functionally mimic the way human brain processes emotions to infer feedbacks and responses. The speech emotion data collection and features extraction methods are basically the contribution from digital signal processing field and the different classifiers mimicking the brain functionalities adopts the neural functionality understanding.

Phase 3: Speech emotion verification

Phase 3 handles the preliminary work of emotion verification and spectrogram-based speech emotion profiling. Emotion verification differentiates the targeted emotion and its complementary class (2-class problem). In order to perform such analysis, feature selection method is implemented to reduce the number of features used. Automatic Relevance Determination (ARD) [MacK94, MacK92] is selected as a principal selection mechanism due to its simplicity, scalability and ability to empirically yield good results [Guyo03]. Suitable network model for the classifier is then selected to ensure that the best accuracy can be obtained.

Phase 4: Cultural dependency analysis

Phase 4 concentrates on the different cultural dependency analysis using three different cultural influences from Europe, America and Asia. The speech emotion data were mixed in the homogeneous and heterogeneous way respectively and segregated into multiple

experiments. Different experimental settings are designed based on the hypothesis that intra-cultural assessment performs better than inter-cultural assessment [Grim07, Bezo83]. Comprehensive analysis and discussion is provided in Chapter 5.

Phase 5: Affective space model approach

Phase 5 introduces an alternative tool for speech emotion analysis using the Valence-Arousal (VA) analysis approach. The idea is derived from the psychologists' understanding of affective space model [Schl54] that emotion can be represented using emotional primitives of valence, activation and dominance. The visualization of speech emotion using the VA approach complements the discrete-class approach described in Chapter 5. Although the discrete-class approach is the current state of art method adopted by many researchers, it is not able to provide further analysis of emotion characteristic. Hence, the VA approach is used as an alternative tool to comprehensively analyze speech emotion.

Phase 6: Correlation of emotion and behavior

Phase 6 focuses on finding the correlation between emotion and behavior particularly for driver behavior state (DBS). The VA approach is adopted due to its ability to visualize the emotion dynamicity and characteristics. It goes beyond the current discrete representation of emotion and is a powerful alternative tool for description of emotion system. It allows the practical implementation of the speech emotion recognition system. Phase 5 will be presented in Chapter 6 and Phase 6 is described in Appendix AIII.

1.6 Research Hypotheses.

The research work is comprised based on the following hypotheses:

- Hypothesis 1: Emotion can be propagated and quantified through speech.
- Hypothesis 2: Cognitive inspired approach can provide better performance than statistical approach for speech emotion recognition system.
- Hypothesis 3: Culture influences emotion conveyance and perception.
- Hypothesis 4: Speech emotion recognition system performs better for intra-cultural speech emotion compared to inter-culture speech emotion recognition.
- Hypothesis 5: Emotion is dynamic and can be represented by emotion primitives of valence and activation.
- Hypothesis 6: Neutral is the base of all speech emotion.
- Hypothesis 7: Emotion can be elicited from the listener if appropriate stimuli are presented.
- Hypothesis 8: Emotion has reciprocal relationship with human behavior.

1.7 Research Contributions.

There are five main contributions desired from this work. They are listed as follows:

1. *Qualifying the qualitative psychologists' understanding of speech emotion.*

To date, psychologists and engineers seem to take different routes in tackling the speech emotion recognition problem. This situation resulted in differing views and approaches. In this work, the attempt is made to bridge the two fields by empirically verifying the psychologists' observation using engineering method of

quantitative measurement hence complementing the psychologists' qualitative study. Psychologists extensively studied emotion as one of the three important mechanisms that contributes to the prioritization and selection of information [Raym09, Lang95] as well as attention and motivation. Emotion as an affective evaluation of stimuli interacts with attention (the immediate task relevance of stimuli) and motivation (the predicted value of stimuli) to cater for the brain's limited capacity to simultaneously handle vast amount of information. However, such psychologists' understanding is a qualitative measure. The psychologists may know the fact that emotion plays an important role but neglects the possibilities to test their hypotheses in numerical descriptions. Hence, this scenario explains the minimal empirical research reported to quantify emotion. Qualitative research can be broadly defined as a kind of research that produces findings based on real-world observation where the phenomenon of interest unfolds naturally [Patt02]. It is also referred to as what, how, when and where of things (its essence and ambience) compared to the quantitative research reference to meanings, concepts, definitions, characteristics, metaphors, symbols and descriptions of things [Berg08]. In this research, computational intelligence methods are used to empower the understanding of affective state in such a way that it can provide a datum for better comparative analysis. It also helps to simplify identification of emotions for machine implementation. The fundamental hypotheses that emotion is propagated through speech and emotion can be measured are the basis of this research work. Multi-disciplinary contributions are used for feature extraction and classification to mimic the brain's functionality for emotion recognition.

Current research in speech emotion recognition demonstrates that many classification techniques as well as large number of features can be used to

automatically segregate different emotions. However, most of the studies are based on statistical approach. The excellent classification performance obtained does not really reflect on the way human brain behaves with the speech emotion. In this work, Mel Frequency Cepstral Coefficient (MFCC) [Slan98] is used as a features extraction method in an attempt to simulate the way the cochlea filter works and selects the important features to be presented to the brain for neural decoding. The neural network of Multi Layer Perceptron (MLP) [Bish95, Rume86, Rose62] and fuzzy neural network of Adaptive Network-based Fuzzy Inference System (ANFIS) [Jang93] and Generic Self-organizing Fuzzy Neural Network (GenSoFNN) [Tung02a] are used as a classifier. This is to functionally mimic the way human brain process information from the selected features. The proposed speech emotion recognition system is able to perform with reasonable degree of accuracy and demonstrates the potential for future implementation.

2. *Empirically demonstrates the importance of culture as one of the key parameters of emotion conveyance and perception.*

Culture is learned since young and it enriches the conversation to be more engaging and lively. Cultural influence can be disseminated through speech thus making speech as one of the most reliable source to understand the effect of culture on the spoken language. However, most researchers tend to by-pass the importance of culture as it is assumed to be embedded in the background (in this case as part of the Long Term Memory (LTM)). In this work, an extensive study investigating the influence of culture on the speech emotion recognition system provides a better understanding of this neglected effect. Three different culturally-influenced datasets representing American, European and Asian cultures are used in the experiments. The study provides different views and perspectives of intra-cultural

and inter-cultural dependency influence in homogeneous and heterogeneous culturally-influenced speech emotion data setting. In addition, the effect of culture can be clearly visualized using the Valence-Arousal (VA) analysis approach especially in driver behavior state (DBS) analysis using the Real-time Speech Driving dataset (RtSD). Hence, such approach is tailored to be more culture-specific to achieve better performance in term of accuracy.

3. *Provides a better generalization of the affective space model by adopting neutral as a common ground for different emotion comparison.*

While understanding culturally-influenced speech emotion data, neutral is found to be the common denominator for all intra-cultural assessments. This is because neutral is assumed to be an emotionless state that is supposed to have zero value for valence and activation axis. Neutral is regarded as equal irrespective of the cultures. It has always been regarded as inconsequential and trivial since it does not have any significant emotional information. Due to this fact, most researchers do not include neutral as one of the basic emotion and considered it as a constant void. However, neutral is located in the center of the model based on the understanding of Affective Space Model. Due to the culture influence, the location for the neutral centroid of the cluster is hypothesized to be the origin of the generated Valence-Arousal (VA) Model. The dynamic arrangement of the axis is proven intuitive to handle the emotion dynamic. From the experimental results, it can be observed that neutral of one culture is unique compared to neutral of the other cultures although its' valence and activation values are supposed to be zero. Based on this observation, it is hypothesized that neutral is the base of all emotion and should be considered as one of the important parameter in understanding culturally-influenced speech emotion data. Neutral in speech emotion is also unique, as it does not exist for other bio-

signal analysis such as EEG signal. This attribute makes speech emotion as one of the relevant tools for understanding the effect of culture in speech emotion.

4. *Introduces an alternative tool of Valence-Arousal (VA) analysis approach to measure and analyze speech emotion derived from the psychologists' affective space model understanding.*

Evaluation of speech emotion profile visualization based on the emotion primitives' value is one of the major contributions of this research work. This approach provides an alternative measurement and analysis to speech emotion. This is very important for dealing with dynamic emotion, which changes over time. Such an approach enables researchers to view emotion in different perspectives and understand the cultural interaction of emotion from the viewpoint of the psychologists'. The culture influence factor can be addressed from the neutral region in the understanding of the VA model. Moreover, the emotion region projection (in this study the calm region of Quadrant 4) can also be derived with the assumption that each emotion; namely: anger in Quadrant 1, happiness in Quadrant 2 and sadness in Quadrant 3, are known prior to the VA model generation. Such analysis is not possible using the current state-of-art discrete-class emotion recognition approach. Hence, combining both approaches may improve the accuracy for speech emotion recognition system as well as providing an in-depth analysis of speech emotion from a different perspective.

5. *Extending the Valence-Arousal approach for in-depth study of human behavior.*

Subsequently, the extension of speech analysis using the VA approach can be extended to understand driver behavior state. This is because driver behavior state is directly influenced by emotion and can be disseminated through speech. Drivers'

emotion states such as: sleepy, talking through mobile phone, laughing as well as normal driving; are used to investigate the correlation between speech emotion and behavior. The in-depth study enables a better understanding of the relationship of these two parameters.

1.8 Organization of the Thesis.

This thesis is organized into seven chapters. Chapter 1 covers the overview of the research work, the present situation as well as the motivation of the research and the research objectives. Chapter 2 presents the literature review of the relevant concepts and works related to the research. Chapter 3 gives the detailed description of the architecture of the proposed speech emotion recognition system that addresses the main components in the system; namely: data collection with listening comprehension survey and corpus generation, features extraction and classification. Preliminary experimental result on the verification of the speech emotion is presented in Chapter 4. Chapter 5 demonstrates the different approaches of speech emotion identification in discrete-class approach using an extensive study of intra-cultural and inter-cultural assessments. Thorough discussion and observation are presented to provide a complete understanding on the effect of culture on the recognition process. Chapter 6 is dedicated to discuss about the Valence-Arousal (VA) analysis approach as a complement to the discrete-class approach in understanding speech emotion. Implementing the proposed VA analysis approach to real world application based on driver behavior state understanding provides a practical application and will be presented in the Appendix AIII. The final Chapter 7 subsequently summarizes the work as well as the future recommendation focusing on the tasks that can improve research in speech emotion related field in general.

CHAPTER 2

LITERATURE REVIEW

Chapter 2 presents the relevant literatures intended to give the reader with a basic understanding of speech emotion recognition. It forms a foundation of this convergent study by defining a number of concepts that are intricately correlated in addressing the speech emotion recognition issues.

It is well known that emotion is the affective aspect of consciousness [Pank05] and this will be discussed in the first part of this chapter. Detailed descriptions of the psychologists' perspective of emotion will be used to define the context of this research, where thought and physiology are inextricably entwined. Many psychologists are motivated to derive a comprehensive emotion theory to aid better understanding. In fact, emotion is such a difficult concept that has multitude of relevant aspects to be adequately defined. Hence, there are many definitions of emotions in the scientific literatures [Prin04, Klei81]. To date, no unified or generally accepted theory of emotion is available. In this chapter, three prominent psychological perspectives of emotion are listed below:

- Relational Model,
- Functionalist Model, and
- Social Constructivist Model

and will be examined to provide a better comprehension and its use in this thesis.

This is followed by a section on the two types of emotion theory defining the number of existing emotion that can be distinguished. The discrete emotion model and multi-dimensional emotion model are presented. Both concepts are used in designing the experiments for further investigation. As human experienced such a wide range of

emotions, a subset of emotions need to be considered for simplification purpose in order to conduct the emotion classification.

The third part of this chapter reviews the summary of similar state-of-art works in speech emotion recognition. It includes the description and taxonomical table of the speech emotion recognition systems that outlines other researchers' findings in the speech emotion recognition field. Although the analysis may not be exhaustive, many explanations of speech emotion recognition works are intended to familiarize readers with the basic assumption in speech emotion. The performance of state of the art speech emotion recognition systems are also provided as an overview of the milestone achieved. A taxonomical table based on the data corpus, features extraction methods, number of emotions identified and the classifiers used in such systems are compared to give a clear yard stick of the systems performance. Subsequently, the on-going cross cultural speech emotion research is discussed. Such description is aimed to give reader some foundation understanding that leads to the motivation of this research work. In addition, a summary is provided to proffer the gist of the overall discussion in this chapter.

2.1 Psychologists' Theoretical Perspectives on Emotion

In this ever-changing everyday life situation, emotion represents a pervasive reaction to major significant event of an individual. It is a complex psychophysical process that arises spontaneously and involuntarily without explicit conscious effort. This reaction evokes either positive or negative psychological response or physical expression in order to cope with the situation, particularly in managing stress. Emotion can be a prime determinant of the sense of subjective well-being. It appears to play a central role in displaying and regulating patterns of behavior that is realized in many human activities.

Thus, many theoretical psychologists concentrated on the question of the scientific elucidation of emotion and resulting in many different theories of emotion to explain what emotions are and how they work. The models of emotion will be described in the subsequent sections.

2.1.1 Relational Model of Emotion

Relational model of emotion attempts to provide a description of the emotion generation process. It consists of a complex reaction that engages both human mind and body [Laza94]. The reactions can be listed as:

- i. *Subjective mental state* such as the feeling of anger or anxiety,
- ii. *The impulse to act* which can be expressed overtly or otherwise; such as: flighting or fighting, and
- iii. *Profound changes in the body*; such as: increased heart rate or blood pressure.

These reactions arise when a stimulus, event or situation is cognitively assessed (appraised) which is determined by personal goals and concerns that have been shaped through the individual's experiences. Thus, an appropriate emotional response is then generated depending on the outcome of these reactions.

Scherer et al. [Sche01b] summarized a working definition of emotion appraisal theory containing complex multifaceted experience with the following components:

- Cognitive appraisal Events are judged or appraised, which have significance to an individual's goals, concerns, values, needs, preferences, or well-being would elicit an emotion.
- Subjective feelings The appraisal process is accompanied by feelings that are

good or bad, pleasant or unpleasant and calm or aroused.

- **Physiological arousal** Emotions are accompanied by autonomic nervous system activity.
- **Expressive behaviors** Emotion is communicated through facial and bodily expressions, postural and voice changes.
- **Action tendencies** Emotions carry behavioral intentions, and the readiness to act in certain ways.

Subsequently, Ekman in his book refined the emotion definition by stating that emotion is a particular kind of automatic appraisal process [Ekma07]. It is influenced by human evolutionary and personal past such as feelings, perceptions or beliefs about elements, objects or relation between them. When one senses that something important to his/her welfare is occurring, a set of physiological changes and emotional behaviors begins to deal with the situation.

2.1.2 Functionalist Model of Emotion

Functional theorist defines emotion as a dynamic interface between an individual and the environment that emphasizes the context under which the emotion is experienced. The idea is that emotions are integral to a person-environment transaction that they function to serve towards an individual's goals. For instance, when an individual feels happy, that particular individual has a sense that his/her life is going well. On the contrary, when he/she is angry, the particular individual knows that something is not right and has the sense that there is something he/she needs to do about it. Functionalist model of emotion focuses on the goal or the end result of emotion rather than the process of emotion elicitation that focuses on the relational theory of emotion described earlier. In layman's terms, emotion provides checks and balances for regulators of human behaviors.

Campos [Camp94] noted that emotion is characterized by: 1) postulation of a close interrelation between emotion and the goals and strivings of the person; 2) its emphasis on emotional expressions as social signals; and 3) the hypothesis that the physiology of emotion, far from involving only homeostasis and the internal milieu, can regulate and be regulated by social processes. Campos also mentioned the following four factors that are significance to an individual experience and thus evoke an emotional response:

- i. The relationship of the event to the goals including what an individual wants in his/her transactions with others.
- ii. The social response others give to that person (e.g. their emotionally expressive reaction which the individual may view as evaluative of himself/herself).
- iii. The hedonic tone of the event (e.g. consider the hedonic differences in an excruciating social exposure versus a tenderly affectionate exchange.
- iv. The individual's memories of similar earlier events.

Frijda's theory views emotions as functional expressions of the individual in response to his environment [Frij88]. Emotions are convergence of possibilities offered by the surrounding or situation with possibilities that the individual has with respect to satisfying his/her concerns. Both '*threat*' as well as the hope of '*satisfying*' concerns set the emotion process in motion. Emotions arise in the '*interaction of the situational meanings and concerns*'. Frijda also proposed that each emotion is paired with action readiness with a quantitative aspect of '*activation*' and qualitative aspect of '*action tendency*'. This conceptualization focused on emotion as a continuous process rather than discrete event. The usage of emotion dimensions [Cowi03, Lang95, Schl54] such as a positive-negative continuum or degree of arousal [Russ85] would be descriptively

relevant to a functionalist position insofar as the hedonic tone of the emotional experience is a notable marker and influenced subsequent behavior. Further description of this multi-dimensional emotion theory will be described further in Section 2.2.

2.1.3 Social Constructivist Model of Emotion

The third model to explain the psychologists' theoretical perspective of emotion begins with the idea that emotions are socially constructed in the sense that social conditioning in a culture accustoms human feeling. The individual participation in social structures are acquired or learned through experience and collective custom of a particular community. Emotion and their expression are regulated by social norms, values and expectations in conjunction with cultural ideologies and beliefs. These values influence the appropriateness of the emotion perceived and the way the emotion is conveyed to the audience.

Averill [Aver80] defined emotion as a transitory social role that includes an individual's appraisal of the situation that is governed by social norms and expectations. A role may be defined as a socially prescribed set of responses to be followed by a person in a given situation and emotions are interpreted by passive responses to that particular situation. For example, when a worker loses his/her job, he/she may yell, cry, and hit the boss or perhaps displaying no overt behavior at all. Social norm dictates how the worker responds to the situation of either suitable or inappropriate to the society acceptance. For instance, it is acceptable for the worker to plead with the boss to get the job back but it is not tolerable for the worker to hit the boss as an outlet for his/her frustration.

Lewis and Michalson [Lewi83] in their work with infants found that the child is able to feel the emotion and convey the emotion after the child is exposed to socialization

influences that coordinate between emotional states and behaviors. Firstly, the children have to learn the way to coordinate their states, experiences and behaviors. Then, they practice to exercise the emotion independently of one another and at will. The learning is highly influenced by culture and exposure that they have at that particular period of time. Gordon [Gord89] also stated that when the children understood the cultural meaning of an emotion, he/she is able to act towards it in term of magnifying, suppressing, or simulating the emotions in themselves as well as evoking or avoiding it in other people. Through the exposure to others, the children learn the emotional behaviors, norms and symbols of their culture as unintended consequences of social interaction. Gordon later viewed emotion as a composition of (1) bodily sensation, (2) expressive gestures, (3) social situations or relationships and (4) emotional culture of a society [Gord90]. Gordon also stressed that the bodily sensations are only significant when it was expressed in behavior guided by cultural norms. Thus, the power of culture is particularly evident in the formation of emotion based on the social constructivist perspective.

2.1.4 Psychological Perspective of the Thesis

In summary, based on the different perspectives, there is no single theory that can boast to comprehensively define emotion. Most theories in this regards are all interconnected and serve their purpose in refining the respective psychologists' interest. Combination of these three psychologists' theoretical perspective of emotion provides a basis of this thesis with three main emphasis; namely: cognitive functions in emotion processing, the dynamic charecteristics of emotion and the effect of culture on speech emotion. The correlation of each psychologist's perspective in this thesis is summarized in Figure 2.1.

Figure 2.1: The Psychological Perspectives of Emotion

Cognitive function in emotion processing. Using Scherer approach of Brunswikian Lens for speech emotion [Sche78], Lazarus's theory is referred [Laza94]. Lazarus theory states that emotion is elicited based on the appraisal of changes in the surrounding that have personal meaning to the agent (in this case: the speaker). Thus, based on this notion, it is hypothesized that emotion will be elicited if the appropriate stimulus are presented to the speaker. The emotion produced can be detected through physiological changes of voice. The changes then can be empirically measured so that the conveyed emotion can be discriminated.

The dynamic characteristic of emotion. Frijda proposed that emotion is a continuous process based on the functionalist model of emotion [Frij88]. The dynamic characteristic of emotion allows it to be mapped onto an affective space model. The usage of affective space model notion was first proposed by Schlosberg in 1954 [Schl54]. In this thesis, patterns of the speech emotion are measured and profiled in such a way that it can complement the current state of the art approaches for better speech emotion understanding. The emotion primitives with emphasis on valence and arousal are studied

in detail to analyse the similarity and dissimilarity measurement in comparison from one emotion to another. In addition, comprehensive study on the effect of culture towards speech emotion system performance is also conducted using this approach as it is able to offer such a powerful visualization tool.

The effect of culture on speech emotion. Culture plays an important role in affecting the conveyence of speech emotion and perception. This notion has been explicitly described by psychologists who believed in the social constructivist model of emotion. Such deep rooted culturally-influenced behaviors, in which an individual has been exposed since young, yielded an automatic psychological response. Emotions are believed to be universal that enable human from different background and geographical boundary to correctly discriminate it. However, there are certain characteristics that are shared among people of similar cultural background that makes it unique to that particular community. Thus, based on the theory inspired from the social constructivist emotion model, the universality and diversity of emotions for different cultures with emphasis on language as a cultural parameter will be investigated in depth. Universality refers to the ability of similar emotion to be recognized across cultures whereas diversity focuses on the ability of different emotions to be identified in a similar culture. An example of universality is that emotion *anger* (same emotion) can be detected in European, American or Asian cultures respectively (different cultures). Subsequently, example of diversity is that emotion *anger*, *happiness*, *sadness* and *neutral* (different emotions) can be segregated in Asian culture (same culture). In this thesis, the inter-cultural assessment is refined towards acknowledging the universality of the emotion and the intra-culture assessment is more directed in finding the diversity of emotion.

In summary, emotion in this thesis is defined as a mental state that arises spontaneously from external stimuli that is often accompanied by dynamic physical

expressions (adopted from relational model of emotion perspective). This is heavily influenced by culture (adopted from social constructivist model of emotion perspective) to serve a purpose of a particular situation in a dynamic manner (adopted from functionalist model of emotion perspective).

2.2 Archetype of Emotion

Researchers from both the cognitive science and engineering have long argued on what constitutes as subset of emotions from the plethora of emotions exhibited. Emotions need to be categorized so that classification is possible and more useful observation and analysis can be conducted. There are two methods to sub-divide emotions; namely: discrete basic emotion and multi-dimensional emotions model.

2.2.1 Discrete Basic Emotions

One traditional approach to simplify the categories of emotion is to postulate a small number of basic or primary emotions from the entire range of emotions that are tinged and enlivened. Due to its subjectivity and dependancy on culture and environment-related factors, there have been many criteria to identify emotion resulting in a several number of basic emotion lists. The disparity between these emotion lists occur because each researcher claims that his/her list of emotion subset covers the most comprehensive description of basic emotion. Nonetheless, researchers in various disciplines agree that there are some basic emotion that can be universally accepted and is fairly consistent across cultures [Buck99, Ekma88]. Hence, it is sound to consider some emotions that are universally accepted for analysis purposes.

Cornelius [Cowi03, Corn96] coined the “*Big Six*” term of the best known list of emotion in speech which consists of emotions; namely: fear, anger, sadness, happiness, surprise and disgust. This notion of six basic emotions was first proposed by Ekman in his research work on facial expression across culture [Ekma71c, Ekma71b, Ekma71a, Ekma69]. Ekman reported that members of different cultures were able to reliably identify expressions of emotion in photographs. Based on this finding, Ekman conceded that there are certain emotions that are universally accepted across cultures. However, this Big Six emotion list is challenged by others [Plut02, Buck99, John92, Ekma88] with the arguments that emotions cannot be thoroughly explained with a limited set of class and it does not allow subjects to access an adequate level of discrimination. Hence, based on such contradiction, many version of basic emotion lists are introduced. Ekman and Heider [Ekma88] divided anger by two subclasses of hot and cold anger. They also compiled contempt as one of the anger subclass as a distinction to a milder anger. Desire was added by Johnson-Laird and Oakley in their list [John92]. Buck [Buck99] distinguished emotion into four affective groups; namely:

- biological affects (happiness, sadness, anxiety, fear, anger, disgust),
- social affects (pride, guilt, envy, pity, arrogance, shame, jealousy, scorn),
- cognitive affects (curious, interested, surprised, bored, burned out), and
- moral affects (moral rapture, moral indignant).

In addition, Plutchik [Plut02] defined eight basic emotions; namely: anger, anticipation, joy, trust, fear, surprise, sadness and disgust, by modelling three-dimensional circumplex color wheel. This is to describe the relation among emotion concepts as depicted in Figure 2.2. The cone’s vertical dimension represents the intensity of emotion and the horizontal circle represents the degree of similarity among the emotions. The eight sectors are designed to indicate that there are eight basic emotions defined in his theory and arranged

in circular manner as four pairs of opposites with their respective angles. In the exploded model, the emotions in the blank spaces are the primary dyads-emotion that are the combination of the two primary emotions. However, Plutchik did not reflect neutral in his list as he defined neutral as an absence of the emotion rather than one of the emotion to be considered. Table 2.1 shows the Plutchik's examples of the stimulus event that is recognized by human and animal which elicited possible feeling state and overt reaction to achieve the needed effect. This observation was derived from the perceived similarity between human and animal behavior when encountering such situations.

Figure 2.2: Plutchik's Emotions Color Wheel [Plut02]

Table 2.1: Examples of Stimulus Events and Possible Reaction and Effect [Plut02]

Stimulus event	Cognition	Feeling state	Overt behavior	Effect
Threat	“Danger”	Fear	Escape	Safety
Obstacles	“Enemy”	Anger	Attack	Destroy obstacles
Gain of valued object	“Possess”	Joy	Retain or repeat	Gain resources
Loss of valued object	“Abandonment”	Sadness	Cry	Reattach to lost object
Member of one’s group	“Friend”	Acceptance	Groom	Mutual support
Unpalatable object	“Poison”	Disgust	Vomit	Eject poison
New territory	“Examine”	Expectation	Map	Knowledge of territory
Unexpected event	“What is it?”	Surprise	Stop	Gain time to orient

2.2.2 Multi-dimensional Emotion Model

In contrast to the idea that emotion is a discrete event, some researchers later pointed to the fact that emotion is a continuous process that will change dynamically over time. This notion has been supported in many current direction of speech emotion research using the multi-dimensional emotion model [Woll09, Goud09, Sche01a]. Based on this model, emotion can be discriminated by specifying a value for each dimension; namely: valence, activation and dominance. Figure 2.3 depicts the affective space model with x, y and z axis representing valence, arousal and dominance respectively. The arousal term is interchangeably used with activation that carries similar meaning throughout this thesis.

The affective space model with 3 dimensions was first proposed by Schlosberg [Schl54]. It consists of three main emotion primitives; namely: valence, arousal and dominance. The emotion primitives are actually the descriptive generic attributes of an

emotion. These attributes function as constituents that act as a fully complementary description of the emotion. The affective space model can be divided into quadrants where each quadrant represents positive or negative value of the emotion primitives. The emotion primitives ranging from positive to negative value; such as: scale of pleasure (+) to displeasure (-) for valance, scale of active (+) to passive (-) for arousal as well as scale of strong (+) or weak (-) for dominance. Figure 2.3 shows the affective space model with several labeled emotions with neutral as a black dot in the middle of the model.

Figure 2.3: Four Different Basic Emotion Quadrants of the Affective Space Model

Cowie et al. [Cowi00] developed a software tool called '*FeelTrace*' to record the perceived emotional content of a time-dependant stimulus from observers using a two dimensional affective space model constructed using the arousal and valence values. Cowie denoted arousal as activation ranging from very passive to very active and valence denoted as evaluation ranging from very negative to very positive in this work. The observers move a pointer in a form of a colored disk in the circle representing the axes as they reported the emotion perceived based on the stimulus. The colored disk approach

was inspired by Plutchik's color wheel approach [Plut02] that can be associated with the relevant emotional state. Figure 2.4 shows the instrument and a mouse movement starting at *afraid*, moving to *interested*, *relaxed*, and finally to the emotional state *content*.

Figure 2.4: Dimensional Labeling in FeelTrace software [Cowi00]

The basis of the speech emotion research in this thesis is to adopt the view that some emotions are more universally accepted than others. This assumption can give a clear direction in developing the proposed speech emotion recognition system. Thus, in this study the focus are directed on three basic emotions; namely: anger, sadness and happiness with neutral acting as a reference for the emotionless state. These emotions are selected based on the psychologists' view [Cowi03] and can be assigned in three different quadrants of the affective space model [Lang95, Schl54] as shown in Figure 2.3. Furthermore, based on Ververidis and Kotropoulos review of 32 emotional speech databases [Verv03], emotions of anger, sadness and happiness are the top most recorded and analyzed emotions by the researchers in the field of speech emotion.

2.3 Related Works in Speech Emotion Research

To date, most research in the areas of speech emotion recognition can be divided into two main focuses. Firstly, the focus is on the impactful speech acoustical features for classification and secondly, is on the suitable classifiers to be employed. In this section some of the works that had been conducted will be briefly discussed to give some idea of the research area that need to be explored and enhanced for the purpose of this thesis. However, the data used for training and testing also plays an important influence in determining the successfulness of the speech emotion recognition system. Ververidis and Kotropoulos [Verv06] introduced three main approaches to data collection for emotional speech research; namely based on: natural speech, elicited speech and acted speech. Their work was later supported by Callejas and Lopez-Cozar [Call08]. These approaches require some fine-tuning and balancing between naturalness of the speech expressed and degree of control in order to obtain optimum classification accuracy. Figure 2.5 shows the correlation of these approaches to the degree of naturalness and control.

Figure 2.5: Different Approaches of Generating the Speech Emotional Corpora [Verv06]

Natural speech data, shown in shade of green in Figure 2.5, can be acquired from emotionally charged real-life situation. Such speech data can be collected through several

sources; such as: call centers, recording audience response in live performances of television talk shows or capturing students' responses to the professor in the lecture theatre. This method offer the highest degree of naturalness but at the same time creates challenge to the researchers to individually annotate the correctly perceived emotion. This is because the emotion conveyed is always unpredictable with no prior information and it is expressed freely by the speaker. In addition, speech in daily life tends to be expressed in moderate emotional states rather than extreme or full-blown basic emotion [Cowi03]. This senario occurs due to personal reservation and modesty due to the culture or collective norm in order to be socially accepted in the community.

The elicited speech approach, shown in shade of red in Figure 2.5, can be acquired by having the subjects to perform certain tasks to induce the intended emotion. Some of the popular methods to encourage the speaker to respond appropriately with the intended speech emotion includes providing scenario-text to the speaker to perform emotional speech under a certain prescribed situation, assigning pre-selected music for the subject to listen to as well as presenting the subjects with a pre-determined video clips to watch. Emotions can also be induced by allowing the subject to play a difficult computer games or by interrogating the suspect to admit his/her guilt. After the subject has been roused to feel the intended emotion, his/her speech responses can then be recorded. Although this method provides better control of the selected emotion compared to natural speech data collection approach, it may compromise the naturalness of the emotional speech response. The subjects may fake the reaction resulting in a misleading emotional speech feedback. Thus it is not suprising if the naturalness of the emotional speech is ambiguous.

Subsequently, the third approach of emotional speech collection is the acted speech which is shown in shade of blue in Figure 2.5. This kind of speech data can be recorded from an actor imitating the intended emotion [Call08]. This method provides the

best balance between control and naturalness in comparison to the other two data collection approaches. The reason is because it is commonly accepted that actors can correctly deliver the appropriate speech emotion although there is a high tendency for exaggeration [Verv06]. The quality of the emotional speech is highly dependant on the actor's skill and an almost natural-like emotional speech data can be obtained from the professional actors who can portray a more realistic emotion. It is also noticed that acted speech enables intense emotional speech portrayal in high-quality audio recording hence simplifying the pre-processing step to remove artifacts and noise. Table 2.2 summarized the different data collection approaches adopted by selected researchers in speech emotion recognition system. The term *simulated* refers to acted speech, *elicited* refers to emotion-rousted speech and *natural* refers to freely-expressed speech emotion data.

After the appropriate speech emotion data are collected, speech emotion features can then be extracted using either a simple spectral analysis or a complex Hidden Markov Model [Verv06, Zhou01, Woma99]. Features are extracted from the pitch, formants, short-term energy and many other acoustic characteristics that have been utilized for emotion or stress recognition. Most of these works provided either a simple discrimination of stress versus neutral behavior of the subjects or limited emotions understanding. Some popular acoustic features used for emotion recognition purposes [Grim07, Verv06] are listed as follows:

- i. ***Fundamental Frequency F_0 (Pitch)*** – also known as glottal waveform because it depends on the tension of the vocal folds and the sub-glottal air pressure. It describes the intonation and speech melody. For instance, it can capture the monotone speech or highly accented syllable.
- ii. ***Formant*** – representation of the vocal track resonances. It can be used to discriminate the improved articulated speech from the slackened speech.

- iii. ***Speaking rate*** – describes the temporal characteristic in the prosody. Its properties include pauses between voiced sounds, lengths of voiced segments and rate of speech (articulation). For instance, it might reveal whether the speech sound is urged or relaxed.
- iv. ***Energy/Intensity*** – used to capture the volume in speech and helps to discriminate loudness and is related to the arousal level of emotions. For instance, it is able to differentiate between sadness and depressed speech.
- v. ***Spectral energy distribution*** – particularly Mel Frequency Cepstral Co-efficient (MFCC). It exploits the human auditory frequency response in the cochlea. It uses a certain number of co-efficient filter banks and specific shape filter function. Its short-term statistics are very useful for phoneme recognition while the long-term statistic indicates voice quality. These features are very broadly used in automatic speech recognition (ASR) applications with high degree of success.
- vi. ***Statistic-measured features*** – Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM). The clustering algorithm is used to estimate the probability density function (PDF). Its mean vector, covariance matrix and mixture weight will be extracted as features. The timing information will be lost because it employs statistics of the prosody features (mean, variance, etc.) and neglect the sampling order. However, this technique may reveal cues about the speaker identity, gender and emotion.

Thus, from the features extracted, a suitable classifier is then employed to classify the emotion. Some machine learning algorithms such as K-nearest neighbours (kNN), neural networks (NN), Hidden Markov Models (HMM) and others have been successfully implemented. Several related works with emphasis on the features extraction methods and

classifiers type adopted as well as its performance are discussed to give an overview of the achievements in speech emotion recognition system.

Many researchers [Call08, Nwe03, Nich00] focus their attention in identifying speech emotion based on the Big Six discrete basic emotions notion proposed by Cornelius [Cowi03, Corn96]. The emotions to be identified are; namely: fear, anger, sadness, happiness, surprise and disgust. Variety of features extraction methods and classification approaches are adopted to obtain acceptable degree of accuracy. Although the researchers may recalibrate the emotion list by adding or removing any particular emotion that they are interested in their study, the fundamental understanding remains that emotion can be empirically measured and speech contains sufficient information on the level of emotion information so that it can be used as an input.

Some researchers reported their findings of the experimental results for the speech emotion recognition system. For instance, Dellaert et al. [Dell96] explored several statistical pattern recognition techniques to classify utterances according to their emotional content. A new method of extracting prosodic features from speech based on a smoothing spline approximation of the pitch contour coupled with Maximum Likelihood Bayes (MLB), Kernel Regression (KR) and k-NN classifiers can obtain classification performance that is comparable to human performance on the task. The human performance presented in the confusion matrix shows that anger yielded the least error (1%) and fear obtained the maximum error of 9%. By applying majority voting of subspace specialists combined with k-NN classifier, the best recorded accuracy in recognition of 4 emotions; namely: anger, sadness, happiness and fear, was 79.5%.

Petrushin [Petr99], in her research, conducted experiments to segregate 5 basic emotions by imitating call-center environment. Anger, sadness, happiness, fear and neutral can be automatically identified based on 30 non-professional actor's conversation.

Total average accuracy of 70% can be yielded by adopting several neural networks recognizers. As an extension, she verified two speech emotion classes of agitation and calm. Three basic emotions of anger, happiness and fear were combined as *agitation* class whereas sadness and neutral were grouped as *calm* class. Based on the ensemble neural networks approach, 77% average accuracy was obtained for the system performance.

Nicholson et al. [Nich00] found that the emotion recognition rate for eight emotions derived from radio actors' speech can reach up to 50% accuracy by applying several neural network approaches coupled with prosodic and phonetic features. The prosodic features selected were pitch and speech power while the phonetic features adopted were Linear Predictive Coefficient (LPC) and its LPC derivative. In another study, Nwe et al. [Nwe03] employed the short time Log Frequency Power Coefficient (LFPC) to be the speech features that represents energy distribution across the frequency spectrum. A four-stage ergodic Hidden Markov Model (HMM) is then employed as the classifier to ensure the effects of the speaking rate and variation of F_0 is minimized. The result of the experiments shows that average recognition accuracy between 77.1% and 89.0% can be achieved in classifying six emotions. Moreover, while studying the drivers' stress speech data, Fernandez and Picard [Fern03] reported their proposed speech recognition system could identify up to 61% recognition accuracy for speaker dependant speech data. The performance however drops to 51% for speaker independent speech data. The results are obtained using the mixture Hidden Markov Model (m-HMM) classifier combined with Teager Energy Operator (TEO) features. This is based on the 4 classes of driver states of FF, FS, SF and SS in the context of driving at variable speeds while engaged on mental tasks of variable cognitive load. The first letter denotes whether the data comes from a *Fast* (F) or *Slow* (S) driving condition; whereas the second

indicates the frequency which the driver was presented with an arithmetic task: namely; every 4 second is considered *Fast* (F) or every 9 second is considered *Slow* (S).

Subsequently, Lee and Narayanan [Lee05] conducted negative and non-negative emotions recognition in a commercially deployed automatic call-center. The effort devoted to negative emotion detection in order to improve the quality of the service in the automated call center application. Instead of letting the negative caller to be handled by computer operator, they were directly assisted by a human operator. This strategy can expedite the waiting time for the caller with negative emotion and served as an attempt to calm their impatient state. It is very crucial to improve the service quality as it was observed that unnecessary late respond may lead to user dissatisfaction, increased user frustration level leading to the user switching loyalty to a competitor company. Two pattern recognition techniques of Linear Discriminant Classifier (LDC) and k-Nearest Neighbor (k-NN) were used to classify the emotion states conveyed by the speech utterances. It was reported that the overall accuracy ranging between 66.1% and 80.0% could be obtained for verifying the negative and non-negative class compared to 50.0% chance probability.

In addition, Callejas and Lopez-Cozar [Call08] carried out some experiments to recognize the negative emotions; namely: anger, boredom and doubtfulness, in facilitating user interaction using Universidad Al Habla – University Speaking (UAH) dialogue system. This research was carried out to detect these negative emotions from the user's verbal interaction with the system to improve the system's quality and ensuring user satisfaction and loyalty. Selected features of fundamental frequency (F_0) or pitch, formant frequency and its bandwidth (F_1 , F_2 , B_1 and B_2), energy and rhythm were coupled with Multi Layer Perceptron (MLP) as the classifier. Such approach was able to yield 69.3% identification performance for the three emotion classes of anger, boredom and

doubtfulness with chance probability of 33.3%. The results can be boosted up to 80.0% when boredom and doubtfulness emotion were combined as one class as opposed to anger emotion with chance probability of 50.0%. This approach yielded an improvement of at least 30% from chance guessing performance.

Enriching the body of research, the performance of standard Gaussian Mixture Model (GMM) and GMM super vector based Support Vector Machine (SVM) were compared in Hu et al. work [Hu07]. This effort is dedicated to recognize 5 emotions of anger, fear, happiness, sadness and neutral. The GMM super vector based SVM were constructed using the Universal Background Model (UBM). The reason of such an approach is to represent the speaker independent distribution of speech features due to the UBM training with speech from a large number of speakers. This attribute can provide priori knowledge thus compensate the lack of training data for GMM. It was reported that accuracy ranging from 77.9% to 93.6% can be obtained using SVM classifier.

In more recent effort to boost speech emotion recognition understanding, Schuller et al. [Schu09] devised an international challenge in INTERSPEECH 2009 for detecting emotion given a standard corpus of the German Aibo Emotion Corpus [Ste09]. The corpus comprised of spontaneous speech from 51 children while interacting with Sony's pet robot, Aibo. The children were led to believe that the Aibo robot was responding naturally to their command to encourage natural responses although the robot was actually controlled by a human operator. The emotion challenge was divided into a two-class verification of *negative* emotion versus *idle* task and a five-class identification task of segregating *anger*, *emphatic*, *neutral*, *rest* and *positive*. The *anger* class consists of angry, touchy and reprimanding whereas *positive* class comprised of motherese and joyful. In this work, two standard classifiers were used; namely: SVM and HMM to ensure that the results were reproducible. The experimental results demonstrated that

38.2% accuracy for 5-class identification task and 67.7% for 2-class verification task were achieved. The challenge thus exhibited the difficulty in dealing with a real-life non-prototypical emotion recognition scenario.

From the selection of related works, it is clear that there are no absolute formula for the most appropriate features extraction method as well as the best classifiers type for speech emotion recognition system. The number of features extracted from the speech signal varies significantly from approximately 10 basic features such as mean value and range in pitch and intensity [Lee01] to 276 in the case of systematic application of functionals to a set of basic trajectories [Schu06]. The number of features investigated even increased to 384 for the 2009 INTERSPEECH emotion challenge [Schu09]. This variation of results indicates that researchers in the speech emotion arena are still investigating the best suited features and classifiers for speech emotion recognition task. It also relies heavily on the data used in training and testing.

Table 2.2 summarized the different approaches of some selected studies for a better comparison. It is ascendingly arranged according to the number of emotion analyzed in the selected works. The number of subject, the type of speech data collection, the feature extraction method as well as the classifier used to discriminate the emotions are listed in Table 2.2. The mean accuracy ranges of the emotions recognition task are also provided to give an overview of the performance of a particular approach for such speech emotion recognition system. Based on the review, Mel Frequency Cepstral Coefficient (MFCC) features extraction coupled with neural network and fuzzy neural network classifiers are selected for further analysis in this thesis.

Table 2.2: Summary of Selected Works in Speech Emotion Recognition System

<i>Reference</i>	<i>Language</i>	<i>Subjects</i>	<i>Recording</i>	<i>Number of Emotion</i>	<i>Features Extraction</i>	<i>Classifier</i>	<i>Accuracy (%)</i>
Petrushin (1999) [Petr99]	English	30	Simulated	2 (Agitation, Calm)	Pitch, Formant (F ₁ , F ₂ , F ₃), formant bandwidth (BW ₁ , BW ₂ , BW ₃), energy, speaking rate (8)	k-NN, ensembles of neural networks	73.0 (kNN) – 77.0 (ensembles of neural network)
Lee & Narayanan (2005) [Lee05]	English	1187 calls	Natural	2 (Negative, Non-negative)	Pitch, energy, duration, formant (10)	Linear Discriminant Classifier (LDC), kNN	86.3 (kNN with acoustic + language features) – 92.1 (LDC with acoustic + language features)
Schuller et al. (2009) [Schu09]	Germany	51	Natural	2 (Negative, Idle)	MFCC, pitch, energy, zero-crossing-rate (ZCR), root mean square (RMS) frame energy, harmonics to noise (HNR) ratio (384)	Hidden Markov Model (HMM), Support Vector Machine (SVM)	62.3 (HMM baseline 1 state) – 67.7 (HMM with diverse pre – processing strategies)
Callejas & Lopez-Cozar (2008) [Call08]	Spanish	60	Natural	3 (Anger, Bored, Doubtful)	Pitch, Formant (F ₁ , F ₂), formant bandwidth (BW ₁ , BW ₂), energy, rhythm (60)	Multi Layer Perceptron (MLP)	49.0 (no context information) – 96.4 (acoustic and dialogue context)

Dellaert et al. (1996) [Dell96]	English	5	Simulated	4 (Anger, Happiness, Sadness, Fear)	Pitch, speaking rate, slope, smoothed pitch rhythm (15)	Maximum Likelihood Bayes (MLB), Kernel regression (KR), k-NN	56.0 (MLB) – 79.5 (kNN + cooperative composition)
Fernandez & Picard (2003) [Fern03]	English	4	Elicited	4 (FF, FS, SF, SS) F - fast S - slow	Teager Energy Operator (TEO) (20)	Dynamic Bayesian network model and mixture-Hidden Markov Model (m-HMM), Support Vector Machine (SVM) and Multi Layer Perceptron (MLP)	39.3 (Hidden Markov Decision Tree HMDT) – 96.4 (m-HMM)
Petrushin (1999) [Petr99]	English	30	Simulated	5 (Anger, Happiness, Sadness, Fear, Neutral)	Pitch, Formant (F ₁ , F ₂ , F ₃), formant bandwidth (BW ₁ , BW ₂ , BW ₃), energy, speaking rate (8)	k-NN, ensembles of neural networks	35.0 (ensemble neural network for afraid) – 83.0 (ensemble neural network for sadness)
Hu et al. (2007) [Hu07]	Chinese	8	Simulated	5 (Anger, Happiness, Sadness, Fear, Neutral)	Mel Frequency Cepstral Coefficient (MFCC), delta MFCC, energy, Gaussian Mixture Model based on universal background model (64)	Support Vector Machine (SVM)	78.4 (fear for female subjects) – 96.5 (anger for male subject)

Schuller et al. (2009) [Schu09]	Germany	51	Natural	5 (Anger, Emphatic, Positive, Rest, Neutral)	MFCC, pitch, energy, zero-crossing-rate (ZCR), root mean square (RMS) frame energy, harmonics to noise (HNR) ratio (384)	Hidden Markov Model (HMM), Support Vector Machine (SVM)	28.9 (HMM without diverse pre – processing strategies) - 38.2 (HMM with diverse pre – processing strategies)
Nwe et al. (2003) [Nwe03]	Burmese	6	Simulated	6 (Anger, Joy, Fear, Sadness, Surprise, Disgust)	Log Frequency Power Coefficient (LFPC) (12)	4-stage ergodic Hidden Markov Model (HMM)	70.8 (S4 female) - 95.8 (S6 female)
Nwe et al. (2003) [Nwe03]	Mandarin	6	Simulated	6 (Anger, Joy, Fear, Sadness, Surprise, Disgust)	Log Frequency Power Coefficient (LFPC) (12)	4-stage ergodic Hidden Markov Model (HMM)	66.7 (S3 male) - 87.5 (S1 male)
Nicholson et al. (2000) [Nich00]	Japanese	60	Simulated	8 (Anger, Joy, Fear, Sadness, Disgust, Surprise, Teasing, Neutral)	Pitch, Speech power, Linear Predictive Coefficient (LPC), derivative of LPC (20)	Neutral network	56.9 (One Class in one neural network OCON) - 72.3 (Learning Vector Quantization LVQ)

2.4 On-going Cross Cultural Speech Emotion Research

Scientists and engineers have been trying assiduously to identify and quantify parameters needed to correlate emotion in speech. Unfortunately many researchers as discussed in the previous section, have been focussing on the features extraction methods and classifiers to improve such speech emotion recognition system. Based on the psychologists' understanding of emotions, it is such a natural progression for researchers to shift their attention towards culture as one of the important variables to be considered. Motivated by the social constructivist model of emotion perspective, it is noted that different cultural backgrounds shape human responses to the social cues in such a way that both emotion conveyance and perception are uniquely conducted and are dictated by the personal experience and community collective acceptance.

2.4.1 Cultural Effect of Emotion

The expression of emotion is governed by the social norm and have significant impact on the intensity of emotion expression [Ekma71a]. It is also known as *display rules*. Display rules can be defined as a cultural convention that dictates the way emotion should be expressed and the strength of that particular conveyed emotion. It relates closely to three major activities of withholding, disguising and exaggerating expression. This scenario correlates intricately to the way culture is being perceived. For instance, American is more obvious and vocal in expressing anger compared to Asian [Mats01]. This is because Americans are more focused on the individual happiness (individualistic culture) compared to Asian that relies on collective group satisfaction such as group conformity (collectivistic culture) [Miya10]. Due to this notion, Asian often associated anger with rudeness which will affect the overall community feeling when overtly expressed. Thus, it

is no surprise to state that culture does influence the conveyence and perception of emotion.

In addition, the contributions of multi-disiplinary fields to this body of research offers a large room for further analysis and exploration. Such collaboration may introduce new variables to be considered. Hence, it is not an easy task for the international community to follow many directions to understand the influence of culture in speech emotion. Due to this complexity, most researchers in speech emotion recognition arena today do not consider the effect of cultural and language differences in the global accuracy of the speech emotion recognition system. As a consequence, the corresponding supporting empirical research findings on cross cultural emotion recognition from speech are scarce [Sche01a].

The importance of understanding cultural diversity and universality is for personalization of service. When a culturally homogeneous user group is envisaged, it is sound to provide a system or service that cater specifically for that particular group's needs and expectations. It can improve the user satisfaction which can leads to long term user's loyalty. User can feel more relax when using such system and the learning curve can be minimized.

There are common tendencies in the acoustical correlates of basic emotions across different cultures. The universality of the acoustic features allow people with widely varying cultural background to recognize emotion conveyed by other speaker regardless of the language or geographical boundary. This notion is supported by the result of the accuracy level greater than predicted by chance guessing [Elfe02, Sche01a]. Yet, there is a clear distinction of the way emotion conveyed and perceived by each culture such that these information is only shared among the people in the same culture [Dewa08]. Beaupre

mentioned in his paper [Beau05] that the reason for individual to accurately recognize expressions by members of their own ethnic group is because of their ability in encoding and decoding the expressions. For instance, there might be subtle differences in expressive style shared between similar members of a particular cultural group that make it more difficult to decode by other different cultural group member. Therefore, it is imperative that both intra-cultural and inter-cultural variation must be investigated in order to understand the cultural effect on speech emotion.

Juslin and Laukka [Jusl03] revealed that overall decoding accuracy for similar culture-influenced expression was equivalent to 70% accuracy score. However, 7% decrement was achieved for cross-cultural expressions in a five response forced-choice task. In an earlier research finding, Bezooijen et al. studied vocal portrayals of emotion produced by 48 Dutch speakers using standard sentences [Bezo83]. A group of 40 judges each from Netherlands, Taiwan and Japan were required to identify the conveyed emotion. From the result, it was noted that the Dutch speaking judges did significantly better than either the Taiwanese or the Japanese in the speech emotion recognition task. However, the latter groups still outperform the expected accuracy rate by chance probability identification.

2.4.2 Intra and Inter Cultural Assessment

Intra-cultural assessment is a similarity measure that enable different emotions to be detected using the same set of cultural parameters. For instance, given similar culturally-influenced dataset, human are able to discriminate anger, happiness and sadness (example of different emotion). On the contrary, inter-cultural assessment is a dissimilarity measure that enable a similar emotion to be detected across different set of cultural parameters. For

instance, anger can be detected in American, European and Asian culturally-influenced datasets (examples of different culturally-influenced datasets). In the context of this thesis, language is selected as the cultural parameter since the focus is on speech rather than any other emotion modalities, such as body language and facial expression. Both the intra-cultural and inter-cultural variations of speech emotion are also addressed in this thesis. In addition, a method to empirically measure the different culturally-influenced speech emotion data is proposed to allow further analysis of qualitative understanding.

2.4.3 3D Affective Space Model

Grimm et al. [Grim07] extended the idea of a 3D affective space model to automatically classify the speech emotion. Features selected were 46 different speech emotion characteristics describing pitch, energy, speaking rate and spectrum on an utterance-based segmentation level. These features were then mapped on to the three-dimensional emotion space of valence, arousal and dominance. Valence was measured using positive or negative emotion level and arousal was measured using the high or low excitation in the speech. On the other hand, dominance indicated the weakness or the signal strength of the speaker's voice and it is very important to recognize between anger and fear. A rule-based fuzzy logic estimator was then adopted as the classifier to identify anger, sadness and the happiness emotion.

Grimm et al. reported that the proposed system yielded accuracy ranging from 63.5% to 91.9% for speech emotion identification. The study also shows that emotion primitives can provide automatic estimation of the speech emotion in the range of human evaluator, which is proven to be more intuitive and accurate. In addition, Grimm et al. also presented an interesting observation through the co-variance plot of the emotion class

mapping based on valence, activation and dominance as shown in Figure 2.6. It is observed from Figure 2.6 that there is a clear correlation between the cluster distances of one emotion to another in the emotion primitive space. From the Figure 2.6, *A* represents anger, *H* represents happiness, *N* represents neutral and *S* represents sadness for 3 speakers of Electromagnetographic Articulatory (EMA) data corpus.

Figure 2.6: Co-variance Plot for the Emotion Classes [Grim07]

Based on Figure 2.6, it is observed that neutral (denotes as N_1 , N_2 and N_3) occupies the $[0, 0, 0]$ coordinate of the co-variance plot. Such finding is in-line with the psychologists' notion that neutral is an emotionless state where the value for the emotion primitives of valence, arousal and dominance is 0. Anger (denotes as A_1 , A_2 and A_3) is dominating in $[-0.5, 0.5, 0.5]$ coordinate as opposed to sadness (denotes as S_1 , S_2 and S_3) in $[-0.5, -0.5, -0.5]$ coordinate. Anger and sadness share similar valence value that indicate both emotions bring negative effect but obviously differ in arousal and dominance value. Subsequently, happiness (denotes as H_1 , H_2 and H_3) occupies the $[0.5,$

0.5, 0.5] position. Such visualization illustrates clearly the different emotion characteristic of a particular emotion.

Grimm et al.'s finding [Grim07] was in line with Bezooijen et al.'s [Bezo83] observation in an earlier inter-cultural speech emotion research. Bezooijen et al. mapped the emotion findings into a two-dimensional space using a multi-dimensional scaling technique. This technique is designed to provide an insight into the nature of the stimulus characteristic dimensions that determines the underlying similarity between objects. Bezooijen et al. used this similarity measure that can be derived from the number of times the emotional categories have been confused with one another. To ensure that this approach is working, the confusion matrices had to be made symmetrical before the analysis could be carried out. It is interesting to note that the locations of the 10 emotions tested are close to each other in the similarity matrix for all three groups of Netherlands, Japanese and Taiwanese subjects. This finding indicates that speech emotion is universal but it is influenced by cultural variances that are exclusive to a particular culture. Hence, the findings emphasize the importance of cultural dependent understanding to achieve a better speech emotion recognition system performance.

In addition, Wang and Guan [Wang08] conducted a study using audiovisual information to segregate 6 basic emotions similar to Ekman's list [Ekma71a]. The emotions are anger, disgust, fear, happiness, sadness and surprise. They found that the combination of audio and visual perception could achieve an overall recognition rate of 82.1% based on a multi-classifier scheme. However, if audio information is used alone, the proposed method was only able to achieve between 51 % to 66% accuracy. They also concluded that culture and language plays an integral part in emotion recognition. This is because they combined 6 languages in their corpus; namely: English, Mandarin, Urdu, Punjabi, Persian and Italian to ensure that the input data is free from cultural biasness.

Subsequently, in a more comprehensive research, Scherer et. al. [Sche01a] evaluated intercultural differences based on several countries using German speakers. The participants from different countries; namely: Switzerland, Britain, Netherlands, United States, Italy, France, Spain, Indonesia and Germany were given listening task that required them to select the perceived speech emotion of anger, fear, joy, sadness and neutral based on their judgements. Meaningless multi-language sentences approach was used in such a way that the sentences were constructed by combining six European languages syllables that are randomly arranged without any semantic meaning. This approach was adopted to ensure the potential bias of the German judges was eliminated as they may have advantage with prior semantic knowledge of the speech compared to the other judges of the other nationalities. Experimental results reported that the accuracy obtained for the German judges ranged from 74% to 88% except for joy which only yielded 48%. This result was mainly due to the frequent confusion of the joy stimuli with the neutral stimuli (34%). This finding also shows that the Indonesian judges yielded the worst accuracy of 52% (Cohen's Kappa = 0.39) as opposed to Germany with 74% accuracy (Cohen's Kappa = 0.67) and the other Western countries judges (Cohen's Kappa ranging from 0.52 to 0.62). It was interesting to note that the rank order of countries with respect to overall recognition accuracy mirrored exactly the decreasing similarity of languages (with exception to the Swiss sample). The Germanic origin language (Dutch and English) obtained the highest accuracy followed by Roman languages (Italian, French and Spanish) and subsequently followed by Indonesian language that has minimal similarity to the meaningless multi-language sentences. Hence, it is a valid observation that language is one of the cultural parameter and culture does play an important role to the performance of speech emotion recognition. This finding also indicates that human from the same cultural background can better perceive emotion given similar culturally

influenced speech data. Moreover, it is noticed that human are able to recognize emotion cross-culturally due to universality of the emotional acoustic features in the speech. This notion motivated further investigation the inter-cultural and intra-cultural effect to the speech emotion in this thesis.

Figure 2.7: Timeline of Selected Works Speech Emotion Recognition.

The summary of the achievement in culture-influenced speech emotion recognition is presented in Figure 2.7. It is interesting to note that researchers start to move towards automatic speech emotion recognition using computer [Grim07, Wang08] compared to manual segregation [Jusl03, Elfe02, Sche01a, Bezo83]. Considerable attentions need to be focused on this effort in ensuring optimum advantage of better speech emotion understanding. The detail description of the computerized culture-influenced speech emotion recognition is provided in Table 2.3. Hence, in this work, the computer-assisted system is adopted to segregate 3 different emotions; namely: anger, happiness, and sadness, with neutral acting as an emotionless state from three different

culturally-influenced dataset of European, American and Asian cultures. Further experiments need to be conducted so that inter-cultural and intra-cultural assessment analysis can be derived. Details of the three different datasets will be discussed in Chapter 3.

Table 2.3: Summary of the Selected Computerized Culturally-influenced Speech Emotion Recognition System

Reference	Language	Subjects	Recording	Number of Emotion	Features Extraction	Classifier	Accuracy (%)
Grimm et al. [Grim07]	Germany	47	Natural	4 (Anger, Happiness, Sadness, Neutral)	Pitch, energy, speaking rate, intensity, Mel Frequency Cepstral Coefficient (MFCC) (46)	Fuzzy logic estimator, k-Nearest Neighbor	73.4 (sad using kNN) – 91.9 (anger using kNN)
Wang & Guan (2008) [Wang08]	English + Mandarin + Urdu + Punjabi + Persian + Italian	8	Simulated	6 Anger, Happiness, Sadness, Fear, Surprise, Disgust	Pitch, slope energy, speaking rate, MFCC, formant (25)	Gaussian Mixture Model (GMM), kNN, neural network (NN), Fisher's Linear Discriminant Analysis (FLDA)	51.4 (neural network) – 66.4 (FLDA)

2.5 Summary

In this chapter the detailed explanation on the psychologist's view of emotion and the achetype of the emotion are presented. Several related works on speech emotion research with emphasis on the type and method of feature extraction and classifiers used have been

discussed to provide some idea on the direction of this PhD work. Issues relating to culture in speech emotion was outlined to emphasize the focus of the PhD work.

Based on the literature review presented, three key issues will be tackled in this thesis are:

- i. Discrete class speech emotion recognition system as a basis in understanding how most speech emotion recognition systems work and their related problems.
- ii. Understanding the influence of culture on speech emotion using three different culturally-influenced speech emotion datasets will be studied to improve the proposed speech emotion recognition system performance.
- iii. Develop an alternative algorithm to analyze and visualize the correlation of culture and speech emotion.

Hence, in the next chapter, three different culturally-influenced datasets will be introduced; namely: the Berlin dataset which is derived from Berlin Database of Emotional Speech (Emo-db) [Burk05], the NTU_American and the NTU_Asian datasets. These different datasets are used to represent European, American and Asian perspective respectively in providing variation for culture influence analysis. This approach is based on the assumption that culture influences the conveyence and perception of the emotion. The overall architecture of the proposed speech emotion recognition system will be outlined and each module will be thoroughly explained. The frequency domain analysis method based on spectral energy distribution of 40 Mel Frequency Cepstral Coefficient (MFCC) and time domain analysis method of 15 Gaussian Mixture Model (GMM) for features extraction coupled with neural network of Multi Layer Perceptron (MLP) [Bish95, Rume86, Rose62] and fuzzy neural network such as Adaptive Network-based Fuzzy Inference System (ANFIS) [Jang93] and Generic Self-organizing Fuzzy Neural

Network (GenSoFNN) [Tung02a] classifiers are adopted and will be explained in the next chapter.

CHAPTER 3

DATA CORPUSES, FEATURES EXTRACTION METHODS AND CLASSIFIERS

Even though emotion in general has been research into for many years, the international speech emotion recognition research community has been trying assiduously to understand the speech emotion for practical applications and implementations as described in Chapter 1. To support such research objectives, rigorous analyses were conducted to discover relevant factors that affect the performance of such systems. Culture has been discussed in details in Chapter 2 and is considered as one of the integral parameters in speech emotion recognition system. Various techniques in a speech emotion recognition system were also discussed in the previous chapter to provide a comprehensive understanding of state of art approaches in tackling various speech emotion issues. The different features extraction methods as well as classifiers are coupled together to discriminate emotions from either discrete-class emotion or multi-dimension emotion space approach. In this chapter, a cognitive-inspired speech emotion recognition system is proposed to identify three basic emotions; namely: anger, happiness, and sadness with neutral serving as an emotionless state. Such a proposal is derived from the fundamental understanding outlined in Section 2.2.1. Furthermore, the proposed system is designed to facilitate an in-depth understanding of culture influence in speech emotion.

Figure 3.1 depicts a block diagram of the proposed speech emotion recognition system architecture. The system comprises of three major modules of: data corpus

generation and dataset pre-processing, features extraction and classification. These three modules can be observed respectively in Figure 3.1 as block (a), (b) and (c).

Figure 3.1: The General Architecture of the Proposed Speech Emotion Recognition System.

Module 1: Data Corpus Generation and Dataset Pre-processing

This initial module deals with the input to be computed for speech emotion analysis. It is illustrated in Figure 3.1(a). The raw speech emotion data can be collected from three different approaches; namely based on: simulated, elicited or natural speech. Brief discussion on the common data collection approaches was provided in Chapter 2 and a comprehensive description of the data corpus generation used in this work will be presented in the next section. Once the relevant raw speech emotion data is available, few pre-processing steps need to be carried out. It is important to ensure that noise and any unwanted artifacts are removed in such a way that the speech emotion data can be considered as ‘*clean*’ to be used in the analysis. *Clean speech emotion data* in this thesis refers to the following criteria:

- There is no overlapping speech uttered by different speakers at a time,
- All unnecessary pauses or long pauses are removed,
- The background noise is minimal so that it will not distort the perception of the listener,
- Only voiced segments are selected, and
- Only full blown emotions are considered. For instance, hidden emotions such as cynical remark or cold anger are not taken into consideration for the scope of this research work.

After the raw speech emotion data are considered “*clean*”, the selected dataset pre-processing steps are implemented. Firstly, the raw data are filtered and windowed to a smaller frame to facilitate analysis. The rationale of such an approach is to ensure the data consistency and homogeneity so that the results of the experiment are reliable and reflective of the actual emotion. Normalization is also applied to ensure uniformity of the data with the speech amplitude set to $[-1: +1]$ Volt and sampled at 8,000 samples per second. This is to simulate a standard land-line speech telephony system (64 Kbit at a sampling rate of 8 kHz and 8-bits per sample) [Self10] covering the bandwidth of 200 Hz to 3.2 KHz. This ratio is sufficient for normal human to decipher [Rabi93]. The assumption that 8 kHz sampling rate is satisfactory is derived from the observation that human can still segregate emotion during telephone conversation [Lee05].

Module 2: Feature Extraction Module

Features must then be extracted from the pre-processed speech emotion data and transformed into an appropriate format for further processing. This is needed in order to effectively recognize the different emotions from the speech signal. The feature extraction module provides essential information generated from the speech emotion data. The

module is illustrated in Figure 3.1(b). These acoustical and statistical features represent the important parameters that characterize the emotional content in the speech. The extracted features are then used as training and testing data for the classification module to discriminate the different emotions.

For the purpose of this study, the speech signals were transformed from time domain to frequency domain based on either spectral related features of Mel Frequency Cepstral Coefficient (MFCC) [Slan98] or statistical related features of Gaussian Mixture Model (GMM) [Reyn95a]. These two features extraction methods are selected based on its popularity in the speech processing research community, as they provide high accuracy. Some of the recent findings using similar features extraction methods were reported in [Schu09, Wang08, Grim07, Hu07] and the performance of such systems have been discussed in Chapter 2. Thorough discussion and reasoning for the selection of these two feature extraction methods in this thesis will be provided in Section 3.2.

Module 3: Classification Module

Once the relevant features had been extracted, the classification module computes the outcome of the training and testing process to measure the performance of the proposed system. This module is visually presented in Figure 3.1(c). The neural network classifier of Multi Layer Perceptron (MLP) [Bish95, Rume86, Rose62] and two fuzzy neural network classifiers; namely: Adaptive Network-based Fuzzy Inference System (ANFIS) [Jang93] and Generic Self-organizing Fuzzy Neural Network (GenSoFNN) [Tung02a] are adopted. The reason for the selection of these three different classifiers is to mimic the way the brain segregate the different emotions from speech emotion features passed from the cochlea to the auditory cortex. Further descriptions of the classification module will be provided in Section 3.3.

The results are then comprehensively analyzed to facilitate a better understanding of the speech emotion research. Such analysis is also dedicated to provide an indication of the suitability of the proposed system to be implemented for future practical applications. In this work, three different data corpora are employed to represent different cultures in order to study the influence of distinct culture on speech emotion perception. The intra-culture and inter-culture assessments are conducted to determine the similarity and dissimilarity measurements of the selected identified emotions and will be comprehensively addressed in Chapter 5. Based on such analysis, the research work is extended into understanding human behavior (particularly drivers' behavior states (Appendix AIII)) towards a realistic and practical implementation of the proposed speech emotion recognition system. The details of such analysis will be presented in Chapter 6.

3.1 Speech Emotion Annotation and Data Corpus

In any computer system application development, the data used for the system are crucial to ensure the correctness of the system implementation. It can be divided into two groups of data; namely: training and testing data. The training data is used to construct or discover a predictive relationship to facilitate a list of appropriate rules generation applicable to the testing dataset. A testing dataset is used as a yardstick for the system performance that captures the system's ability to identify the intended outcomes. Hence, it is important to have a well-generalized training data so that it can represent correct input for the classifier to generate universal rules.

In this thesis, three culturally-influenced datasets; namely: Berlin, NTU_American and NTU_Asian datasets are employed to represent European, American and Asian cultures respectively. These datasets are used for intra-cultural and inter-cultural effect

analysis on speech emotion recognition. The justification of the three different culturally-influenced datasets usage is to provide different cultural perspectives for the empirical verification of the culture dependency influence on the proposed speech emotion recognition system.

The Berlin, NTU_American and NTU_Asian datasets were gathered through the acted speech data collection approach to provide some control on the perceived emotion. This data collection approach can enable the performance of the proposed system to be scrutinized and offer better understanding to the speech emotion recognition system. The NTU_American and NTU_Asian corpora were collected at The Center for Computational Intelligence (C²iLab), Nanyang Technological University, Singapore whereas the Institute of Communication Science of the Technical University of Berlin collected the Berlin dataset [Burk05]. The Berlin dataset is considered as a standard dataset in speech emotion recognition task and widely known by the researchers in the speech emotion community. Details of the adopted datasets used in the experiments are provided in the following sub-sections.

3.1.1 The NTU_American Dataset

The NTU_American database is our own speech emotion database collected at the Center for Computational Intelligence (C²iLab), Nanyang Technological University, Singapore. It is specifically constructed by collecting speech emotion data samples from movies and television sitcom video clips downloaded from the Internet. The database includes short utterances contributed by four to six speakers covering the basic emotions; namely: anger, happiness, disgust, surprised, sadness with neutral serving as the emotionless state. All speakers are using the American English language as a medium of interaction. The

emotions portrayed by the speakers were manually identified based on the speech semantic, interlocutor facial expression and the basic understanding of the video clip's plot. These video clips were then converted to audio wave file at 8,000 samples per second, mono-streamed and its amplitude value is normalized to $[-1: +1]$ V. To ensure 'pure' speech emotion data, all other artifacts; namely: background music, long pauses and non-verbal sound (sighs, sobbing and etc.) were removed. Figure 3.2 summarized the data collection processes. Human perception test was then needed to verify the correctness of the emotion perceived for the emotional speech samples. Example of the questionnaire used by the perception test to validate the emotion perceived is provided in Appendix A1.

Figure 3.2: The Summary of Data Collection Processes of NTU_American Dataset.

Heterogeneous Human Perception Test

In order to ensure that the video clips obtained for the NTU_American dataset were correctly perceived, heterogeneous human perception test was subsequently carried out. Heterogeneous human perception test refers to different culturally-influenced speech

emotion data were perceived by judges from different cultural backgrounds. In this survey, judges (participants) with different cultural background (Asian cultures) are required to listen to six emotions from different culturally-influenced speech emotion data (American culture). This is to test the hypothesis that human regardless of their culture can recognize emotion through speech.

Participants

In this test a total of 40 healthy human judges ($N=40$) volunteered to provide their perceptual assessment of the culturally-influenced speech emotion audio files and gave written consent according to the university ethical rules and regulations. They were comprised of 11 judges from Nanyang Technological University, Singapore (referred to as the *Validation Group* (VG)) and 29 judges from the International Islamic University Malaysia (referred to as the *Control Group* (CG)). Most of the judges were undergraduate and post-graduate students from the respective institutions with age ranging from 19 and 50 years (VG mean age: 23.3 years and CG mean age: 20.1 years). The ethnicity of both groups of VG and CG judges are Asian and English is not their first language used in their daily conversation. The detailed fragmentation of the judges' ethnicity is presented in Figure 3.3. The rationale of designing VG is to verify the result from CG to ensure a correct understanding of the heterogeneous human perception test. This is because although both groups are from similar cultural background, which is Asian, both groups have different ethnicity that may have its own distinction in perceiving the target emotion.

In CG, 51.7% (15 judges) are male and the remaining 48.3% (14 judges) are female whereas in VG, the male judges dominated about 81.8% (9 judges) compared to 18.2% female (2 judges). Once both of the groups combined, the judges can be divided

into 60% of which is male and the remaining 40% is female. It is noted that a gender difference is hypothetically assumed to be one of the factors that allow one to perceive and convey emotion accurately. Many researchers reported that female are superior to male in emotion recognition and expression [Mont05, Hall04, Hall00]. However, in more recent finding, Hoffman et al. [Hoff10] found that although women were more accurate than man in recognizing subtle facial displays of emotion, there was no difference between male and female participants when recognizing highly expressive stimuli. Hence, some of the researchers explained such scenario by looking into the distinction of the brain's underlying neural network [Dern10, Kemp04, Lee02]. For the purpose of this work, since full-blown speech emotion stimuli is the focus, normal adult from both genders are assumed to have general ability to recognize emotion. Such assumption eliminates the importance of gender segregation in analyzing the human perception test result. Gender grouping should be considered when one handles subtle emotion for a high precision application; for instance: the emotional intelligent quotient (EQ) recognition system. Subsequently, all of the participated judges reported that they experienced neutral emotion prior to the commencement of the human perception test.

Figure 3.3: The Detailed Ethnicity Fragmentation for Different Judges' Group in the Heterogeneous Human Perception Test.

Procedures

The human perception test was conducted in a laboratory environment where the judges can listen to the speech emotion audio files with minimal distraction. This is to ensure that distractions such as visual, auditory or kinesthetic effects that may hinder the judges from making the appropriate selection are strictly controlled. The judges would sit in front of a blank-screen computer and the culturally-influenced speech emotion audio files were played via a headphone. Such measures are taken to ensure that the judges can listen to the speech emotion audio files without prejudice. As the human judges listen to the speech emotion audio file, they have to select one emotion that they strongly perceived from the questionnaire provided. The questionnaire sheet was designed in a 6 force-choice format representing the 5 emotion classes; namely: anger, happiness, sadness, surprised and disgust with the addition of neutral serving as the emotionless state. In order to avoid any mislead perception, each speech emotion audio file's name was labeled using file number that has no relation to any of the respective emotions. In addition, the sequential arrangement of the similar emotion files is not administered to avoid any prior prediction of the emotion pattern based on the audio files arrangement. The human judges are allowed to repeatedly listen to any of the speech emotion audio files to facilitate them in making the appropriate decision. In the final stage, the result from the judges were collected and summarized in a form of confusion matrices.

Results

Table 3.1 shows the results of the heterogeneous human perception test for (a) Control Group (CG) and (b) Validation Group (VG) in perceiving the different emotions of NTU_American dataset. The experimental results showed the CG and VG performance is consistent and yielded almost identical pattern. It can also be observed that most judges

from both CG and VG were able to identify sadness, anger, happiness and neutral with accuracy of at least 76.4%. On the contrary, surprised and disgust suffer badly with accuracy ranging from 20.0% to 60.0%.

Table 3.1: The Confusion Matrix for Heterogeneous Human Perception Performance
based on the NTU_American Dataset

	Happiness	Anger	Disgust	Surprised	Sadness	Neutral
Happiness	76.6	0.0	2.1	13.1	0.0	8.3
Anger	0.0	91.0	4.8	0.0	3.4	0.7
Disgust	0.0	44.8	37.2	5.5	2.8	9.7
Surprised	9.0	0.7	11.0	60.0	0.7	18.6
Sadness	0.7	0.0	0.7	1.4	95.9	1.4
Neutral	1.4	0.0	2.8	0.0	1.4	94.5

(a) Control Group (CG)

	Happiness	Anger	Disgust	Surprised	Sadness	Neutral
Happiness	76.4	0.0	0.0	9.1	0.0	14.5
Anger	0.0	89.1	3.6	0.0	3.6	3.6
Disgust	1.8	29.1	20.0	14.5	3.6	30.9
Surprised	5.5	7.3	1.8	56.4	3.6	25.5
Sadness	0.0	0.0	1.8	0.0	98.2	0.0
Neutral	0.0	0.0	0.0	0.0	1.8	98.2

(b) Validation Group (VG)

From the result in Table 3.1(a) and (b), *Disgust* yielded very low recognition of 20% in VG and 37.2% in CG. This strongly suggests that the judges were confused with

its definition that they might perceive disgust as mild anger thus resulting in 7.6% and 9.1% higher percentage for anger in the CG and VG results respectively. Similarly, *surprised* also scored fairly low perception performance due to the judges' mixed perception. Some of the judges categorized surprised as neutral as they feel no emotion while listening to the speech emotion audio file. On the other hand, other judges reported that they perceived negative surprised as disgust or anger. It is interesting to note that *sadness* is the most correctly perceived emotion with 95.9% and 98.2% recognition performance in CG and VG experimental result respectively. It is due to the fact that sadness has the most acoustically distinct features compared to the other emotions. Based on Table 3.1(a) and Table 3.1(b), the descending arrangement of correctly perceived emotion in the heterogeneous human perception test is as follows:

Sadness → Neutral → Anger → Happiness → Surprised → Disgust

Such result is in-line with the focus of this work that emphasizes on anger, happiness, sadness and neutral. For the NTU_American dataset representation of the 4 respective emotions, 235 utterances were selected for further analysis in this PhD work.

3.1.2 The NTU_Asian Dataset

The NTU_Asian dataset was collected to provide different culturally-influenced dataset based on the local perspective. In this dataset, Malay language was selected based on the immediate availability of the data and judges whose first language is Malay. Such approach was employed to administer homogeneous human perception test in which the judges from similar culture background can perceive the emotion of the culturally-influenced speech emotion data. This is to give a comparison with the results obtained

from the heterogeneous human perception test. Hence, the data collection process was then extended to gather speech emotion data from Asian culture.

Procedures similar to the NTU_American dataset data collection were implemented. The NTU_Asian dataset was constructed from Malaysian television drama series downloaded from the Internet. Six speakers contributed for each speech emotion class and all of the speakers use Malay language to interact. Subsequently, the manual selection of speech emotion data and pre-processing steps were conducted to segregate the speech emotion audio file to its respective emotion class.

Homogeneous Human Perception Test

This test was conducted to test the hypothesis that human from similar culture background are able to correctly identify emotion from culturally-influenced speech emotion data with high accuracy.

Participants

Forty Asian human judges ($N=40$) aged 19 to 32 (mean age of 22.4 years) participated in the study. They were recruited via advertisements posted at the International Islamic University Malaysia (IIUM). The majorities of the subjects is the undergraduate and post graduate students and were not previously involved in the NTU_American human perception test. The rationale of this decision is to make sure that the judges are not biased from the previous culturally-influenced human perception test. The homogeneous human judges group were comprised of 35% male (14 subjects) and 65% female (26 subjects). As discussed previously in heterogeneous human perception test, gender is inconsequential in analyzing perception on highly expressive speech emotion data. The

ethnicity of the homogeneous human judges group is dominated by Malay with 87.5% are Malaysian Malay (35 subjects) and 12.5% are Indonesian Malay (5 subjects). All of the judges are using Malay language as their mother tongue and frequently using it to communicate daily.

Procedures

Similar procedure as the previous heterogeneous human perception test was adopted. This is to ensure consistency and correctness of the results obtained. It is imperative to ensure that similar procedures are followed accordingly in order to eliminate biasness and misinterpretation of the result.

Results

Table 3.2: The Confusion Matrix for Homogeneous Human Perception Performance
based on the NTU_Asian Dataset

	Happiness	Anger	Disgust	Surprised	Sadness	Neutral
Happiness	70.6	1.7	10.6	10.6	1.2	5.3
Anger	1.3	86.3	4.6	4.1	2.0	1.7
Disgust	8.0	13.0	46.5	5.0	5.5	22.0
Surprised	3.3	6.7	8.3	68.8	3.3	9.6
Sadness	1.1	0.8	1.5	1.3	87.8	7.5
Neutral	3.6	2.1	4.6	1.1	3.6	85.0

Table 3.2 shows the confusion matrix of the NTU_Asian dataset human perception test. It is observed that sadness, anger, neutral and happiness were the top

emotions correctly perceived with at least 70% accuracy. Surprised and disgust were recorded with much lower performance. Such outcome is consistent with the result yielded from the heterogeneous human perception test for the NTU_American dataset. However, the distribution of the result in Table 3.2 seems to have more spread compared to the NTU_American dataset human perception test. It shows that the homogeneous human judges are more sensitive to the inconspicuous changes in the similar culturally-influenced speech emotion data. This could be due to the human judges' pre-conceived semantic knowledge of the utterances as well as the refinement of the perception skill in the communication using similar day-to-day language. Hence, it can be concluded that there is an intricate relationship between culture and performance of the emotion perception. Although human in general are able to categorize different emotions, the intra-culture and inter-culture factors influence the level of sensitivity of the emotion perception. Human from similar culture background can better detect the subtle changes in the utterance of the homogeneous culturally-influenced speech emotion data when compared to the other culturally-influenced speech emotion data.

Subsequently, both the homogeneous and heterogeneous human perception test results validated this thesis focus to use the three basic emotions of: anger, sadness and happiness as well as neutral since human can easily identify them with at least 70% accuracy. Neutral was also selected to represent emotionless state in this study. Hence, 276 utterances were selected from the NTU_Asian dataset to represent the respective emotions for further analysis.

3.1.3 The Berlin Database of Emotional Speech (Emo-db)

The Berlin Database of Emotional Speech [Burk05] collected by the Institute of Communication Science of the Technical University of Berlin is a standard dataset used for speech emotion recognition analysis. It comprises of 10 sentences which cover 7 emotion classes; namely: anger, fear, happiness, sadness, disgust, boredom and neutral. The speech data were excerpt from everyday communication and could be interpreted in all emotional contexts without semantic inconsistency. The spoken content is pre-defined and presented by 5 male and 5 female native German professional actors. The recording was conducted under studio conditions with high-quality recording equipments in an anechoic chamber. The complete database has been evaluated through human perception test with 20 human judges with respect to the emotion perceived and the degree of naturalness. The degree of naturalness reflects the listeners' confidence level of the spoken speech emotion. This was to rule out any exaggerated speech. It was reported that 84.3% mean accuracy was recorded for all emotion in human perception test. For the analysis, 5 to 8 speakers' speeches were used for the emotion class with 385 utterances.

Table 3.3: The Confusion Matrix for Homogeneous Human Perception Performance
based on the Berlin Dataset [Burk05]

Emotion	Recognition level (%)	Naturalness level (%)
Anger	100.00	98.00
Sadness	98.00	82.12
Happiness	100.00	84.60
Neutral	99.00	95.50
Mean	99.25	90.06

The emotion recognition result for the selected data and its degree of naturalness is presented in Table 3.3. Other speech emotion data were discarded as the focus is only on basic emotion of anger, happiness and sadness with neutral acting as emotionless state in this study.

3.2 Features Extraction Method

Features are the representation of the important characteristics extracted from the pre-processed speech emotion data. Many state of the art features extraction techniques and methods adopted by the researchers to find the optimal value as input to the classifier had been discussed in Chapter 2. From this understanding, the frequency domain analysis using the Mel Frequency Cepstral Coefficient (MFCC) [Slan98] and the time domain analysis using the Gaussian Mixture Model (GMM) were employed [Reyn95a].

In both selected features extraction methods, the input signals were first framed and filtered. This is to ensure that the time information can be captured as it carries a lot of emotion information. Human are able to categorize emotion given sufficient length of speech emotion data. Assuming an input speech data sequence s , the speech vector ($\mathbf{S}$) is given by Equation (3.1)

$$\mathbf{S} = \cdots + [s_{(n-1)N+1+\lambda}, s_{(n-1)N+2+\lambda}, \dots, s_{nN}, s_{nN+1}, \dots, s_{nN+\lambda}]^T + \cdots \quad (3.1)$$

$$\cdots + [s_{nN+1}, \dots, s_{(n+1)N}]^T + \cdots + [s_{(n+1)N-\lambda+1}, \dots, s_{(n+1)N}, s_{(n+1)N+1}, \dots, s_{(n+2)N-\lambda}]^T + \cdots$$

where: N is the number of samples per block of data that is processed,

n is the block number,

λ is the number of samples that is overlapped between the present and the previous block of data, and

T is the transpose of the matrix.

The first set of data in Equation (3.1) represents the previous frame of data followed by the present frame of data and subsequently followed by the next frame of data. It can be seen from Equation (3.1) that when there is no overlap (i.e. $\lambda = 0$), the string of data s is just a data sequence with N number of samples per frame of data.

Further *smoothing process* was implemented as an attempt to reduce the noise due to data framing at the overlap between the adjacent frames. In smoothing, the data points of a signal are modified so that individual points that are higher than the immediately adjacent points (presumably because of noise) are reduced, and points that are lower than the adjacent points are increased. This naturally leads to a smoother signal. Equation (3.2) describes the different shapes for the smoothing algorithms where M_1 and M_2 are $\lambda \times \lambda$ diagonal matrices. The M_1 matrix is used for the previous frame overlap ($n-1$ frame) and the M_2 matrix is used for the current frame (n frame). Smoothing shapes in this case refers to the different smoothing function that can be used to minimize the artifacts due to abrupt change of the input speech signal. Different smoothing shapes; namely: triangular, cosine and exponential decay are presented in Equations (3.3) to (3.5). The need of different smoothing shapes is dependent on the data to be smoothed.

$$M_1 = \begin{bmatrix} m_{11} & 0 & \cdots & 0 \\ 0 & m_{22} & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & m_{\lambda\lambda} \end{bmatrix} \text{ and } M_2 = \begin{bmatrix} 1-m_{11} & 0 & \cdots & 0 \\ 0 & 1-m_{22} & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1-m_{\lambda\lambda} \end{bmatrix} \quad (3.2)$$

where

m_{xx} defines the different shapes of the smoothing function and can only take value from 0 to 1.

To achieve a triangular function, m_{xx} is defined as Equation (3.3):

$$\text{triangular } m_{xx} = 1 - \frac{x}{\lambda}, \quad \text{for } x = 1 \dots \lambda \quad (3.3)$$

The cosine smoothing function is described by m_{xx} of the form:

$$\text{cosine } m_{xx} = \frac{1 + \cos\left(\frac{x}{\lambda}\pi\right)}{2}, \quad \text{for } x = 1 \dots \lambda \quad (3.4)$$

and for exponential decay, m_{xx} is defined as:

$$\text{exponential decay } m_{xx} = \frac{e^{-\left(\frac{x}{\lambda}\pi\right)} - e^{(-1)}}{1 - e^{(-1)}}, \quad \text{for } x = 1 \dots \lambda \quad (3.5)$$

Based on Equations (3.3) to (3.5), the overlapped data sequence can be handled in various ways. However, for the purpose of this study, the moving average method is adopted. An array of raw data, $\mathbf{S}$ can be converted to a new array of smoothed data by taking the average value of the overlapped sequence of the previous data frame, $\mathbf{S}_{n-1}$ and the present data frame, $\mathbf{S}_n$. This method is simple to implement and sufficient to suppress any unwanted artifacts in the speech emotion data.

3.2.1 Mel Frequency Cepstral Coefficient (MFCC)

The discussion of the most relevant features to discriminate speech emotion is one of the topics addressed by many researchers. The most common acoustical features to be extracted can be grouped to four categories; namely: pitch, energy, MFCC and duration. Although there was no clear evidence that one was more important than the other, the MFCC features are always observed to be a good feature extraction method for acoustic features in speech emotion [Bat11, Wu10, Mowe09, Grim07]. Wu et al. reported that the

relative importance of acoustic features is arranged in the order of MFCC > Energy $\approx$ Pitch > Duration [Wu10]. Such observation is consistent with the result reported by Batliner et al. [Bat11] and Mower et al. [Mowe09]. Therefore, it is sound to assume that MFCC features are relevant for the task of speech emotion discrimination.

To date, human auditory system outperforms current machine-based system for speech emotion recognition system. Human are known to be better at categorizing different emotion in almost automatic response. Motivated by this fact, Mel Frequency Cepstral Coefficient (MFCC) features extraction method was adopted as it is based on the approximation of critical bands in the human auditory system.

MFCC exploits the human auditory frequency response. The features capture the most important parts of the spectral envelope perceptually in the audio signals. The MFCC algorithm was derived from the idea of homomorphic filtering in the human cochlea [Skow02]. Cochlea is the organ that translates the sound energy into the nerve impulses for the neural decoding. It does not just acts as an input device but also filters the sound, extracts the features and passes the relevant features (features selection) to the brain. The importance of cochlea function can be seen among patients suffering from cochlea impediment. The patients were observed to have decreased ability in perceiving speech due to his/her inability to differentiate voice and background voice although hearing aid is used [Hear10].

The cochlea structure consists of fluid-filled caverns (the Scala Vestibuli and Scala Tympani) separated by Basilar Membrane as shown in Figure 3.4 and illustrated in Figure 3.5(e). Figure 3.5 shows the general anatomy of human ear.

Figure 3.4: The Schematic Diagram of Cochlea

Figure 3.5: Anatomy of Human Ear [Live07]

Consider the scenario of an interlocutor talking to a listener. The emotion is propagated using the sound signal. This acoustic wave is transmitted from outer ear (Figure 3.5(a)) via Auditory Canal (Figure 3.5(b)) to the middle ear (Figure 3.5(c)). The stapes (Figure 3.5(d)) converts the acoustic energy to mechanical energy and excites the oval window on the outside of cochlea. This excitation vibrates the liquid in the Scala Vestibuli (upper chamber of cochlea) and in turn vibrates the liquid in Basilar Membrane and the Scala Tympani as depicted in Figure 3.4. The Basilar Membrane acts like a Fourier Transform with each location tuned to resonate at a particular frequency, ranging from high frequency at the base end to low frequency at apex end. Therefore, the cochlea

can be viewed as a cascade of filters, each tuned to a different center frequency [Li09, LaMa06]. Basilar Membrane velocity is then detected by hair cells (Figure 3.5(f)). The hair cells are the acoustic sensor cells that convert the vibration into an auditory nerve firing rate. These nerves are compressed and only the appropriate nerves are fired and decoded by the auditory cortex in the brain.

MFCC is derived from the Fourier Transform of the audio signal where the frequency bands are logarithmically positioned on the Mel scale. It is inspired from the Basilar Membrane function in the cochlea that can be summarized as a bank of band-pass filters tonotopically ordered along the length of cochlea [Yang92]. It is developed in such a way to closely approximate the response of the human auditory system when compared to a linearly spaced frequency band of Fourier Transforms. The spectral analysis performed by MFCC is implemented based on a short-time spectrum where the input speech emotion signals are divided into frames of size ranging from 128 to 2048 samples. Each frame is filtered using a Hamming window and transformed to the relative frequencies using the Fast Fourier Transform (FFT) to compute the frequency spectrum. The Fourier basis audio signals are decomposed into superposition of the finite number of sinusoids. The power spectrum bins are grouped and smoothed according to the perceptually motivated Mel-frequency scaling. The spectrum is segmented into 27-critical bands by means of a filter bank that typically consists of overlapping triangular filters in such a way that low frequencies are given more weight than high frequencies. This is to conform to the cochlea mechanism of extracting appropriate features for neural decoding. Finally, a Discrete Cosine Transform (DCT) is applied to the log of the filter bank outputs. Such technique results in the vectors form of the de-correlated MFCC features. The need arises when the processed frequency decomposed signals need to be converted

back and to ensure that no information is lost through the forward transforms [Li09].

Figure 3.6 shows the block diagram of the MFCC features extraction method.

Figure 3.6: The Block Diagram of MFCC Features Extraction Method.

There are many implementations of MFCC by several researchers for different applications with different focuses [Skow04, Slan98, Youn95, Davi80]. These approaches vary in several parameters; such as: number of filters, filters' shape, bandwidth and the manner in which the spectrum is warped, 'interest' frequency range, selection of actual subset, and number of MFCC co-efficient derived. A study by Ganchev et al. [Ganc05] found that Slaney's approach [Slan98] gives slightly better performance than others. Thus, for the purpose of this study, the approach by Slaney's Auditory Toolbox in Matlab [Slan98] is adopted.

The MFCC implemented by Slaney are calculated using 40 coefficient traingular-shaped filter banks. The first 13 bands are positioned linearly between 100 Hz to 500 Hz with step size of 33.33 Hz. The remaining 27 bands are logarithmically spaced in the range of [536:3200] Hz with a step given by logStep of Equation (3.6). For the purpose of this study, a 8 Khz sampling frequency is used with a frequency ranging from 66 Hz to 3427 Hz as shown in Equation (3.6) as described in [Slan98].

$$\logStep = \exp\left(\ln\left(\frac{f_{c_{40}}}{1000}\right) / numLogFilt\right) \quad (3.6)$$

where :

$numLogFilt$ is the 27 logarithmically spaced filters, and

$f_{c_{40}}$ is the centre frequency of the last logarithmically spaced filters
(3200Hz),

Each one of these equal area triangular filters is defined by Equation (3.7).

$$H_i(k) = \begin{cases} 0 & \text{for } k < f_{b_{i-1}} \\ \frac{2(k - f_{b_{i-1}})}{(f_{b_i} - f_{b_{i-1}})(f_{b_{i+1}} - f_{b_{i-1}})} & \text{for } f_{b_{i-1}} \leq k \leq f_{b_i} \\ \frac{2(f_{b_{i+1}} - k)}{(f_{b_i} - f_{b_{i+1}})(f_{b_{i+1}} - f_{b_{i-1}})} & \text{for } f_{b_i} \leq k \leq f_{b_{i+1}} \\ 0 & \text{for } k > f_{b_{i+1}} \end{cases} \quad (3.7)$$

where:

i is the i-th filter, with value of = 1, 2, 3, ..., M ,

f_{b_i} are $M + 2$ boundary points that specify the M filters,

k corresponds to the k-th co-efficient of the N-point FFT with value of = 1, 2, ..., N

f_{b_i} is the center of frequency position,

$f_{b_{i-1}}$ is the cut-in frequency , and

$f_{b_{i+1}}$ is the cut-off frequency.

The boundary points are expressed in the form of the following terms: f_{b_i} , $f_{b_{i-1}}$ and $f_{b_{i+1}}$ as specified in Equation (3.7). The key to equalization is the area under the filters, defined as Equation (3.8).

$$\frac{2}{(f_{b_{i+1}} - f_{b_{i-1}})} \cdot \quad (3.8)$$

The filter bank is normalized in such a way that the sum of co-efficient for every filter is equal to one. Thus, the i-th filter satisfies Equation (3.9).

$$\sum_{k=1}^N H_i(k) = 1 \quad \text{for } i = 1, 2, \dots, M \quad (3.9)$$

Next, the equal area filter bank is employed in the computation of the log-energy output using Equation (3.10).

$$X_i = \log_{10} \left(\sum_{k=0}^{N-1} |X(k)| \cdot H_i(k) \right) \quad \text{for } i = 1, 2, \dots, M \quad (3.10)$$

Finally, the DCT in the following equation provides the Slaney's MFCC parameters as in Equation (3.11).

$$C_j = \sum_{i=1}^M X_i \cdot \cos \left(j \cdot (i - 0.5) \cdot \left(\frac{\pi}{M} \right) \right) \quad \text{for } i = 1, 2, \dots, J \quad (3.11)$$

3.2.2 Gaussian Mixture Model (GMM)

GMM in this thesis is referred to the spectral GMM that is widely used by the researchers in speech emotion field. In the recent survey [ElAy11], El Ayadi et al. reported that GMMs are more appropriate for speech emotion recognition when only global features to be extracted from the training utterances. However, GMMs cannot model temporal

structure of the training data since all the training and testing equations are based on the assumption that all vectors are independent. In addition, El Ayadi et al. also had included many other researchers work using GMM indiscriminating emotion with accuracy ranging 74.8% to 81.9% [ElAy11]. Therefore, it is possible to use spectral GMM to detect emotion as reported in the preliminary work of the thesis.

Gaussian Mixture Model (GMM) can represent general speech spectral shapes and can provide a smooth approximation to the underlying sample distribution of the speech signal. Such characteristic is capable of reflecting some general speaker dependant vocal tract configuration which explains its succesfulness in the speaker identification field [Reyn95b]. One of the advantage of using GMM is its capability to model arbitrary densities [Reyn00, Reyn95a]. The GMM assumes that the probability distribution of the mixture density used for the likelihood function of a D-dimensional feature vector, x is defined as Equation (3.12).

$$p(x|\lambda) = \sum_{i=1}^M w_i p_i(x) \quad \text{for } i = 1, 2, \dots, M \quad (3.12)$$

where:

w_i are the mixture weights,

$p_i(x)$ are the multivariate normal distribution, and

M is the number of the components.

Each component densities are parameterized as Equation (3.13)

$$p_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_i)^T (\Sigma_i)^{-1} (x - \mu_i) \right\} \quad (3.13)$$

where:

Σ_i is $D \times D$ covariance matrix, and

μ_i is a mean vector.

The mixture weight, w_i however, must satisfy the constraint $\sum_{i=1}^M w_i = 1$.

The complete density model parameters for $i = 1, 2, \dots, M$ can be written as Equation (3.14). M is also sometimes known as the gaussian mode.

$$\lambda = (w_i, \mu_i, \Sigma_i) \quad (3.14)$$

where:

w_i is the mixture weight,

μ_i is a mean vector, and

Σ_i is the covariance matrix.

Mean vector, μ_i of the w_i component density can represent the shape of the i th relative frequency of the time domain signal and variation of the histogram can be represented by covariance matrix Σ_i .

The process of GMM begins with the speech signal inputs being divided into frames with size ranging from 512 to 2048 samples. Preliminary empirical study shows that a 80% data overlapping between frames provides the best performance. Each frames are filtered using a Hamming window and subsequently transformed to the relative frequencies using the histogram method [Figu02]. From the histogram, selection of the number of bin will affect the resolution of the plot. The higher the number of bins used, the finer the pattern plot. The GMM method is then used to approximate the Probability

Density Function (PDF) of the histogram. The number of GMM components selected plays an important role in determining the accuracy and training time for a specified application [Waha07, Robe98]. A minimum of 2 Gaussian components ($M = 2$) are required to have a meaningful gaussian estimation while a maximum of up to 6 gaussians ($M = 6$) are sufficient. The overfitting problem would occur if too high a number of Gaussian components is generated to estimate the histogram. On the contrary, underfitting problem would occur if insufficient number of Gaussian components are used.

There are several methods available for estimating the parameters of a GMM and the most popular method is the maximum likelihood (ML). The ML parameter can be iteratively determined using the Expectation Maximization (EM) algorithm and the best gaussian distribution fit can be derived [Duda00, Demp77]. The EM algorithm relies on the existence of immediate variables, called latent data, in the estimation problem. It starts off with an initial model $\bar{\lambda}$ that must conform to $P(X | \bar{\lambda}) \geq P(X | \lambda)$ state. The model $\bar{\lambda}$ then becomes the initial model for the next iteration and the process is repeated until the specified convergence threshold is reached. The EM computes the expected value of the latent variables in *Expectation (E)* step and also computes the maximum likelihood estimates of the parameters in *Maximization (M)* step. The mixture weight, w_i , mean vector, μ_i and covariance matrix $\sum_i$ of each p_i component density forms the features extracted per Gaussian component λ_i . Therefore, in the GMM method, the number of features extracted depends on the number of Gaussian components used to estimate the Probability Density Function (PDF). Figure 3.7(b) shows a typical histogram plot of a speech emotion signal (Figure 3.7a) sampled at 8 KHz with 2048 frame size with an 80%

overlap using 100 bins. The GMM estimation of the speech PDF using 5 GMM components are also depicted at the bottom of Figure 3.7(c). Further discussion on the number of features extracted using this method is discussed in details in Chapter 4 for speech emotion verification task.

(a) Time Domain Signal of the Selected Speech Emotion Frame

(b) Histogram of the Selected Speech Emotion Frame

(c) The Estimated Gaussian Distributions of Selected Speech Emotion Frame

Figure 3.7: Example of GMM Implementation.

3.3 Classification

As described in the general architecture of the proposed speech emotion system of Figure 3.1, classification module provides the performance information of such system. The extracted features from culturally-influenced speech emotion data are trained and tested with the selected classifiers. Subsequently, the result of each classifier are analyzed and compared for detailed analysis.

Many researchers in speech emotion field have implemented various types of classifier and some selected results had been discussed in Section 2.3. Statistical analysis classifiers such as Radial Basis Function (RBF), naïve Bayes, Decision Tree and others had been considered in many literatures as presented in Chapter 2. However, current trend shows that researchers are looking actively at Neural Network (NN) as an alternative solution [Call08, Fern03, Nich00, Petr99] (For a review, refer Table 2.2). The reasons for NN popularity are due to several factors; namely: 1) NN can achieve reasonable degree of accuracy, 2) NN is able to model non-linear functions with large numbers of variables and 3) NN is also intuitively appealing as it is derived from the modeling of a crude low-level model of biological neural system that allows learning by examples.

NN attempts to mimic the fault-tolerance and capacity to learn of biological neural systems [Patt96]. It tries to emulate the brain architecture. The brain is principally comprised of 10 billions of interconnected neurons which are able to propagate an electrochemical signal. When a neuron is activated, it fires an electrochemical signal to the other interconnected neurons. The changes of the synaptic connection are postulated as learning [Hebb49]. Similarly, NN receives a number of inputs via a connection that has a weight. These weights correspond to synaptic efficacy in a biological neuron. The adjustment of the gradient are iteratively changed until it meets the pre-set target goal.

Hence, in this work, a neural network classifier of Multi-Layer Perceptron (MLP) is employed to simulate the way the brain processes the extracted acoustical information for neural decoding in order to discriminate the underlying emotion in the speech signals.

However, NN has its own disadvantages. One of the major drawbacks is that NN treats the whole nonlinear process as a black box. As a consequence, best achievable degree of accuracy may not be achieved. Such a drawback can be resolved using the Fuzzy Neural Network (FNN) approach. FNN is actually a combination of neural network and fuzzy inference system. It offers an attractive compromise between adaptability of the neural network (NN) and interpretability of a Fuzzy Inference System (FIS). The FNN can self-adjust the parameters of the fuzzy system using neural-network-based learning algorithm and the fuzzy logic system components can handle vagueness by making inferences through fuzzy rules in a structurally transparent manner for better generalization [Tung02a]. The hybrid of these two systems can provide strong reasoning schemes with a structure which can learn rules. In this study, the focus is given on two FNN systems; namely: Adaptive Network-based Fuzzy Inference System (ANFIS) and Generic Self-organizing Fuzzy Neural Network (GenSoFNN). ANFIS are able to generate a huge number of rules from the input data hence it can be very difficult to use such rules in a practical manner. On the contrary, GenSoFNN can generate a lesser number of rules than ANFIS and the rules can be easily interpreted in the IF... THEN rule manner.

It is noted that all classification in this work are conducted off-line as opposed to on-line learning. The off-line learning refers to the training data that are stored first before the training exercise can be executed. This is imperative because the speech emotion data that are used need to be labeled prior to its usage (supervised learning). As the emotion is

commonly expressed freely, the data labeling process is always needed to be conducted manually before it can be used in training. The off-line learning is based on a computation of the objective function hence error and confidence interval can be measured. Such situation is unsuitable to on-line learning. In on-line learning, each case is discarded after it is processed and the weights are updated. Hence, the objective function as well as error and confidence interval on the validation set for a fixed set of weights cannot be computed since these datasets are always changed with the new addition of data to the training data set.

3.3.1 Multi Layer Perceptron (MLP)

Multi Layer Perceptron (MLP) [Bish95, Rume86, Rose62] is adopted due to its ability to determine non-linear boundaries separating the target class. Target class here refers to the speech emotion instances that are labelled manually prior to the training exercise for supervised learning. The MLP uses a combination of several perceptron layers that are interconnected to each other. It exhibits a high degree of connectivity which is determined by the synapses of the network.

Figure 3.8: A Multi Layer Perceptron Network Architecture

Generally MLP consists of three main layers that are known as input layer, hidden layer and output layer as depicted in Figure 3.8. In the input layer N_{input} , the number of neuron must be equal to the number of the features used. For instance, in the case of MFCC features extraction method, 40 coefficients are generated hence the network needs 40 neurons to compute the data. Each input data will be given a weight by the network to be used by the hidden layer N_h to compute the non-linear element using an appropriate activation function. The output from the hidden layer were then aggregated by the output layer N_{output} . These outcomes are used to determine the error between the output and the expected values. The mean square error (MSE) is applied to adjust the weights to minimize the error. MLP adjusted and updated the gradient of each neuron using the back-propagation algorithm to minimize the MSE. The iteration for learning are stopped once the goal or MSE value is met or the number of iteration has reached its maximum.

Consider the MLP network with d inputs and m outputs. The input and output vectors are respectively written as Equation (3.15):

$$\begin{aligned} x &= [x_1, x_2, \dots, x_d]^T \\ o &= [o_1, o_2, \dots, o_m]^T \end{aligned} \quad (3.15)$$

where d and m are the number of input features.

Let w_{ij} denotes the weight that connects output neuron i with input neuron j , the activation value for output neuron i can be shown as in Equation (3.16)

$$v_i = \sum_{j=1}^d w_{ij} x_j \quad (3.16)$$

After receiving the activation, the neurons will perform the processing of the activation signal. The output o_i is simply a transformation of nonlinear activation function to the v_i in Equation (3.17)

$$o_i = \varphi(v_i) = \varphi\left(\sum_{j=1}^d w_{ij}x_j\right) \quad (3.17)$$

The network will be tuned for convergence by using back-propagation algorithm that used mean square error (MSE) to evaluate performance.

However, it is important to note that the error function used in training MLP is dependent on the response of nonlinear elements. Globally optimum solution can be reached with the use of iterative first order gradient descent technique of the backpropagation algorithm. The algorithm however suffers from the drawbacks of gradient descent methods, such as slow convergence due to its tendency to oscillate in the proximity of the solution. In addition, due to its susceptibility to local minima, the technique may converge to a spurious solution different from the true global minimum. Hence, the network training could be a lengthy process with no guarantee that the state of the network obtained at convergence is the optimum for the intended application.

3.3.2 Adaptive Network-based Fuzzy Inference System (ANFIS)

Adaptive Network-based Fuzzy Inference System (ANFIS) [Jang93] was adopted as one of the fuzzy neural network classifiers to compare the different classifiers performance in identifying emotion from speech. ANFIS is a hybridization of neural network and fuzzy system that optimizes the parameters of the given fuzzy inference system by applying learning procedure to the training dataset. It uses the hybrid algorithm, which is the mixture of least mean square (LMS) error minimization to determine the consequent parameters and back-propagation to learn the premise parameters. Thus two passes are required. In the first pass, the premise parameters are assumed to be fixed and the optimal consequent parameters are estimated by the iterative Least Mean Square procedure. In the

second pass the consequent parameters are assumed to be fixed and back-propagation is used to learn the premise parameters. This method will ensure fast convergence and the global minimum consequence parameter can be reached.

ANFIS applies the first-order Sugeno model architecture due to its efficiency and transparency. It also promotes the use of fuzzy inference system as the engine to make a rule-based approach from the given input-output relationship. It consists of five layers with each layer performing different processes; namely: fuzzification, multiplication, normalization, execution and de-fuzzification respectively. The five-layer architecture of ANFIS is illustrated in Figure 3.9.

Figure 3.9: The ANFIS Architecture

Consider two input variables, x and y and a single output f_i . A rule in the first order Sugeno Fuzzy Inference System (FIS) has the form as shown by Equation (3.18)

$$\begin{aligned} \text{IF } x \text{ is } A_1 \text{ AND } y \text{ is } B_1, \text{ THEN } f_1 &= p_1 x + q_1 y + r_1 \\ \text{IF } x \text{ is } A_2 \text{ AND } y \text{ is } B_2, \text{ THEN } f_2 &= p_2 x + q_2 y + r_2 \end{aligned} \quad (3.18)$$

where:

A_i, B_i are the membership functions, and

p_i, q_i, r_i are the consequent parameters.

Every node i in the first layer have node function as in Equation (3.19)

$$O_i^1 = \mu_{A_i}(x) \quad \text{for } i = 1, 2, \dots, m \quad (3.19)$$

where O_i^1 is the membership function (MF) of A_i and it specifies the degree to which x satisfies the quantifier A . Each input variable m denoted as the total number of MF. Tuning the values of these parameters will vary the MFs and change the FIS behavior.

Layer 2 multiplies the outputs of Layer 1 and sends the outcome to the next layer. Its output node w_i represents the firing strength of a rule given in Equation (3.20). m MFs are associated with each input, so the input space is partitioned into m^2 fuzzy subspace which is governed by a fuzzy If-Then rule.

$$w_i = \mu_{A_i}(x) \mu_{B_i}(y) \quad \text{for } i = 1, 2, \dots, m^2 \quad (3.20)$$

Normalization of the MFs are then conducted in Layer 3 as shown in Equation (3.21) where the ratio of the j^{th} rule's firing strength to the sum of all rules' firing strengths is computed.

$$\bar{w}_i = \frac{w_i}{\sum_{i=1}^{m^2} w_i} \quad (3.21)$$

Layer 4 executes the consequent part of the fuzzy rules where $\bar{w}_i$ is the output from Layer 3 and $\{p_i, q_i, r_i\}$ is the parameter set. The node function is given in Equation (3.22).

$$O_i^4 = \bar{w}_i f_i(x) = \bar{w}_i (p_i x + q_i y + r_i) \quad (3.22)$$

The final output of ANFIS is obtained from Layer 5 as in Equation (3.23) by summing up the outputs in Layer 4.

$$f = \sum_{i=1}^{m^2} \bar{w}_i(x) f_f(x) \quad (3.23)$$

ANFIS applies clustered data for input instead of crisp raw data for the fuzzification process. This is because the clustered data enables ANFIS to approximate the input value and generate generalized rules hence resulting in faster processing time. ANFIS offers three data clustering approaches; namely: grid partitioning, subtractive clustering and fuzzy c-means. These clustering algorithms require different initialization value for the input. For instance, the grid partitioning approach is only useful if the number of features is no more than 6 or 7. This is because ANFIS will generate an exponentially large number of rules that may be irrelevant and hence produce overfitting if the number of input features is too large.

The fuzzy c-means approach requires the number of clusters for the initialization of the input data. However, since no prior knowledge on the optimal number of clusters is available, specifying the number of cluster information would pose a challenge while using brute-force approach to determine an optimal number of cluster is time-consuming.

The other clustering approach is subtractive clustering. It is a fast, one-pass algorithm that is able to estimate the number of fuzzy clusters and cluster center based on the radii value provided. Each data point is considered as a potential cluster center and the clusters are formed based on the density measure of the defined neighborhood radius covered within the cluster radii. It would take the input-output training data and generate the Sugeno-type fuzzy inference system (FIS) that models the data behavior. This

estimation is used for FIS initialization by ANFIS as a prior information. The high radii value covers a large portion of the data space hence generated a small number of cluster. On the other hand, small radii value means a high number of cluster will be formed. Optimum number of cluster is dependant on the data used for the analysis. In the initial stage of subtractive clustering, the first cluster center is assigned to the point that has the highest density measure. However, the density measure for each data point is repeatedly revised until a stopping criterion is met. The detailed explanation of this technique can be found in [Chiu94]. For the purpose of this study, subtractive clustering [Yage94] approach is selected to estimate the optimum number of cluster of the input data.

Figure 3.10: Block Diagram of the N-class mANFIS.

One of the major drawback in ANFIS is that it can only generates one output at a time. Thus, an equivalent numbers of ANFIS classifiers with the number of the output decision threshold would be required to classify multiple output system. Multiple ANFIS or better known as mANFIS aggregates many independent ANFIS as a complete system network where the output of each individual ANFIS will be separately generated and combined using a combiner and a decision threshold to determine the output. Figure 3.10 shows the block diagram of an n output mANFIS using n ANFIS to determine the n outputs. Each individual ANFIS network is separately trained and the combiner cum

decision threshold will select the most likely output. Only one of the n outputs of the decision threshold will be '1'. Thus an n emotions identification system using mANFIS can be implemented as shown in Figure 3.10.

3.3.3 Generic Self-organizing Fuzzy Neural Network (GenSoFNN)

The Generic Self-organizing Fuzzy Neural Network (GenSoFNN) [Tung02a] employs an incremental and dynamic approach to the identification of knowledge based on fuzzy rules that defines its connectionist structure. It is developed to emulate the brain functionalities by extracting a set of high-level semantic If-Then fuzzy rules from the presented numeric training data [Oent08, Oent06]. The GenSoFNN structure is presented in Figure 3.11.

Consider vector $X = (x_1, \dots, x_i, \dots, x_{n1})^T$ representing the input, vector $Y = (y_1, \dots, y_i, \dots, y_{n1})^T$ representing the output and vector $D = (d_1, \dots, d_i, \dots, d_{n1})^T$ representing the desired network outputs required to the GenSoFNN. The variable i, j, k, l and m refer to arbitrary nodes and the output of a node is denoted as Z and its subscripts specify the origin. In Layer 1, each single input x_i is passed to the input node IV_i . It would be subsequently passed to Layer 2 where it contains the parameters of the input fuzzy set. The parameters are interpreted as the corner of the trapezoidal-shaped fuzzy sets; they are l and r representing left and right support points respectively and u and v representing left and right kernel points. Based on the parameter $IL_{i,l}, \dots, IL_{n1,jn1}$, rules are generated in Layer 3 and are denoted as $R_1, \dots, R_k, \dots, R_{n3}$. It would be followed by Layer 4 of the output term. These layer 2, 3 and 4 act as a hidden layer of GenSoFNN and are created and tuned as training commences. Subsequently, in Layer 5, the output node OV_m where $m \in \{1, \dots, n5\}$,

computes the single output denoted as y_m . GenSoFNN adopts the Mamdani's [Mamd77] fuzzy model and the k th fuzzy rule R_k has the form as defined in Equation (3.24).

$$R_k : \quad \text{IF } (x_1 \text{ is } IL_{(1,j)}) \text{ AND } \cdots (x_i \text{ is } IL_{(i,j)}) \text{ AND } \cdots (x_l \text{ is } IL_{(l,j)}) \\ \text{THEN } (y_1 \text{ is } OL_{(l,1)}) \text{ AND } \cdots (y_m \text{ is } OL_{(l,m)}) \text{ AND } \cdots (y_M \text{ is } OL_{(l,M)}) \quad (3.24)$$

where $IL_{(i,j)_k}$ is the j th fuzzy label of the i th input that is connected to R_k and $OL_{(l,m)}$ the l th fuzzy label of the m th output to which R_k is connected. The use of Mamdani fuzzy model provide an intuitive and easy to understand rule-base as compared to the Takagi-Sugeno-Kang (TSK) rule structure that have been adopted in ANFIS.

Figure 3.11: The GenSoFNN Architecture.

GenSoFNN is one of the classifiers that can provide consistent result due to its ability to do self-clustering process of the training data by employing the Discrete Incremental Clustering (DIC) [Tung02b] technique. In contrast to the ANFIS clustering approach that needs prior information, GenSoFNN requires no initial knowledge of the input data. The DIC ability of GenSoFNN can separate input clusters from noisy/spurious data that have poor correlation to the genuine data. DIC also allows dynamic clustering where the data points can move from one cluster to another, under varying conditions. These abilities made GenSoFNN robust even under very noisy environment.

3.4 Summary

The proposed speech emotion recognition system was introduced and discussed in details in this chapter. It was derived from the fundamental understanding of different approaches to cater to speech emotion issues and task as presented in Chapter 2. A selected review of similar speech emotion recognition system was presented in Taxonomic Table 2.3. Inspired by human auditory system with hypothesis that human can recognize emotion from speech, the proposed speech emotion recognition system comprises of three main modules was introduced. The first module is the speech emotion data collection and pre-processing. This module addressed the data used in constructing the proposed speech emotion recognition system. Detail description of the data collection has been explicitly presented in this chapter whereas the general data collection approaches had been discussed in Chapter 2. Three different cultural influence datasets; namely: Berlin, NTU_American and NTU_Asian will be employed to represent European, American and Asian culture respectively. This chapter continued with the discussion of two different features extraction methods based on the frequency domain

analysis of Mel Frequency Cepstral Coefficient (MFCC) and statistical time domain analysis of Gaussian Mixture Model (GMM). This second module of features extraction handles the important data transformation to map the correct emotion to be further processed using the selected classifiers. Subsequently, the classification module caters the training and testing of the pre-processed extracted features. Performance of different classifiers will be compared and analyzed to measure the accuracy of the proposed speech emotion recognition system ability. Based on the hypothesis that human can categorize different emotions, three different neural network and fuzzy neural network classifiers are applied. The Multi-layer perceptron (MLP) [Bish95, Rume86, Rose62], Adaptive Network-based Fuzzy Inference System (ANFIS) [Jang93] as well as Generic Self-organizing Fuzzy Neural Network (GenSoFNN) [Tung02a] are selected for the classification module. Hence, an emotion verification experiment will be conducted using the proposed speech emotion recognition system. Issues such as features selection and classifier model selection will be discussed in detail. The preliminary experimental results of emotion verification will be presented in Chapter 4. In addition, visualization approach will also be provided to enable one to see the differences as well as similarity of one culturally-influenced speech emotion data to the other.

CHAPTER 4

SPEECH EMOTION VERIFICATION: A PRELIMINARY WORK

The general architecture of the proposed speech emotion recognition system was introduced in the previous chapter and illustrated in Figure 3.1. It consists of three important modules; namely: data collection and pre-processing, feature extraction and classification. The development of such a proposed system was derived from the fundamental understanding outlined in Chapter 2. Several selected speech emotion recognition systems implemented by other researchers were also reviewed and summarized in Table 2.3. It is observed that there are many attempts to combine different feature extraction methods and classifiers to tackle speech emotion recognition tasks. However, the performance of a speech emotion recognition system does not solely rely on the correct combination of features and the classifiers. The selection of important features to represent the emotion information in speech and the optimal network model need to be considered for the best emotion recognition performance. In this chapter, a preliminary work on speech emotion verification is presented with emphasis on features and classifier model selection for the proposed speech emotion recognition system. The different decibel analyses based on the pre-processed speech emotion spectral signal will also be introduced to enable researchers to visualize the similarity and dissimilarity pattern of one culturally-influenced speech emotion data from another. Such visualization approach can facilitate better understanding of speech emotion in particular the cultural influence on speech emotion that is currently not offered by most speech emotion recognition system to date.

4.1 Feature Selection

The number of features generated from the selected feature extraction techniques can be large especially when one starts to combine multiple feature extraction approaches. The use of a large number of features in a classifier can result in lower classification accuracy and longer computation time. Such situation could be due to the contradictory responses of different feature combination or repeatedly overlapping information. Hence, such situation can be addressed by incorporating the use of discriminatory or relevant features.

Feature selection refers to the technique of selecting a subset of relevant features by eliminating irrelevant and redundant features with minimal correlation in producing the output for building a robust learning model. By doing so, it can significantly improve the prediction performance of the system by minimizing the complexity of the task to be handled by the classifier, speeding up the learning process, improving the model interpretability as well as promoting superior data understanding [Somo10, Guyo03]. In this section, a preliminary verification experiment for NTU_American dataset is conducted for both Mel Frequency Cepstral Coefficient (MFCC) [Slan98] and Gaussian Mixture Model (GMM) [Rip196] features.

4.1.1 MFCC Feature Selection using Automatic Relevance Determination (ARD)

Somol et al. [Somo10] mentioned that the Best Individual Features (BIF) is the simplest approach for feature selection. A score of each feature is computed based on its relevancy and subsequently a feature ranking is generated. Subsets are then selected according to the scores of the best individual features. Such approach yielded the fastest result compared to the other feature selection methods. Moreover, the feature ranking method is selected in this study as a principal selection mechanism due to its simplicity, scalability

and ability to yield good empirical results [Guyo03]. Hence, the problem of selecting the most relevant features of the Mel Frequency Cepstral Coefficient (MFCC) is focused by applying a feature ranking method based on the Automatic Relevance Determination (ARD) [MacK94, MacK92] as a comparison to the standard Slaney's MFCC arrangement [Slan98]. Although more advanced methods such as Monte Carlo Evaluative Selection [Quah07], Separability-Correlation Measure [Fu03] and Mutual Information based Feature Selection [Batt94] are likely to produce better result, the ARD is selected due to its simple Bayesian structure learning.

The ARD is introduced by MacKay [MacK94] and later extended by Neal [Neal98, Neal96]. The objective of ARD implementation is to discover the relevancy of the features in the dynamics of the system. It can be implemented as a form of Bayesian structure learning where each feature has its own prior variance parameter based on a Gaussian distribution weight. Small variance value suggests that the particular feature have little influence on the subsequent value. On the contrary, a large variance value indicates that the particular feature is important in describing the input data. Hence, the ARD in this work is used to generate MFCC feature ranking based on its relevancy on the speech emotion verification. Selective data elimination is then implemented to study the importance of the selected features to the accuracy of the proposed speech emotion recognition system.

Based on [Qi04], consider a linear classifier that classifies a point x according to $t = \text{sign}(w^T x)$ for some parameter vector w . The two classes point to 0 and 1 value consecutively. Given a training set $D = \{(x_1, t_1)(x_2, t_2) \dots (x_N, t_N)\}$, the likelihood for w can be written as Equation (4.1).

$$p(t | w, X) = \prod_i p(t_i | x_i, w) = \prod_i \Psi(t_i, w^T \phi(x_i)) \quad (4.1)$$

where $t = \{t_i\}_{i=1}^N$, $X = \{x_i\}_{i=1}^N$ and $\Psi(\cdot)$ is the cumulative distribution function for a Gaussian. Step or logistic function for $\Psi(\cdot)$ can also be implemented. The basis function $\phi^T(x_i)$ allows the classification boundary to be nonlinear in the original feature space. This is the same likelihood used in the logistic regression and Gaussian-processed classifiers. Given a new input, x_{N+1} the approximation of the predictive distribution is as given by Equations (4.2) and (4.3).

$$p(t_{N+1} | x_{N+1}, t) = \int p(t_{N+1} | x_{N+1}, w) p(w | t) dw \quad (4.2)$$

$$\approx p(t_{N+1} | x_{N+1}, \langle w \rangle) \quad (4.3)$$

where $\langle w \rangle$ denotes the posterior mean of the weights and is called the Bayes Point [Herb01]. The basic idea in ARD is to give the feature weights independent Gaussian priors to Equation (4.4).

$$p(w | \alpha) = \prod_i N(w_i | 0, \alpha_i^{-1}) \quad (4.4)$$

where $\alpha = \{\alpha_i\}$ is a hyper parameter vector that controls how far away each weight is allowed to go from zero. The hyper parameters α are trained from the data by maximizing the Bayesian evidence $p(t | \alpha)$. The training can be achieved using a fixed-point algorithm or an Expectation Maximization (EM) algorithm by treating w as a hidden variable [MacK92]. The outcome of this optimization is that many elements of α go to infinity in such a way that w would have only non-zero weight w_j . This naturally prunes irrelevant features in the data. Hence, the ARD rank is selected as a yardstick to

position the relevancy of the individual MFCC features in comparing the performance of the standard MFCC arrangement derived from Slaney's approach [Slan98].

NTU_American Verification Based on Different Individual MFCC Feature Arrangements

In this preliminary work, only the *anger* emotion will be used. Anger verification refers to the ability of the proposed system to determine whether the testing instance is an anger emotion or otherwise. *Anger* from the NTU_American dataset is selected as the target class as opposed to its complementary class (denoted as *Not_Anger*). The *Not_Anger* class comprised of *sadness*, *happiness* and *neutral* instances. Subsequently, the different individual MFCC arrangements were then designed as an attempt to investigate the effect of the individual MFCC sequence to the performance of the proposed speech emotion recognition system. Two different classifiers of WEKA-based Multi Layer Perceptron (WEKA-MLP) [Witt99] and Adaptive Network-based Fuzzy Inference System (ANFIS) [Jang93] were adopted to categorize both classes. The WEKA-MLP was employed based on its ability to self-tune hence eliminating the need for classifier model network selection. On the contrary, the ANFIS model in this experiment used the fixed radii value of 0.78. Such a radii value was derived from the initial preliminary experimental result. The block diagram in Figure 4.1 depicts the general architecture of the experiment conducted using different individual MFCC feature arrangements for NTU_American anger verification.

The process of individual MFCC feature arrangement starts with the transformation of the pre-processed NTU_American speech emotion dataset into the 40

coefficient MFCC features. 1000 anger instances were randomly selected from the NTU_American anger dataset to represent the *Anger* class whereas the *Not-Anger* class was represented by a random selection of 333 instances of sadness, 333 instances of happiness and 334 instances of neutral. Then, the selected instances of *Not-Anger* class were randomly arranged to normalize the different emotion sequence. The *Standard-MFCC* feature arrangement refers to the exact individual MFCC coefficient sequence output derived from the Slaney's MFCC transformation approach [Slan98].

Figure 4.1: Different Individual MFCC Feature Arrangements for NTU_American Anger Verification.

Conversely, the *ARD-MFCC* feature arrangement was generated from the restructuring features sequence based on the feature ranking produced by ARD. The features were ranked from the most impactful to the least impactful based on the weight assigned to each features. The feature ranking for different emotion of anger, happiness

and sadness as opposed to neutral are presented in Appendix AII. From the result in Appendix AII, it is observed that the verification result for anger versus neutral and happiness versus neutral yielded similar individual features position in the feature ranking. However, the individual feature position for sadness versus neutral feature ranking differs slightly. Hence, the sequence of individual MFCC features of the *ARD-MFCC* features in this experiment is based on the anger versus neutral ranking presented in Appendix AII. It is also noted that although there have been many works on feature selection; the work in this thesis is restricted to the usage of ARD ranking for the 40 MFCC features.

Once the data for both *Standard-MFCC* and *ARD-MFCC* sequences were accordingly arranged, they were grouped into different number of features; namely: 10, 20, 30 and 40 features. Such an approach is designed to facilitate the understanding of the optimal number of MFCC features to be implemented in the proposed speech emotion recognition system. Subsequently, a five-fold cross-validation technique adopted for global accuracy. This is to ensure that the classifier do not see the testing data prior to the training exercise. 80% of the instances (800 instances from *Anger* class and 800 instances from *Not-Anger* class) were used for training while the remaining 20% (100 instances from *Anger* class and 100 instances from *Not-Anger* class) were used for testing. The training-testing pairs are used until all pair has been trained and tested. The equal number of training data from both *Anger* and *Not-Anger* class were used in order to eliminate biasness towards the class with a larger number of instances.

Figure 4.2 shows the experimental result for the anger verification of NTU_American dataset for different individual MFCC arrangement. Figure 4.2(a) illustrates the overall anger verification accuracy yielded using WEKA-MLP classifier

and Figure 4.2(b) depicts the ANFIS classifier performance. From the observation of both Figure 4.2(a) and 4.2(b), the *Standard-MFCC* feature arrangement consistently performed better than the *ARD-MFCC* feature arrangement for both classifiers. Such result is contradictory to the prior expectation that the *ARD-MFCC* feature arrangement would perform better due to the fact that the features are ranked based on the individual features degree of impactfulness. Such result is obtained because the sequence of *Standard-MFCC* arrangement manages to keep the integral information of emotion derived from the low-frequency to high-frequency signal layout. However, similar information cannot be retained in the *ARD-MFCC* arrangement as the individual MFCC sequence is no more based from the frequency order but on the Bayesian weight derived from the ARD ranking. In Figure 4.2(b), the maximum disparity between these two feature arrangements is clearly noticed in the 40 features experiment using ANFIS classifier with 21.2% variance. Such scenario happens because ANFIS relies heavily on feature arrangement to form the necessary cluster for rules generation. Since the vital information of emotion is lost from the re-arrangement of the individual MFCC features, the *ARD-MFCC* sequence accuracy suffers badly. Hence, the *Standard-MFCC* arrangement is selected for emotion recognition task in this work.

Nonetheless, it is also noted that the different number of features used in the anger verification from the empirical result in Figure 4.2(a) and 4.2(b) does not strongly affect the performance of the proposed speech emotion recognition system. The *Standard-MFCC* performance for 40-features of both WEKA-MLP and ANFIS classifiers reach the highest accuracy with 92.9% and 93.7% respectively. However, the decrease of the accuracy is observed to be directly proportional with the reduction of the number of features adopted. The differences in values ranging from 4.5% to 7.6% for WEKA-MLP and 3.1% to 6.1% for ANFIS are recorded in comparing the performance of different

number of features used. Thus, 40 features *Standard-MFCC* arrangement is selected for further speech emotion analysis since 40 features is still considered as a reasonable number of features to be handled and managed in order to yield reasonable accuracy.

(a) WEKA-MLP

(b) ANFIS

Figure 4.2: ARD-MFCC and Standard-MFCC Arrangement Verification Performances

4.1.2 GMM Feature Selection

The window size and the number of Gaussian fit are the two important parameters that need to be tuned in order to obtain optimum performance for speech emotion recognition.

Hence, preliminary experiment is conducted by employing the WEKA-MLP [Witt99]

classifier to determine the optimum value for both parameters. Although more advanced techniques for optimization are available; such as: Genetic Algorithm (GA) [Holl92] and Differential Evolution (DE) [Stor97], they do not really suit our need as they focus on high-dimensional data and can be time consuming.

Determination of the Optimum Window Size.

The rationale of this experiment is to determine the optimal window size for the NTU_American anger verification task. Three different values of window size of 1024, 2048 and 4096 were selected. Based on the preliminary understanding, 80% overlapping was implemented to ensure that there were no abrupt changes occurring between windows. Table 4.1 shows the different accuracy recorded based on the three different window sizes using WEKA-MLP classifier. It is clearly observed that window size parameter has minimal impact on the accuracy with performance variance of $[-1: +1]$ V from one window size to another. Thus, the 2048 window size with 80% overlap is selected as it offers sufficient number of instances for classification and able to provide comparable accuracy.

Table 4.1: Performance Accuracy for Different Window Size for GMM Features

Extraction.

Window size	Accuracy (%)
1024	65.1
2048	65.0
4096	64.0

Determination of the Optimum Number of GMM Fit.

GMM is based on the assumption that the data to be clustered are drawn from one of several Gaussian distributions. It has been reported that Gaussian mixture distribution could approximate any distribution up to arbitrary accuracy as long as a sufficient number of components are used [Rip196]. It is important to estimate the correct number of mixtures as it affects the overall performance. Too few will cause the under-fitting problem while too large will introduced over-fitting. A common approach for estimating the optimal number of mixtures is by adopting the Expectation Maximization (EM) algorithm [Demp77]. The EM algorithm is used to determine the maximum likelihood estimation of Probability Density Function (PDF).

The NTU_American speech emotion data was selected for anger verification using different number of Gaussian fit. The pre-processed speech emotion data were framed using 2048 window size with 80% overlap based on prior understanding of optimal window size experiment described in the previous section. From the individual windowed speech emotion data, the histogram of the probability function was generated. Subsequently, a different number of GMM fit estimations based on the respective histogram are derived to observe the optimal value that is suitable to represent the anger characteristic for NTU_American anger verification experiment. Different features were extracted from individual GMM estimation; namely: mean vectors, mixture weight and covariance matrix. Detailed explanation was provided in Section 3.2.2.

Figure 4.3 shows the empirical result of different Gaussian fit with different accuracy in the anger verification task using the NTU_American dataset. It is observed five Gaussian components are sufficient to provide a reasonable accuracy. Hence, 5 gaussian components per emotional speech window with 15 features respectively (5 mean

vectors, 5 mixture weights and 5 covariance matrixes) are selected to represent the GMM features for further analysis of the speech emotion recognition performance.

Figure 4.3: WEKA-MLP NTU_American Anger Verification Result

4.2 Speech Emotion Verification

As an extension to the anger verification preliminary experiment discussed in Section 4.1, comprehensive speech emotion verifications are conducted namely; anger, sadness and happiness verification. The verification in this thesis refers to a task of comparing the mutually exclusive membership of the input instance by which an instance either belongs to the target class (usually denoted as 1) or its complement (usually denoted as 0). The objectives of the speech emotion verification in this study are: 1) to study the ability of the proposed system to differentiate emotion, 2) to select a classifier network model that can yield optimum accuracy and 3) to observe the specific emotion verification performance. Two different types of features extraction methods based on the Mel Frequency Cepstral Coefficient (MFCC) [Slan98] and the Gaussian Mixture Model (GMM) [Figu02] with configuration described in Section 4.1 were coupled with neural network of Multi Layer Perceptron (MLP) and fuzzy neural network of Adaptive Network-based Fuzzy Inference System (ANFIS) and Generic Self-organizing Fuzzy

Neural Network (GenSoFNN) classifiers. Such approaches were implemented to verify three discrete emotions; namely: anger, happiness and sadness in NTU_American dataset.

4.2.1 Experimental Set-up for Speech Emotion Verification

Based on the preliminary experiments described in Section 4.1, 40 features were extracted using Slaney's MFCC approach [Slan98] and subsequently 15 features were extracted using the GMM approach [Figu02] for each emotion from NTU_American; namely: anger, happiness, sadness with the addition of neutral serving as the emotionless state. The instances value were normalized between $[-1: +1]$ and labeled according to the emotion represented. 1000 random instances were selected to represent respective emotions using MFCC features whereas 300 random instances were selected to represent GMM features. The disparity of sample size between MFCC and GMM features was due to the different optimal window size. The MFCC features using Slaney's approach was based on *Window Size* = 256 whereas the GMM features was based on *Window Size* = 2048 as described in preliminary experiment in Section 4.1. In order to build the complement class for the particular emotion, similar number of instances consists of other emotions were selected to ensure that the classifiers do not have bias towards class with larger number of instances. For instance, the anger complementary class (*Not_Anger*) consists of one third of *happiness*, *sadness* and *neutral* instances respectively. Such data organization setting is also followed by sadness and happiness verification experiment with appropriate modification according to the emotion to be verified. The arrangement of the different emotions in the complement class was also randomized to ensure that the emotion arrangement of the dataset does not influence the classifiers ability to verify emotion. Figure 4.4 depicts the organization for emotion verification training and testing

class data. The values used in the diagram represent the 1000 randomly selected instances of MFCC features for *happiness*, *sadness* and *neutral* emotion. For the GMM features, 300 instances are selected for each emotion and its complementary class

Figure 4.4: The MFCC Verification Training and Testing Class Data Organization

Once the emotion and its complement classes were ready, the classifiers were employed to verify the respective emotion. The 5-fold cross-validation technique was adopted to determine the global accuracy. 80% of the target and its complementary class were combined and used for training (1600 and 480 instances for MFCC and GMM features respectively) and the remaining 20% data were used for testing (400 and 120

instances for MFCC and GMM features respectively). Each training-testing dataset was iteratively used until all pairs were subjected to the verification experiment. The result from each training-testing pair was combined to construct a complete confusion matrix for each classifier. The performance of the combination of selected features extraction method and classifier can be compared for further analysis.

For this speech emotion verification task, a hardware specification of Intel® Core™2 Duo Processor E6750 with clock speed of 2.66 GHz and 3 GB RAM was used. Subsequently, a software specification of an Operating System of Microsoft Windows XP Professional Version 2002, Service Pack 2 coupled with MATLAB application Version 7.5.0.342 (R2007b) was applied for training and testing.

4.2.2 Classifiers Network Model Selection

An optimal performance outcome can be achieved using a tuned classifier network model. Since each classifier had its own individual parameters, the difficulties of tuning the classifiers are well known. Hence in this section, a selective brute-force problem solving technique was applied where a series of possible values for the important parameters were implemented to study the different accuracy obtained. Such technique was selected based on the simplicity of its implementation. Although it is exhaustive and time consuming, the technique is able to reflect the actual data characteristic.

The MLP network model selection

The hidden layer in MLP architecture is very important because it handles the non-linearity computation [Bish95]. Hence, it is sensible to tune an MLP for the optimal number of the hidden layers and the number of the neurons needed in each hidden layer

respectively to ensure the MLP convergence. The MLP from WEKA toolset [Witt99] adopted in the previous experiments described in Section 4.1 is not implemented in this experiment. This is because the WEKA-MLP self-adjusts the hidden layer parameters automatically which restricts the manual tuning process to be conducted. Hence, a MATLAB-based MLP was employed because MATLAB offers flexibility for hidden layer parameters modification.

A rough estimation of optimal number of hidden layers as well as the number of neurons in the respective hidden layer was computed using local validation technique. Local validation technique refers to the implementation of similar dataset for training and testing. It is employed because of its ability to give indication of the selected classifier network model capability to memorize the input patterns. The accuracy of above 90% can be considered as a good indicator for good verification result.

Several experimental setups are implemented for consistency. The performance goal of Mean Squared Error (MSE) is set to 0.01 with a maximum training epoch of 1000 cycles. The learning rate, which determines how much of the weight can be adjusted, is set to 0.001. It is important to properly set the correct value for the learning rate because it controls the speed of the network error convergence. If the selected learning rate value is too small, the algorithm will take a long time to converge and conversely, if it is too large (maximum value is 1), the classifier may converge to a local minimum. A local minimum is a point at which the network stabilizes on a solution which may not be the most optimal global solution. The number of MLP input neurons is equal to the number of features selected; namely: 40 input neurons for MFCC features and 15 input neurons for GMM features. As for output layer, only one neuron is needed as the output is in the form of regression value ranging from 1 to 0. Threshold value of 0.5 is used to determine the

target emotion class or otherwise. The *Logarithmic Sigmoid (logsig)* function is used as the activation function in the hidden layers and the *Pure Linear (purelin)* activation function is employed in the output neuron. The rationale of the use of *logsig activation function* is to cater for unipolar binary inputs [0, 1] so that the outputs represent posterior probabilities estimates of the non-linear elements in the hidden layer. In addition, the *purelin function* is used to aggregate the output from hidden layers with zero mean and unit variance in the output layer.

Table 4.2: MFCC-MLP Model Selection Local Validation Training Time

Emotion	1 hidden layer (sec)						2 hidden layer (sec)		
	10	15	20	25	30	35	10	15	20
Anger	247.1	1056.3	229.4	87.4	85.9	108.6	7.8	14.1	28.3
Happiness	245.6	586.3	859.2	178.4	71.1	156.6	9.8	39.7	67.5
Sadness	270.9	323.8	338.4	59.8	42.9	78.2	5.8	15.9	37.2

Based on the experimental setup described, different MLP network models with various numbers of hidden layers and neurons values are tested using the MFCC features derived from the NTU_American dataset. Table 4.2 shows the different training time needed for the three different emotions verification convergence; namely: anger, happiness and sadness. From the observation of the result in Table 4.2, the MLP network model with 10 neurons in 2 hidden layers recorded the fastest computation time compared to the other MLP network model for all three different emotions. It is also observable that the two hidden layers MLP regardless of the number of neurons performs better than one hidden layer MLP.

Subsequently, similar experiment is also conducted using the GMM features. Figure 4.5 depicts the timing performance of different MLP network model for anger verification using NTU_American dataset. It is observed that 10 neurons MLP regardless of the numbers of hidden layer are able to compute the fastest convergence. The convergence time is longer when more than one hidden layer MLP is used. Hence, for the purpose of emotion verification, the MLP network model using 10 neurons in 2 hidden layers is selected as the results are consistent for both MFCC and GMM features. Although such approach may not guarantee to be the most accurate, it has empirically demonstrated the MLP high memorization ability within the fastest convergence time.

Figure 4.5: GMM-MLP Model Selection Local Validation Training Time based on NTU_American Anger Verification.

The ANFIS network model selection

ANFIS relies heavily on the input data. The fuzzification stage will cluster the data as an initial Fuzzy Inference System (FIS) [Jang93]. The number of cluster relates

directly to the rule generation which influences the overall ANFIS performance. Since no prior knowledge is available for the emotion data clustering, subtractive clustering technique is employed coupled with a brute-force algorithm to determine the optimum radii value. The subtractive clustering method is able to produce scattering partitions based on the density of the dataset and able to give sufficient input-output mapping precision. In the ANFIS network model selection experiment, the Root Mean Square Error (RMSE) is set to 0.1 and the step size is fixed at 0.5. Radii values ranging from 0.77 to 1.70 are implemented for both MFCC and GMM features to observe the optimum radii value for each fold in a 5-fold validation approach of the NTU_American emotion verification.

Figure 4.6 depicts the accuracy range obtained using the different radii value for NTU_American emotion verification using MFCC features; namely represented by: Figure 4.6(a) for anger, Figure 4.6(b) for happiness and Figure 4.6(c) for sadness emotion respectively. It can be observed that radii values ranging from 0.77 to 1.70 are suitable for selective brute-force algorithm as it gives the highest peak of the accuracy for each emotion in the NTU_American emotion verification. In addition, Figure 4.7 illustrated the different accuracy yielded from the usage of different radii value for NTU_American emotion verification using MFCC features for each fold per emotion.

Figure 4.6: Different Accuracy Obtained by Different Radii Value for different NTU_American emotion verification using MFCC features

b) NTU_American Sadness Verification

c) NTU_American Happiness Verification

Figure 4.7: Different Accuracy Obtained By Different Radii Value for NTU_American

Emotion Verification Using MFCC Features

From the Figure 4.7, it is observed that the different radii values even for similar validation fold data, yielded different accuracy. Each fold has its own unique optimal radii value hence different radii value was acquired for different training and testing data-pair to gain best accuracy. The best radii value for each fold for anger verification are; 1.07 for Fold 1, 1.19 for Fold 2, 1.31 for Fold 3, 0.92 for Fold 4 and 1.25 for Fold 5 as illustrated in Figure 4.7(a). The pattern however is not similar to sadness and happiness verification. Based on Figure 4.7(b), the best radii value for each fold of sadness verification is near to each other; namely: 1.19 for Fold 1, 1.15 for Fold 2, 1.22 for Fold 3, 1.23 for Fold 4 and 1.37 for Fold 5. Happiness verification result in Figure 4.7(c), shows a large distribution of the best radii value from one fold to another; 0.96 for Fold 1, 1.40 for Fold 2, 0.84 for Fold 3, 1.34 for Fold 4 and 1.57 for Fold 5. Hence, this empirical result demonstrates that the implementation of the selected brute-force approach is able to provide better understanding of the radii value needed for respective emotion verification.

The GenSoFNN network model selection

GenSoFNN is a neural fuzzy network [Tung02a] that is able to self-tune its parameter using the Discrete Incremental Clustering (DIC) technique [Tung02b]. Hence, the network model automatically generated by the GenSoFNN is adopted. There are two configuration used for this classifier:

- Generalization Rules: the slope of the cluster is set to be gentle so that it could have a wider range of membership, and
- Specialization Rules: the clusters have a steeper slopes.

The first configuration allows the GenSoFNN network to be able to recognize the data supplied during the testing even though the data supplied had never been trained by the network. In addition, the subsequent setting is focused on memory recall phase where the data supplied for testing had been used for training. The Mean Square Error (MSE) is set to 0.01 and the maximum epoch is limited to 20 cycles. This value is selected based on the observation during the experiment that after 20th cycle, the total error is reduced with a small margin (less than 0.01).

4.2.3 Speech Emotion Verification Result and Discussion

Based on the selected network classifier information gathered through the empirical experiments described in Section 4.2.2, the speech emotion verification experiments were conducted. The 5-fold cross-validation technique was adopted to allow the classifier to train and test with different dataset for global validation. Global validation or better known as generalization refers to the ability of the classifier to make prediction on unknown data. In contrast to the prior experiments of using similar data for training and testing, this k-fold cross-validation technique can provide the actual performance of the emotion verification system.

Figure 4.8 shows the speech emotion verification result using GMM and MFCC features respectively with mean squared error threshold value set to 0.01 for all three classifiers of MLP [Bish95], ANFIS [Jang93] and GenSoFNN [Tung02a]. Subsequently, Table 4.3 and 4.4 show the training and testing time needed for the speech emotion verification experiments.

Figure 4.8: Emotion Verification Result Using Different Features Extraction.

It is observed from Figure 4.8(a) that GMM features are able to produce the accuracy of emotion verification ranging from 64.2% (recorded by happiness verification using GenSoFNN) to 79.1% (recorded by sadness verification using ANFIS) with mean accuracy of 72.2%. However, based on accuracy result in Figure 4.8(b), MFCC verification result consistently exceed the GMM accuracy by at least 13.3% with minimal accuracy of 77.5% yielded by happiness verification using MLP and maximum accuracy of 94.7% achieved by anger verification using GenSoFNN. The mean accuracy of MFCC verification is recorded at 88.6%. Happiness consistently yielded the worst emotion verification performance compared to anger and sadness for both GMM and MFCC features with mean accuracy of 70.8% and 84.1% for GMM and MFCC features respectively. On the contrary, sadness recorded highest accuracy in emotion verification result; see Figure 4.8(a), using the GMM features with mean accuracy of 76.2%. Anger yielded the highest performance with mean accuracy of 92.0% by using MFCC features as depicts in Figure 4.8(b). The three classifiers yielded mixed performance hence no classifier can claim superiority over the others. ANFIS obtained slightly better

performance with mean accuracy of 83.0% followed by 80.8% and 77.5% of GenSoFNN and MLP respectively.

Table 4.3: Convergence Computation Time Performance using MFCC Features.

Emotion	Training time (sec)			Testing time (sec)		
	MLP	ANFIS	GenSoFNN	MLP	ANFIS	GenSoFNN
Anger	6.54	17.16	160.45	0.03	0.03	2.20
Happiness	11.00	11.87	167.53	0.04	0.03	2.44
Sadness	9.14	12.78	168.07	0.05	0.03	2.30

Table 4.4: Convergence Computation Time Performance using GMM Features.

Emotion	Training time (sec)			Testing time (sec)		
	MLP	ANFIS	GenSoFNN	MLP	ANFIS	GenSoFNN
Anger	19.14	14.71	6.67	0.03	0.02	0.15
Happiness	11.99	17.56	6.70	0.04	0.02	0.10
Sadness	18.58	15.52	7.00	0.05	0.02	0.16

The training time shown in Tables 4.3 and 4.4 refer to the duration of the computation needed by the classifier to converge from learning. Moreover, the testing time refers to the duration the classifier needs to match the instance with the correct outcome. The fastest training and testing time for the classifiers in the Table 4.3 and 4.4 are highlighted with bold writing. It is interesting to observe that GenSoFNN needs at least ten times longer than either MLP or ANFIS to train the MFCC features. On the contrary, GenSoFNN is the fastest classifier to converge using the GMM features. This is because GenSoFNN needs to generate more rules for MFCC features (1000 instances X 40 features) compared to the GMM features (300 instances X 15 features). The

consequence of the large number of rules generated for MFCC features results in the longer testing time needed. GenSoFNN needs at least 2 second to test the data as opposed to less than 0.05 second needed by the other classifiers. However, since training is only required once, GenSoFNN is still relevant for real-time emotion verification system provided the application allows ample training time prior to the practical deployment of the system. MLP and ANFIS yielded reasonably consistent computation time regardless of which the features are used.

From the experimental results in Figure 4.8 and Tables 4.3 and 4.4, it can be summarized that both MFCC and GMM features coupled with the three different classifier; namely: MLP, ANFIS and GenSoFNN demonstrated a huge potential for emotion verification. The accuracy of at least 64.2% is recorded within reasonable computation time of less than 3 minutes for training and less than 3 seconds for testing. In all cases, the MFCC features yielded better performance in comparison to the GMM features with more than 10.0% variance. This result provides indication of the practicality of the emotion verification system using MFCC features coupled with either neural network (MLP) or fuzzy neural network (ANFIS and GenSoFNN). Based on these findings, extensive study will be conducted using MFCC features and will be subsequently presented in Chapter 5 and 6.

4.2.4 Culture Influence in Speech Emotion Verification

So far, in the previous experiments, only the NTU_American dataset has been used for speech emotion verification. The empirical results recorded a reasonable accuracy hence shown the potential of the proposed approach. However, culture is hypothesized as one of the factor that can influence the way emotion is being conveyed and perceived. Detailed

discussion of this hypothesis is provided in Section 2.4. Hence, it is valid to extend such a hypothesis to understand the effect of culture influences in speech emotion recognition system.

Many findings reported that human from similar culture are able to recognize emotion with better accuracy when compared to those from a different culture. However, human also have the capability to recognize emotion correctly across culture. Thus, it is important to verify the proposed speech emotion verification system performance for NTU_American dataset by comparing the system performance using a standard and well-known Berlin dataset. As a preliminary study, three different emotions; namely: anger, happiness and sadness from both NTU_American and Berlin dataset were trained and tested for emotion verification task. The MFCC feature extraction coupled with ANFIS classifier were employed. The objective of such a study is to analyze the disparity of the two different datasets performance although similar set of emotions are recognized. In addition, the robustness of the proposed system can also be observed irrespective of the different cultural-influenced data employment.

Similar experimental setups for the data organization as described in section 4.2.1 are applied. In this experiment, only MFCC features coupled with ANFIS classifier were employed based on its ability to yield optimum accuracy. Generalization using the 5-fold cross-validation approach was conducted in order to obtain global accuracy of the proposed system. Table 4.5 shows the speech emotion verification result for both dataset of employing two different classifier networks respectively.

It is observed from the result in Table 4.5 that NTU_American dataset yielded a mean accuracy of almost 90% while Berlin dataset obtained mean accuracy of 80.9%. All of the emotion in NTU_American can be recognized with at least 85.3% and minimal

verification result of 76.4% can be achieved using Berlin dataset. It is interesting to note that happiness is consistently recorded the lowest accuracy compared to anger and sadness with accuracy of 85.3% and 76.4% for NTU_American and Berlin datasets respectively. In addition, anger in NTU_American yielded the highest accuracy with difference of 3.8% and 7.7% compared to sadness and happiness respectively. However, the highest emotion accuracy for Berlin dataset is recorded by sadness. Sadness yielded better accuracy than anger and happiness with variance of 6.6% and 10.1%.

Table 4.5: Speech Emotion Verification Result for NTU_American and Berlin Dataset

Emotion	NTU_American dataset (%)	Berlin Dataset (%)
Anger	93.0	79.9
Happiness	85.3	76.4
Sadness	89.2	86.5
Mean	89.2	80.9

Based on the ANFIS performance for speech emotion verification system using NTU_American and Berlin datasets, it can be summarized that the proposed system has a potential in recognizing the emotion in speech with more than 75% minimum accuracy recorded regardless of the culturally-influenced dataset used. The NTU_American dataset result indicates that the data was comparable to the Berlin dataset which is considered a standard dataset for speech emotion recognition task. It is also noted that there are disparity on the result of similar emotion using different culturally-influenced datasets. More analysis on the cultural influence is needed to facilitate the understanding on the impact of such parameter in the speech emotion. Hence, detailed discussion on the effect of culture influence will be further discussed in Chapter 5.

4.3 Summary and Conclusion

A preliminary work on speech emotion verification was described in this chapter. Experimental results show potential of the proposed techniques to analyze speech emotion. Chapter 4 can be segregated into two main sub-sections; namely: feature selection and classifier model selection. The information and findings gathered from this preliminary findings provides a clearer research direction that is needed to be undertaken in speech emotion understanding and analysis.

Two types of feature arrangements are used for the MFCC feature selection. The *Standard-MFCC* feature arrangement derived from Slaney's approach [Slan98] and an *ARD-MFCC* feature arrangement are adopted to test the best features for speech emotion analysis. It is found that the 40 *Standard-MFCC* feature arrangement performed consistently better in comparison to the *ARD-MFCC* feature arrangement. In addition, variation of window size and the optimal number of GMM fit were also derived for the GMM features. From the result, MFCC performs better than GMM hence the 40 MFCC has been selected and will be used through out this thesis for speech emotion recognition. The best statistical-based GMM features can be yielded using window size of 2048 and 5 GMM fit.

Prior to speech emotion verification experiment, the selection of optimal classifier network model experiment was conducted. An MLP with 10 neurons in 2 hidden layer was selected for optimal MLP network and the ANFIS with radii value ranging from 0.7 to 1.7 was used for optimal ANFIS network. In addition, GenSoFNN utilizes the DIC approach for self-optimizing model. The experimental result shows a potential of using both MFCC and GMM features coupled with the three classifiers of MLP, ANFIS and GenSoFNN. Minimum accuracy of 64.2% with reasonable computation time of less than

3 minutes for training and less than 0.05 seconds for testing can be achieved based on the proposed method. However, MFCC features are able to yield accuracy of at least 10.0% better than the GMM features. Hence, MFCC features will be used in further analysis.

In conclusion, speech emotion can be verified using MFCC features coupled with the three different classifiers of MLP, ANFIS and GenSoFNN. Based on the finding from speech emotion profiling, there are certain similar characteristics of the speech emotion shared across cultures and at the same time has its own unique characteristics, which enable different emotion to be recognized within the same dataset. Hence it is envisaged that the findings obtained would help to boost the understanding of speech emotion and enhance machine speech emotion recognition capabilities. The next chapter will introduce the proposed speech emotion recognition system and further discuss the influence of culture to speech emotion recognition system and the importance in considering and detecting culture in such a system.

CHAPTER 5

CULTURAL DEPENDENCY ANALYSIS

FOR UNDERSTANDING SPEECH

EMOTION

Understanding the interlocutors' emotion and recognizing the listeners' perception is the key to boost communication effectiveness and interaction. Nonetheless, the perceived emotion is subjective and very much dependent on culture, environment and the pre-emotional state of the listener. Attempts have been made to understand the influence of culture in speech emotion and researchers have reported mixed findings that lead us to believe that there are some common acoustical characteristics that enable similar emotion to be discerned universally across different cultures. Yet there are unique speech attributes that facilitate exclusive emotion recognition within a particular culture. Some of the works were discussed in Chapter 2. Thus, understanding cultural dependency is important to the performance of the speech emotion recognition system.

In this chapter, the speech emotion recognition problem is addressed by emphasizing both inter-cultural and intra-cultural assessments. Three different speech emotion databases; namely: Berlin Emo-db [Burk05], NTU_American and NTU_Asian dataset are selected to represent three different cultures of European, American and Asian respectively. These are to provide cultural variation in analyzing and recognizing the emotion of anger, happiness and sadness with neutral acting as a reference for emotionless state. These emotions are selected based on the psychologists' views [Cowi03] that are commonly accepted by many researchers and they can be assigned in three different quadrant of the affective space model [Lang95, Schl54]. Furthermore, based on Ververidis and Kotropoulos review of 32 emotional speech databases [Verv03],

the top most emotions recorded and analyzed by the researchers in the speech emotion field are the emotions: anger, sadness and happiness. Detailed description on the datasets employed was provided in Chapter 3.

Different data arrangements in accordance to varying degree of cultural dependencies are designed for the experimental setup to provide better understanding of inter-cultural and intra-cultural effect in recognizing the speech emotion. Features are extracted using Mel Frequency Cepstral Co-efficient (MFCC) method and classified with neural network (Multi Layer Perceptron (MLP)) and fuzzy neural networks; namely: Adaptive Network Fuzzy Inference System (ANFIS) and Generic Self-organizing Fuzzy Neural Network (GenSoFNN) representing precise and linguistic fuzzy rule conjuncts respectively. Based on the assumption that culture does influence the conveyence and perception of the emotion, a two-stage culture-sensitive speech emotion recognition approach for culture detection and subsequently followed by speech emotion identification is proposed. It is hypothesize that if the culture of a particular speech emotion instance is known, emotion can be identified based on the network trained using similar culturally-influenced speech emotion data. Such an approach can boost the global accuracy of the speech emotion recognition system due to the intra-cultural influence/bias.

5.1 Experimental Set-up for Intra-cultural and Inter-cultural Assessment

The datasets employed in the experiment as well as the overall architecture of the proposed speech emotion recognition system was described in Chapter 3. In this work, a discrete-class classification system approach is implemented. The classification system in this thesis refers to mapping of the testing data class with minimal error based on example

provided by the training data (supervised learning). It handles emotions as discrete classes that have mutual exclusive membership for a class. For example, if instance x is classified as anger, it negates the other classes membership. The proposed classifiers computed the weight of each feature and produced the result with minimal Mean Square Error (MSE) value specified during training.

Figure 5.1: Block Diagram of Discrete-Class Classification System Approach

Figure 5.1 shows the block diagram of the discrete-class classification system approach. The input dataset are construed by the 40 Mel Frequency Cepstral Coefficient (MFCC) features of four emotions; namely: anger, happiness, sadness and neutral. The classifier with four networks are trained for specific emotion and yielded the output of O_{e_A} , O_{e_H} , O_{e_S} and O_n . Although one classifier network is sufficient to identify different emotions with proper threshold values, four classifier networks architecture is selected. This is based on the preliminary experiment finding that four networks works better than one classifier network. From the output, a normalization process is subsequently conducted where the total of all emotion weight is 1 ($O_{e_A} + O_{e_H} + O_{e_S} + O_n = 1$). The

identified emotion class is determined using the maximum normalization value of a network trained using a particular emotion. However, this approach is only suitable for supervised learning where the target class is known prior to testing.

5.1.1 Data Structuring for Different Types of Experiments

The normalized speech data is appropriately arranged to ensure the different mode of experiments for intra, inter and mixed-cultural assessments can be achieved. Normalized speech refers to a pre-processed speech so that it is considered *clean* from noise or artifacts in order to make the emotion information in the speech easily recognized. In this chapter the experiments are designed based on 3 main modules; namely: type of validation, homogeneity of training dataset and culture variance of testing dataset as illustrated by Figure 5.2.

The first module determined the validation method used and in this case only two types of validation are needed to be considered; mainly: localization and generalization. Localization referred to *in-sample* type of experiment where the same data is used for training and testing. This approach manages to empirically measure the capability of the proposed speech emotion recognition system to memorize the data. However, localization approach may not be sufficient as most practical system will encounter unseen data in real life application. Hence, the generalization approach is introduced to enable a more detailed study of the proposed approach for global validation through *out-sample* type of experiment. This approach ensures the network does not see the data prior to training. K-fold cross-validation technique is used so that the performance of the system measured will be more unbiased.

The second module would test the performance of the system based on the homogeneity in the training dataset. Homogeneous training data addressed the performance of the speech emotion recognition system based on a single culture speech emotion training dataset. This data arrangement enables disparity analysis between different emotions that are within the same culture so that variation of emotion within the culture can be studied in-depth. In this thesis, heterogeneous mixed-cultural training data combined two or three datasets of different cultures (mixed-cultures); namely: Berlin, NTU_American and NTU_Asian. The combination of different cultures in the training data is geared towards generalization of the emotion universality. This notion is based on the understanding that emotion is universal thus its acoustical characteristics are similar regardless of their culture. For instance, it is assumed that anger in Berlin dataset is similar to NTU_American and NTU_Asian dataset although the cultures of these three datasets are different. This kind of experiments also allows further investigation for the speech emotion recognition cultural dependency given mixed cultural dataset for training.

Figure 5.2: Summary of Experimental Design for the Different Experimental Set-up

The third module would analyze the culture variance of the testing dataset. This experiment detailed study of the intra and inter-cultural influence effect in speech

emotion. In this chapter intra-cultural refers to speech emotion data based on similar culture used for both training and testing whereas inter-cultural applies to different culture speech emotion data used for training and testing. Example of intra-cultural experiment would be when Berlin dataset is used for training and testing. On the contrary, an example of inter-cultural experiment would be when Berlin dataset is used for training and tested with NTU_American or NTU_Asian datasets. The inter-cultural experiment can also be described as blind testing for different culture speech emotion data. This is because by arranging the testing data in this manner, it is fully independent from the influence of the cultural context of the training dataset. Hence, minimal culture effect can be examined in detail. Figure 5.2 shows the summary of the module for the different experimental design adopted in this thesis.

Further explanations on the experiments conducted is presented in Figure 5.3. The block diagram of the data structures shown in Figure 5.3 illustrated the different experimental setup to determine the effect of culture on recognizing emotion through speech. SW1 is a three-way switch to select the dataset A, B or C to be used as input to the classifiers for training and testing. SW2A and SW2B are four-way gang switches to select four different types of experimental set up. SW3 is a three-way gang switch to select the correct data set to be used for training and testing for the inter-cultural experiment.

In all the homogeneous intra-cultural experiments, switches SW2A and SW2B will be directed to terminal 1 and SW1 will select the input dataset A, B or C respectively. For the heterogeneous (80/20) mixed-cultural experiments, switches SW2A and SW2B will be connected to terminal 2 and for the heterogeneous (50/50) mixed-cultural experiments, switches SW2A and SW2B will be connected to terminal 3. Subsequently,

switches SW2A and SW2B will be connected to terminal 4 and switch SW3 will be used to select the respective input dataset for the inter-cultural experiment.

Homogeneous intra-cultural experiment (Switch SW1)

Homogeneous intra-cultural experiment refers to speech emotion identification task for a single cultural setting. The objective of this experiment is to analyze the disparity between emotion within the same culture so that the intra-cultural effect can be studied in detail. In this thesis, both localized and generalized homogeneous intra-cultural experiments are conducted to enable local validation and global validation comparison of the proposed speech emotion recognition system performance. Each dataset of Berlin, NTU_American and NTU_Asian consist of 4000 instances with 1000 instances per emotion. The instances which represent the expected emotion class is randomly arranged to provide minimal bias towards a particular emotion. For localized homogeneous intra-cultural experiment, same data is used for training and testing. However, the randomized dataset pool is segregated into five fold for cross validation purposes for generalized homogeneous intra-cultural experiment. 80% of the dataset (3200 instances) is used for training and the remaining 20% of the dataset (800 instances) is used for testing. Experiments on each training-testing dataset pair are repeated until all pair can be fully utilized. The overall contingency matrix is then computed by taking the mean of each fold experimental results.

Heterogeneous (50/50) mixed-cultural experiment

In order to study in detail the effect of partial culture dependency on speech emotion recognition, mixed culture-influenced dataset are combined as a single pool of

training/testing data pair. Each data pool consists of 8000 instances for two dataset combined and 12000 instances for three dataset combined respectively. The 50/50 rule is applied as a control experiment to ensure the proposed system does not bias towards the higher number of culture-influenced instances. 50% of the data pool (4000 and 6000 instances for two-dataset and three-datasets combined respectively) is used for training and the other half is used for testing. Experiment on each training-testing dataset pair are repeated until both has been trained and tested. Result from each fold are combined to produce overall contingency matrix.

Figure 5.3: Block Diagram for the Different Experimental Set-up

Heterogeneous (80/20) mixed-cultural experiment

In addition to the control experiment, heterogeneous (80/20) mixed-cultural experiment is conducted to analyze the proposed speech emotion recognition system performance given culture bias training-testing data pair. The dataset is combined together randomly and segregated into 5 fold. 80/20 rules is applied where 80% of the data (6400 and 9600 instances for two-dataset and three-dataset combined respectively) is used for training and the other 20% of the data (1600 and 2400 instances for two-dataset and three-dataset combined respectively) is used for testing. Such experimental set up is designed based on the hypothesis that emotion is universal hence similar emotion from different culture shares similar acoustic characteristics.

Inter-cultural experiment (Switch SW3)

Inter-cultural experiment refers to different dataset used for training and testing. In this thesis, both homogeneous and heterogeneous inter-cultural experiments are conducted. In homogeneous inter-cultural experiment, a single culture dataset is trained while different culture dataset is used for testing. For instance, Berlin dataset is used for training and NTU_American or NTU_Asian dataset is used for testing and vice-versa. Subsequently, heterogeneous inter-cultural experiment applies to combination of culture-influence data used in training and tested with different dataset. For example, combination of Berlin and NTU_American is used as training dataset to be tested with NTU_Asian dataset. The results are presented in a contingency matrix. By designing experiments in this manner, the testing dataset is totally independent from the culture that influenced the training dataset and can be regarded as blind testing. This is to access the inter-cultural impact to the global accuracy and to demonstrate that culture does influence the speech emotion recognition system performance.

5.2 Speech Emotion Identification Result and Discussion

Speech emotion recognition analysis consists of either verification or identification task. Verification distinguishes the existence of one emotion in the speech emotion testing instances from the rest. It is a two-class classification problem that handles the task of inferring either the testing instances is the intended emotion (denote by value 1) or its absence (denote by value 0) as discussed in Chapter 4. In contrast, identification task focuses on the ability of the proposed speech emotion recognition system to discriminate the specific emotion from a series of target classes given a set of testing data instances. Hence identification would require more than one class to be classified, which presented a greater challenge compared to verification task.

The emphasis of this thesis is to identify intra-cultural and inter-cultural effect to the proposed basic speech emotion recognition system. Four basic emotion classes are used for labeling the different basic emotion of anger, sadness, happiness and neutral acting as emotionless state. Inputs to the classifiers are examples of different speech emotion instances that represented each target emotion class for training exercise. To test the classifiers' performance, the unknown input speech emotion instances are identified based on the target emotion class that best matches the input speech. For the experiment, the anger emotion class is labeled as '*angry*', the happiness emotion class is labeled as '*happy*' and sadness emotion class is labeled as '*sad*'. These terms will be used interchangeably throughout the thesis and refers to the same meaning.

5.2.1 Speech Emotion Identification Result for Homogeneous Intra-cultural Assessment

Homogeneous intra-cultural assessment uses the same culture-influenced dataset for training and testing. The localized homogeneous intra-cultural experiment focuses on the proposed speech emotion recognition capability to memorize the speech emotion data pattern whereas the generalized homogeneous intra-cultural experiments are conducted to test the proposed speech emotion recognition system performance to globally discriminate different emotions from the same dataset. It provides information about the degree of variation for specific emotion for intra-cultural analysis. Tables 5.1 and Table 5.2 show the consolidated contingency matrix of the performance obtained from the localized and generalized homogeneous intra-cultural experiments. The three speech emotion datasets with different culture are used to provide a comprehensive understanding of the proposed speech emotion recognition system capability in handling emotion identification task within the same culture.

Table 5.1. Accuracy for Localized Homogeneous Intra-cultural Experiment (%)

Emotion	Berlin			NTU_American			NTU_Asian		
	MLP	ANFIS	GenSo	MLP	ANFIS	GenSo	MLP	ANFIS	GenSo
<i>Angry</i>	73.4	95.7	100	77.7	96.6	100	69.8	87.7	100
<i>Happy</i>	68.0	91.5	100	65.4	97.1	100	76.2	92.1	100
<i>Neutral</i>	74.0	95.7	100	76.9	96.4	100	72.2	88.0	100
<i>Sad</i>	82.1	97.0	100	81.7	98.3	100	82.3	90.6	100
Mean	74.4	95.0	100	75.4	97.1	100	75.1	89.6	100

Table 5.2. Accuracy for Generalized Homogeneous Intra-Cultural Experiment (%)

Emotion	Berlin			NTU_American			NTU_Asian		
	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>
<i>Angry</i>	69.9	65.9	56.7	68.9	75.8	70.7	62.8	68.9	61.6
<i>Happy</i>	48.8	51.7	47.2	63.0	70.2	69.8	69.7	72.2	61.9
<i>Neutral</i>	66.0	66.4	59.7	75.0	75.2	71.1	61.6	69.9	64.5
<i>Sad</i>	75.8	76.1	65.5	74.7	79.9	77.8	63.9	67.6	61.6
<i>Mean</i>	65.1	65.0	57.3	70.4	75.3	72.4	64.5	69.7	62.4

Result in Table 5.1 shows that GenSoFNN yielded perfect accuracy since the classifier is able to completely discriminate locally the different emotion in the different dataset. It demonstrates that GenSoFNN can well memorize the emotion patterns provided by training dataset due to the GenSoFNN ability to construct specific and complex rules. ANFIS exhibited high accuracy result for localized homogeneous intra-cultural experiment with mean accuracy of 93.9% for the three datasets used. This result shows a slight decrement from GenSoFNN result because ANFIS uses subtractive clustering method to fuzzify the input data. Such approach enables most of the emotion pattern covered by rules generation hence explained the slight degradation of the proposed system performance. However, MLP shows the lowest performance with a mean accuracy of 75% for the three datasets. It reflects that MLP is not able to memorize the emotion pattern as well as GenSoFNN and ANFIS. This may be due to MLP characteristic that treats its learning process as a black box. In addition, sadness is the most easily discriminated emotion compared to the other emotion in all datasets with exception in the ANFIS result using NTU_Asian. Hence, the result in Table 5.1 shows

that the proposed speech emotion recognition system is able to discriminate emotion with at least 65% accuracy for local validation given homogenous intra-cultural dataset.

Table 5.2 demonstrates the ability of the proposed speech emotion recognition system to identify emotion from unknown instances after trained with homogeneous intra-cultural dataset. From Table 5.2 it is observed that the accuracy of the Berlin dataset is within 47.2% (happy-GenSoFNN) to 76.1% (sad-ANFIS) range with mean accuracy of 62.5% while for the NTU_American dataset it falls between 63.0% (happy-MLP) to 79.9% (sad-ANFIS) range with mean accuracy of 72.7%. In addition, NTU_Asian performance yielded between 61.6% (shared by neutral-MLP, angry-GenSoFNN and sad-GenSoFNN) to 72.2% (happy-ANFIS) range with a mean accuracy of 65.5%. Thus, the performance using the NTU_American dataset is slightly better than the Berlin and the NTU_Asian datasets with differences of 10.2% and 7.2% accuracy respectively. It is noted that the classification results of the Berlin dataset is more stable with an average variance of 2.5% followed by the NTU_American (3.5%) and the NTU_Asian (5.6%) datasets. Happiness consistently yielded the lowest accuracy in all experiment of the three datasets.

On the contrary the sadness emotion produced the highest correctly identified emotion with the exception of the NTU_Asian dataset. The results are consistent with findings reported by other researchers that happiness is more difficult to classify compared to anger and sadness. However, it is interesting to note that the NTU_Asian result does not fit in the same pattern as Berlin and NTU_American datasets. It is suspected that Asian culture (specifically Malay culture in this case) is biased towards happy emotion due to local custom of politeness hence their speech reflects the underlying happiness which dominate their emotion. The results also show that as far as

the classifiers are concerned, no classifier can claim superiority since there is no obvious difference of global validation accuracy between one to the other. GenSoFNN obtained a mean accuracy of 64.0% followed by 66.7% and 70.0% for MLP and ANFIS respectively for the three datasets. The degradation in GenSoFNN result can be a result of the complex and specific rules generated thus it restricts the GenSoFNN capability to infer the correct emotion for the generalized experiment but is particularly useful in localized experiment.

Figure 5.4: Summary for Generalized Homogeneous Intra-cultural Experimental Result using Different Dataset and Classifier Pairing.

Based on both localized and generalized homogeneous intra-cultural experimental result, it can be observed that the performance of the NTU_American and the NTU_Asian dataset are comparable to the Berlin dataset, which is considered a standard speech emotion dataset. Results obtained from this experiment also support the other research findings that has identified that it is easier to determine sadness and anger compared to happiness with exception of the NTU_Asian dataset. Subsequently, this finding also demonstrates that combination of MFCC as feature extraction method coupled with the three classifiers of MLP, ANFIS or GenSoFNN can achieve reasonable degree of

accuracy with at least 55% performance for global validation as summarized in Figure 5.4. The proposed basic speech emotion recognition approach has potential in recognizing emotion in speech intra-culturally with accuracy two times better than chance probability (25% for 4 classes).

5.2.2 Speech Emotion Identification Result for Heterogeneous Mixed-cultural Assessment

Based on the assumption that emotion is universal, it is hypothesized that emotion in speech signals shares common acoustical features regardless of the cultural influence to certain extent. For instance, anger in the speech Berlin dataset is assumed to be similar to anger in the speech NTU_American and NTU_Asian datasets and vice versa. It is also assumed that by combining the datasets, the inter-cultural effect can be minimized. Thus, similar emotion instances from different datasets are grouped together as a data pool and arranged randomly to avoid bias towards any dataset in this experiment. Both of the training and testing dataset share similar mixed culture influence since the different cultures have been combined together in the data pool which provides mixed-cultural assessment. Hence, the rationale of this experiment is to study the partial dependency culture effect on speech emotion.

The 50/50 and 80/20 rule experiments are applied to demonstrate different cultures influence effect on the proposed speech emotion recognition system performance. The 50/50 rule experiment is conducted to provide a basis of comparison for equal cultural influence of the data pool that represented an ideal scenario. However, such data arrangement is quite difficult to obtain in real-life situation. Nonetheless, the

80/20 rule experiment is developed to emulate the situation when the data pool is bias towards a particular culture.

Table 5.3. Accuracy for Localized (In-sample) Heterogeneous (50/50) Mixed-Cultural Experiment (%)

Emotion	COMBINATION OF 2 CULTURES									COMBINATION OF 3 CULTURES		
	<i>Berlin + NTU_American</i>			<i>NTU_American + NTU_Asian</i>			<i>Berlin + NTU_Asian</i>			<i>Berlin + NTU_American + NTU_Asian</i>		
	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>
<i>Angry</i>	75.2	85.5	100	74.9	77.4	100	75.7	75.3	100	73.7	73.6	100
<i>Happy</i>	66.5	76.6	100	72.7	75.4	100	69.3	66.6	100	67.6	64.7	100
<i>Neutral</i>	73.1	85.2	100	72.8	78.2	100	71.8	72.6	100	71.4	71.3	100
<i>Sad</i>	80.5	91.8	100	75.9	79.9	100	75.2	78.6	100	78.3	76.2	100
<i>Mean</i>	73.8	84.8	100	74.1	77.7	100	73.0	73.3	100	72.7	71.5	100

Table 5.4. Accuracy for Generalized (Out-sample) Heterogeneous (50/50) Mixed-Cultural Experiment (%)

Emotion	COMBINATION OF 2 CULTURES									COMBINATION OF 3 CULTURES		
	<i>Berlin + NTU_American</i>			<i>NTU_American + NTU_Asian</i>			<i>Berlin + NTU_Asian</i>			<i>Berlin + NTU_American + NTU_Asian</i>		
	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>	<i>MLP</i>	<i>ANFIS</i>	<i>GenSo</i>
<i>Angry</i>	63.4	64.3	55.6	57.9	63.6	58.3	63.0	61.7	49.7	58.1	60.5	52.0
<i>Happy</i>	51.5	55.3	50.6	58.2	61.2	59.7	50.3	53.4	46.0	47.3	51.5	50.4
<i>Neutral</i>	60.8	62.2	52.9	60.4	64.2	60.4	59.4	60.2	54.6	57.0	60.4	53.9
<i>Sad</i>	66.6	72.7	60.3	64.4	69.3	65.9	63.6	68.3	55.9	61.6	65.6	56.9
<i>Mean</i>	60.5	63.6	54.8	60.2	64.6	61.1	59.1	60.9	51.5	56.0	59.5	53.3

Tables 5.3 and Table 5.4 present both localized (in-sample) and generalized (out-sample) heterogeneous (50/50) mixed-cultural experimental result respectively. The degradation of the proposed speech emotion recognition system performance of this experiment compared to the homogeneous intra-cultural experiment is expected, as the fusion of datasets would introduce more non-linear components to the input data that complicate the classifier to discriminate the emotion and also partly due to the suppressed inter-cultural effects. The heterogeneous mixed-cultural training dataset seems to be worse off than the homogeneous training dataset experiments by about 2.6% to 8.5% range. Thus, it is shown that culture is one of the parameter affecting the performance of the speech emotion recognition system.

One of the interesting observations is that GenSoFNN yielded perfect performance score for local validation experiment regardless of the cultural combination. The reason is because the complexity of the rules generated by GenSoFNN facilitates it to memorize the emotion pattern present in the training data. However, this characteristic hinders the classifier to achieve good accuracy in generalization experiment due to its rigidity. ANFIS obtained a mean accuracy ranging from 71.5% to 84.8% and MLP yielded a mean accuracy between 72.7% and 74.1% respectively for local validation in the heterogeneous mixed-cultural experiment.

From the results in Tables 5.3 and Table 5.4, happiness is the hardest emotion to determine with a lowest performance with mean accuracy of 80.0% for local validation and mean of 53.0% for global validation. On the contrary, sadness obtained the highest mean accuracy of 86.4% for local validation and 64.3% for global validation. This result pattern is similar to the previous homogeneous intra-cultural assessment. All of the classifiers presented a comparable performance with small variances from one classifier

to another. In the generalized heterogeneous (50/50) mixed-cultural experiment, ANFIS yielded the highest accuracy for all data pool combination with a mean of 62.2% followed by MLP with accuracy differences of 3.2% and GenSoFNN with 7.0% respectively. However, GenSoFNN shows a more consistent result of identifying one emotion over another with very small variance of about 2.8% as compared to ANFIS with mean variances of about 3.8% and MLP with emotion variances of about 3.9%.

(a) Berlin + NTU_American

(b) NTU_American + NTU_Asian

(c) Berlin + NTU_Asian

(d) Berlin + NTU_American + NTU_Asian

Figure 5.5. Generalized Heterogeneous (80/20) Mixed-Cultural Assessment Result for
Different Dataset Combination

The heterogeneous (80/20) mixed-cultural experiment is conducted to investigate the proposed system performance when trained with biased training data pool. The performance of such experiment is estimated to be almost similar to the generalized heterogeneous (50/50) mixed-cultural experiment based on the hypothesis that as long as classifiers are trained with sufficient number of a particular culturally-influenced speech emotion dataset, the classifiers are able to discriminate the different emotion with reasonable degree of accuracy even for the minority culture in the data pool. Figure 5.5 shows the experimental result for the heterogeneous (80/20) mixed-cultural experiment for different data pool mixture. It is observed from the result in Figure 5.5 that ANFIS performs with slightly better accuracy with a mean accuracy of 65.0% compared to MLP with 61.8% and GenSoFNN with 60.1%. Mean overall accuracy for 2-cultures combined is slightly higher than the mean overall accuracy for 3 cultures with difference ranging from 1.2% to 6.4% ratio. This result confirms the initial assumption that a higher mixture of the number of cultures may result in higher non-linear element in the data pool hence complicates the classifier to correctly identify the emotion. Happiness is consistently the most difficult emotion to be identified correctly in all data pool mixture dataset cases with the worst performance recorded for Berlin + NTU_American + NTU_Asian data pool using MLP as classifier with 49.5% accuracy whereas sadness yielded the highest accuracy with best performance recorded at Berlin + NTU_American data pool using ANFIS as classifier with 74.8% accuracy.

In order to fully understand the partial dependency effect of culture on the speech emotion recognition system, both the generalized 50/50 and 80/20 heterogeneous mixed-cultural experimental results are compared. Each rules representing the culture bias weight of the data pool. The generalized heterogeneous 50/50 mixed-cultural experiment is representing equal weight culture-influenced data pool whereas generalized

heterogeneous 80/20 mixed-cultural experiment representing biased culture-influenced data pool. Figure 5.6 shows the summary of the experimental result of both heterogeneous mixed-cultural assessments. It can be clearly observed that the generalized heterogeneous (50/50) mixed-cultural result suffers a slight inclination between 1.6% (recorded in Berlin + NTU_Aasian data pool using MLP classifier) to 5.3% (recorded in the Berlin + NTU_American data pool using GenSoFNN classifier). This result is due to the culture bias influence towards a certain culture in the data pool which caused the classifier to favor the majority of culture-influence dataset in the data pool.

Figure 5.6. Result Comparison for Overall Accuracy of 80/20 and 50/50 Rule Generalized Heterogeneous Mixed-Cultural Assessment

The result of both generalized heterogeneous 80/20 and 50/50 mixed-cultural assessment obtained a ratio of 51.5% to 67.6% accuracy which demonstrates that the proposed speech emotion recognition system is able to identify speech emotion given sufficient amount of culturally-influenced training dataset. Hence it can be concluded that there are common acoustical characteristics for similar speech emotion across cultures that enables a reasonable accuracy of at least 50% even though a mixture of cultures is lumped in the data pool.

5.2.3 Speech Emotion Identification Result for Inter-cultural Assessment

Inter-cultural assessment is conducted by using a different culturally-influenced dataset for training and testing. The rationale of this set of experiment is to empirically demonstrate the impact of fully independent cultural dependency to speech emotion recognition performance. Moreover, this kind of assessment can also be regarded as blind testing experiment. Based on the hypothesis that emotion is universal, a particular emotion in one culture is supposed to be acoustically similar across cultures. Hence, the inter-cultural experiment is also conducted to confirm the understanding that culture influences the performance of the speech emotion recognition system performance and need to be taken into consideration as an important design parameter.

Inter-cultural experiment in this thesis is divided into two types based on the homogeneity of the training data. Both experiments only focus on generalized experiment since only different training-testing dataset pair is needed. Homogeneous inter-cultural experiment refers to a single culturally-influenced dataset used for training and tested with a different culturally-influenced dataset. For instance, the Berlin dataset is used for training and tested with the NTU_American or the NTU_Asian and vice-versa. This is to

study the similarity and dissimilarity measure of the particular emotion across cultures. If the accuracy obtained is high, it means that the particular emotion for both training-testing dataset is acoustically similar which enables the classifier to correctly identify it. On the other hand, it can also be deduced that both training-testing data pair have minimal similarity if the accuracy is low. Subsequently, heterogeneous mixed-cultural experiment applies to combination of cultures dataset used for training to test with a single culturally-influenced dataset. An example of this experiment is when a combination of the Berlin and the NTU_American dataset is used for training and tested with the NTU_Asian dataset. The objective of this experiment is to analyze the correlation between the mixed culture datasets for a particular emotion since it is assumed that the effect of inter-culture is suppressed when the different culture-influenced datasets are combined.

Figure 5.7 shows a radar chart of the proposed speech emotion recognition system performance for homogeneous inter-cultural experiment. A radar chart is a useful way to display multivariate observations in the form of two-dimensional chart. The radar chart is also known as star plot [Cham83]. In this thesis, the radar chart is used to show the inclination pattern of the dataset used for testing towards a particular emotion. The value represents the accuracy achieved of speech emotion identification task while the spike or radii represent a particular identified emotion.

From radar diagram of Figure 5.7, it is noticed that the closure of each radar eye shows the poor performance of the particular emotion in training-testing dataset pair. In almost all cases, the homogeneous inter-cultural experiments yielded very poor overall accuracy that is almost as low as chance guessing (25% for 4 classes).

(a) Homogeneous Inter-Cultural Experimental Result for Testing with Berlin Dataset

(b) Homogeneous Inter-Cultural Experiment Result for Testing with NTU_American Dataset

(c) Homogeneous Inter-Cultural Experiment for Testing with NTU_Asian Dataset

Figure 5.7: Radar Chart of the Performance for Homogeneous Inter-Cultural Experiment with Different Training-Testing Dataset Pair

Based on Figure 5.7, some of the specific emotion accuracy can be as low as 8.1% for happiness in train-NTU_American and test-Berlin dataset pair using MLP classifier. However there are emotions that scored high accuracy. Best-recorded result is yielded when the train-NTU_American and test-NTU_Asian dataset pair coupled with the ANFIS classifier. This result demonstrates that there are emotions that are universally accepted across the three cultures and yet there are emotions that are exclusive for a particular culture.

In addition, referring to the result in Figure 5.7, it is clear that neutral provides the highest mean accuracy by all classifiers that yielded 33.7% for GenSoFNN, 35.4% for MLP and 39.6% for ANFIS respectively. This finding is indeed significant as it confirms that neutral is collectively accepted by all three datasets as universal such that neutral can still be easily identified regardless of the different culture training-testing dataset pair.

Result of homogeneous inter-cultural experiment using Berlin dataset as testing data in Figure 5.7(a) seems to be equally distributed among all three emotions of anger, happiness and sadness. On the contrary, result of homogeneous inter-cultural experiment using the NTU_American dataset as testing dataset in Figure 5.7(b) seems to be biased towards neutral and happiness whereas result of homogeneous inter-cultural experiment using NTU_Asian dataset in Figure 5.7(c) seems to be biased towards neutral. This observation can be noticed from the radar eyes opening presented in Figure 5.7. When looking more closely at Figure 5.7(a), sadness obtained high accuracy for all classifier when training-NTU_Asian and testing-Berlin dataset pair used with 33.3%, 43.3% and 56.6% for GenSoFNN, MLP and ANFIS classifier used respectively. It means that sadness in the Berlin dataset is acoustically similar to sadness in NTU_Asian dataset.

(a) Tested NTU_Asian

(b) Tested Berlin

(c) Tested NTU_American

Figure 5.8. Radar Chart of the Accuracy for Heterogeneous Mixed-Cultural Experiment
with Different Training-Testing Dataset Pair

In Figure 5.7(b), happiness and sadness obtained high accuracy for training-NTU_Asian and testing-NTU_American with 33.1% for GenSoFNN, 40.9% for MLP and 43.1% for ANFIS in discriminating sadness emotion and 26.9% for GenSoFNN, 42.3% for ANFIS and 54.3% for MLP in identifying happiness emotion. This can be explained that happiness and sadness in NTU_American are acoustically similar with the same emotion in NTU_Asian.

In addition, the radar chart of Figure 5.8 shows heterogeneous mixed-cultural experimental result for Figure 5.8(a) of combining the Berlin and the NTU_American as training dataset coupled with the NTU_Asian testing dataset, Figure 5.8(b) of combining the NTU_American and NTU_Asian datasets coupled with the Berlin testing dataset and Figure 5.8(c) of combining the Berlin and the NTU_Asian training datasets coupled with NTU_American testing dataset. From the result, neutral in both Figure 5.8(a) and Figure 5.8(b) yielded the highest accuracy while happiness yielded the highest accuracy as obtained in Figure 5.8(c). This result can be deduced that neutral in NTU_Asian is acoustically similar to neutral in the combination of Berlin+NTU_American and neutral in the Berlin dataset is acoustically similar to the combination of NTU_American+NTU_Asian. It is also observed that happiness in NTU_American is acoustically similar to happiness in the combination of Berlin+NTU_Asian dataset.

Based on the experimental results of homogeneous inter-cultural and heterogeneous mixed-cultural assessments, some interesting observations can be noted that lead us to conclude certain emotion which shares common characteristics among different cultures indicating the universality of the emotion and yet the distinct acoustic features of each culture exist for some emotion allowing diversity of the emotion. Hence, culture does influence the speech emotion recognition system performance.

5.2.4 Observation

It is obvious that the proposed basic speech emotion recognition system cannot appropriately address the correct emotion for a particular speech emotion instances if its culture is not explicitly introduced to the classifier as empirically demonstrated in inter-cultural assessment. Inter-cultural effect makes the overall accuracy degraded compared to intra-cultural effect. This result leads us to believe that culture does play an important role in influencing emotion in speech.

Figure 5.9: The Summary of Performance for Different Cultural Dependency Assessment.

The effect can be clearly seen in the inclination of accuracy from homogeneous intra-cultural experiment which represents full cultural dependency, followed by heterogeneous (80/20) mixed-cultural and heterogeneous (50/50) mixed-cultural intra-cultural experiment which represents partial culture dependency and lastly by heterogeneous mixed-cultural inter-cultural experiment which represents full-culturally independent data structure. A summary of the result is depicted in Figure 5.9.

The experimental results also provide an interesting observation that although generalized heterogeneous (50/50) intra-cultural experiments used only partial cultural dependency of a particular dataset for training, it still yielded comparable accuracy with the homogeneous intra-cultural experiments that is fully dependent on the particular culture. This finding is in line with the initial assumption that if the classifier is provided with sufficient number of instances during training for a particular culture, it could provide reasonable degree of accuracy. However, inter-cultural experimental results are very poor that in almost all cases the accuracy is close to chance or guessing (25% for 4 classes) except for certain emotion like neutral. This finding partly supported the reason that some successful speech emotion recognition system fails spectacularly when implemented in different region of the world due to the intra-cultural and inter-cultural effect. Moreover, this finding cemented the notion of culture dependency of speech emotion recognition performance that if the cultural effect can be minimized, better performance can be achieved. Thus, to boost up the proposed speech emotion recognition system accuracy, a two-stage culture-sensitive speech emotion approach is proposed to address culture recognition issue and will be discussed in the next section.

5.3 Two-Stage Culture-sensitive Speech Emotion Approach for Culture Detection

As discussed earlier, culture is often regarded as an exiguous parameter by many researchers hence the implementation of automated speech emotion recognition system for global use is still a challenge. Pell et al. [Pell09a] in their study shows that speakers of four distinct languages; namely: English, German, Hindi and Arabic, exhibited many similarities in how they express and identify vocal expressions of emotion despite important differences in the language they are speaking and in their linguistic-cultural

background. However, Pell et al. [Pell09b] in their other study reported that they tested 61 monolingual Spanish with a subset of the English, German and Arabic utterances [Pell09a] and found that the judges could identify six basic emotion of anger, sadness, disgust, happiness, surprised, fear as well as neutral with overall accuracy ranging from 56% to 59% while scored 64% overall accuracy when presented with native Spanish utterance. Based on these finding, it is hypothesized that if cultural information is known before speech emotion recognition is conducted, the inter-cultural effect can be suppressed and gain better accuracy due to intra-cultural advantages. Thus, in this thesis, a two-stage culture-sensitive speech emotion approach is proposed to recognize culture in speech emotion data.

Figure 5.10: Block Diagram of the Proposed Two-Stage Culture-Sensitive Speech Emotion Approach

Figure 5.10 shows the block diagram of the two-stage culture-sensitive speech emotion approach to detect culture and emotion. At the first layer, all emotion data of the respective culture is assembled and denote as culture data and used as a training data. The classifier would classify the group of data (or dataset) based on the culture identification. The outcome of the first layer is a culture that would determine the second layer network for emotion recognition based on the classifier trained using the intra-cultural setting. This approach will ensure that the instance will be classified by a trained network accordingly to its culture-influenced speech emotion dataset.

5.3.1 Culture Detection Experimental Result and Discussion

Similar to the proposed basic emotion recognition system approach, two types of experiment are conducted using the generalized heterogeneous (80/20) mixed-cultural assessment and heterogeneous (50/50) mixed-cultural assessment for the first stage of the culture-sensitive speech emotion approach. Generalized heterogeneous (80/20) mixed-cultural assessment for culture detection refers to culture identification task combination of the three culture-influenced datasets using 80/20 rules where 80% of the dataset (9600 instances) is used for training and the remaining 20% (2400 instances) used for testing. These data pairs will repeatedly be used until all of them are trained and tested. Overall performance is calculated by taking the mean of each fold accuracy result. Generalized heterogeneous (50/50) mixed-cultural assessment for culture detection applies to the combination of the three cultures with equal culture-influenced weight as training data where 50% of the data pool (6000 instances) is used for training and the other half (6000 instances) is used for testing and vice-versa. Each culture represented by the three different dataset is used as a label for the target class.

Figure 5.11(a) shows the culture detection result for generalized heterogeneous (80/20) mixed-cultural assessment for culture detection experiments. It can be observed that the three classifiers yielded high accuracy ranging between 59.5% and 98.8%. European culture which is represented by the Berlin dataset is consistently to be easily identified by all classifiers with accuracy of 88.1% for MLP, 93.1% for GenSoFNN and 98.8% for ANFIS and an overall mean accuracy of 93.4%. This is followed by the the American culture represented in the NTU_American dataset which yielded a mean accuracy of 83.6% and followed subsequently by the Asian culture with a mean accuracy of 80.6%. As the performance of each classifier, ANFIS obtained the highest accuracy with a mean accuracy of 97.4% followed by MLP with a mean accuracy of 85.9% and GenSoFNN with a mean accuracy of 74.4%. ANFIS is also observed as the most consistent classifier with minimal variance between the cultures detected with mean variance value of 1.1% and followed closely by MLP with 1.8% variance value. However, GenSoFNN does not perform well for the culture detection task for the NTU_American and the NTU_Asian dataset which can be observable in Figure 5.11(a). This could be because MFCC features are not really suitable for GenSoFNN for the generalized heterogeneous (80/20) intra-cultural assessment for culture detection.

Figure 5.11(b) shows result of generalized heterogeneous (50/50) mixed-cultural assessment for culture detection. The Berlin dataset which is representing the European culture obtained the highest score of overall mean accuracy of 93.5%. Both the American and the Asian cultures yielded 82.6% and 81.8% overall mean accuracy respectively. However, for consistency of classifier performance, MLP leads the rank with mean variance value of 0.8%, followed by ANFIS with 1.0% mean variance value and GenSoFNN with 13.3% mean variance value. It seems that GenSoFNN can only identify European culture with high accuracy than the other two cultures of American and Asian.

(a) 80/20 Rules

(b) 50/50 Rules

Figure 5.11: Experimental Results for Culture Detection for Different Heterogeneous Mixed Cultural.

Based on this result in both Figure 5.11(a) and Figure 5.11(b), it is obvious that the proposed technique are able to obtain a good result for culture detection in stage 1 of the two-stage culture-sensitive speech emotion approach. Once the culture can be detected, the homogeneous intra-cultural assessment (fully culture dependant) can be conducted as described in the earlier experiment of Section 5.2.1. The accuracy will improve at least 30% compared to the heterogeneous mixed-cultural assessment (culture independent) as illustrated in Figure 5.9. Hence, it is plausible to implement such approach in the speech emotion recognition system to boost up the global accuracy.

5.4 Summary and Conclusion

In this chapter, two main issues of cultural influence on speech emotion recognition system and culture detection from speech emotion dataset are addressed. In order to analyze both issues, intra-cultural and inter-cultural assessments are conducted based on a proposed basic speech emotion recognition system that comprised of MFCC features extraction coupled with three classifiers; namely: MLP, ANFIS and GenSoFNN to

discriminate anger, happiness and sadness emotion with neutral acting as emotionless class. The block diagram of the proposed system is illustrated in Chapter 3. Three different datasets; Berlin, NTU_American and NTU_Asian datasets, are used to introduce a fusion of cultures in the analysis. The Berlin dataset is collated from German speakers representing European culture and the NTU_American dataset from American English representing American culture. The NTU_Asian dataset gathered from Malay language is used to represent Asian culture.

From the generalized homogeneous intra-cultural assessment of emotion identification experiments, it is found that the proposed basic speech emotion recognition system could identify the emotions well regardless of the input data used with mean accuracy of at least 57.3%. The heterogeneous mixed-cultural experiment is then conducted by combining two or three different dataset together with assumption that emotion is universal thus the acoustical characteristic in speech are supposed to be similar. In addition mixing the multiple culture-influenced dataset is assumed to reduce the effect of the intra-cultural differences. The rationale of this experiment conducted is to study in details the effect of partial culture dependency to speech emotion recognition task. Result of the experiments showed that the heterogeneous mixed-cultural assessment is able to obtain a reasonable accuracy although it is lower than homogeneous intra-cultural experimental result. This result is expected as the combination of the dataset will introduce more non-linear element to the data pool which complicates the classifier task to correctly identify the speech emotion. This factor surely gives impact to the overall accuracy of the proposed basic speech emotion recognition system.

More experiments are carried out to investigate the completely independent cultural fusion effect on the performance of the basic speech emotion recognition system.

The inter-cultural assessment is carried out to analyze the degree of cultural effect by implementing different training-testing dataset pair to the proposed speech emotion recognition system in order to understand the universality of the basic emotion. From the experimental result it seems that the performance of the proposed speech emotion recognition system suffers badly to almost chance guessing accuracy (25% for 4 classes) due to lack of culture information. Hence, it is empirically verified that culture is important and has to be properly addressed in designing speech emotion recognition system as it influences the performance of such system.

Based on the finding in the intra-cultural and inter-cultural assessments, the idea of basic speech emotion recognition is extended to a two-stage culture-sensitive speech emotion approach to recognize culture and emotion. The experimental result of culture detection in stage 1 is presented in Section 5.4. The result shows that it is possible to detect culture with mean accuracy of at least 60% from three different cultures. Hence, the proposed approach can be implemented with speech emotion recognition system to firstly identify the culture before channeling the testing instances to its respective culturally influenced basic speech emotion recognition system. By identifying the culture, homogeneous intra-cultural assessment can be conducted (fully culture dependant). Relative 30% improvement can be obtained in comparison to heterogeneous mixed-cultural inter-cultural assessment that is independent completely from culture information. Based on the empirical results, this approach can improve the overall accuracy of such system and can be used as a step towards globally accepted speech emotion recognition system and can also be applied to many speech-processing applications to improve its time performance.

In conclusion, speech emotion is very much influenced by culture. By understanding the cultural effect in speech can help improved many speech emotion application systems in the future including understanding and analyzing driver behavior while driving for comfort and safety. In this chapter, it is observed that the effectiveness of the speech emotion recognition system within the same culture in recognizing emotion is fairly practical and can be easily attained. Thus the biggest challenge will be to analyze and understand the intrigue of cultural influence on speech emotion so that a more generalized system can be achieved. In the next chapter the Valence and Arousal (VA) analysis tools that allows visualization, analysis and understanding of the effect of culture and dataset variation with respect to the classifier's performance is proposed based on the Affective Space Model (ASM). Subsequently, the VA technique is employed to analyze a drivers' behavior state for practical implementation of speech emotion recognition system as an example of practical implementation. Detail description will be presented in Appendix AIII.

CHAPTER 6

AFFECTIVE SPACE MODEL APPROACH FOR SPEECH EMOTION UNDERSTANDING

From the social constructivist model of emotion perspective as discussed in Chapter 2, it is theorised that culture plays an important role in influencing the conveyance of emotion. The assimilation of culture is conditioned by the acceptance of society. Hence, it is not surprising that researchers start to comprehensively explore the correlation between emotion and culture [Grim07, Bezo83]. In addition, the fact that emotion is known to be propagated through speech justifies the employment of speech as one of the cheap and effective input for culturally-influenced emotion recognition system. Detailed explanation of the intra and inter-cultural assessment in Chapter 5 explicitly presented the culture influence on speech emotion. Chapter 5 is indeed a refinement work from preliminary study undertaken in Chapter 4. In Chapter 4, an emotion verification experiment to segregate emotion and its complementary class was conducted to give an overview of the performance of the proposed system in discriminating the emotion of interest. From the result of NTU_American emotion verification using MFCC features coupled with ANFIS classifier, see Table 4.5, it is clear that a specific emotion can be recognized with accuracy ranging from 85.3% (happiness) to 93.0% (anger).

Many researchers adopt the popular discrete-class approach for emotion recognition as discussed in Chapter 2. In fact, similar approach had also applied for intra-cultural and inter-cultural understanding as presented in Chapter 5. However, few weaknesses are encountered. Firstly, the discrete-class approach is only suitable for supervised learning where the testing instance must have mutual exclusive membership of

the training class. For instance, since the training class consists of basic emotions; namely: anger, happiness, sadness and neutral classes, the testing instance must belong to either one of this respective class. It is also found that if the correct training classes are not properly accounted for, valuable information is discarded resulting in inconsistent estimation and biased outcome. Furthermore, the dynamic characteristic of emotion is ignored although it is known that emotion is not a discrete event as proposed by the functionalist model of emotion. The rigidity of such approach restricts any other variation of class that is not trained to be recognized hence limiting the detailed study of emotion primitives and dynamics.

As mentioned earlier, culture plays an important role in emotion display and perception in speech. From the empirical result in Chapter 5, the proposed speech emotion recognition system which is tailored to a specific emotion can boost the overall accuracy of the system. However, finding a world-wide complete set of correctly annotated speech emotion is a challenge. An alternative solution to this issue is to develop a method that is customized to the data available (data-driven approach). Such solution can also be applicable to human behaviour analysis due to the fact that emotion is manifested through behaviour in relating to the situation and surrounding. This is based on the Cannon-Bard theory [Cann27] that suggested emotion is the catalyst that prompts one to react to stimuli. Hence, it is rational to deduce that behaviour is the outcome from emotion processing. Based on this notion, it would be beneficial if speech emotion analysis can be conducted using dynamic approach of the affective space model as it can describe the emotion primitives. Such analysis facilitates the understanding of the core attributes of emotion rather than limiting the focus to the customary discrete emotion classes.

The Valence-Arousal (VA) analysis approach is proposed in this chapter to complement the discrete-class approach. This approach is inspired from the psychologists' hypothesis based on the affective space model understanding that emotion can be represented in two different axes of the Valence (V) and Arousal (A). These VA axes can be graphically illustrated in such a way that similarity and dissimilarity measurement of the emotion pattern can be expressed and visualized as Euclidean distance between points in this affective space. The difference values of such axis can be used to automatically segregate one emotion from another. The advantages of such approach are:

- i. It provides a visualization tool for researchers to understand emotions in terms of the *effect* of the emotion (gathered from the degree of valence) and the *activation* of the emotion (obtained from the degree of arousal).
- ii. It allows the emotion profiles to be analyzed in details using the neutral recalibration based on data-driven approach. Neutral is always assumed to have zero values for Valence and Arousal emotion primitives of the affective space model. Unfortunately in almost all discrete-class approaches for emotion recognition system, the variations of neutral due to cultural and environmental effects have not been considered. This is because neutral is regarded as insignificant and irrelevant to speech emotion by most researchers.
- iii. It also allows the projection of the unknown speech emotion regions based on the mapping of the affective space model. This is derived from the understanding of the neutral-emotion boundary and the recalibration of the origin of the affective space model generated from a selected culturally-influenced dataset.

- iv. The dynamics of emotions can be easily visualized from the VA approach, which is not available in most speech emotion recognition system.
- v. In addition, the proposed approach can be applied to profile human behaviour based on hypothesis that behaviour is directly related to emotion in the VA domain. This is currently missing in the existing speech emotion recognition system.

Psychologists added Dominance as an axis in the 3D affective space model due to the fact that the Valence and Arousal axes are unable to fully discriminate certain important emotional distinctions [Grim07, Lang95, Schl54]. This situation arises when discriminating anger and fear. Both emotions are unpleasant and have high arousal thus both emotions occupy similar position in the 2D affective space model. Therefore, Dominance primitive (z-axis) is added. Dominance dimension represents the degree of control that the audience had while interacting. It is symbolized by the changes in the size of Self-Assessment Manikin (SAM); a large figure indicates maximum control in the situation and vice-versa [Brad94]. However, it is reported in the same paper that the Dominance factor has accounted for the least variance in affective judgment. Such finding implicated the impact of Dominance dimension on emotion recognition is considerably lower compared to Valence and Arousal dimensions. Hence, Dominance is not critical in discriminating emotion and will be ignored in this thesis. The result is consistent with Russell's earlier finding [Russ79, Russ77] that Valence and Arousal are accounted for by far the major proportion of variance in the affect scales while Dominance only accounted for a very small amount. Although Schlosberg stated that emotion can be represented in a 3D space [Schl54], a 2D affective space model is deemed sufficient for this work. It is also noted that the emphasis of this study is only using three basic emotions of anger,

happiness and sadness with neutral acting as the emotionless state (serving as a reference state). These three basic emotions can clearly be used to represent the three different quadrants of the 2D affective space model as described by Figure 2.3 of Chapter 2.

In this chapter, two issues will be investigated. Firstly, an alternative approach is developed based on the proposed VA analysis to complement the discrete-class approach so that a comprehensive study of speech emotion can be conducted. The implementation of VA approach can provide critical information and a clearer view from different perspective of the culturally-influenced speech emotion datasets. It is comprehensively based on the available data (data-driven) so that the effect of culture can be readily addressed and accessed. Experimental result presented in Chapter 5 for the discrete-class approach will be used to validate and complement the performance of the proposed VA approach.

Secondly, the proposed VA analysis will be employed to profile human behaviour state based on his/her speech emotion (Appendix AIII). A detailed study of human behaviour while driving based on the speech emotion understanding will be conducted. This study focuses on the driver behaviour state (DBS) analysis using the Real time Speech Driving (RtSD) dataset. It also serves as a practical example in implementing the VA speech emotion recognition analysis. The correlation experiment will be conducted by testing the RtSD dataset with the classifiers that are trained with different speech emotion datasets. Such experiments attempt to provide a detailed understanding of the particular DBS based on the different culturally-influenced emotion. These understanding may facilitate in designing practical application systems. An emotion-sensitive driver state indicator can be one of such system where perturbed driver can be given an early warning so that accident can be avoided. Perturbed driver can be emotionally quantified

by measuring his/her Valence and Arousal primitives' values. If one has negative value for both primitives, he/she has low activation (passive) and unpleasant feeling. Such situation is not conducive for driving as he/she needs to stay alert and be positive to drive. Detailed description of the RtSD dataset as well as the experimental set up and discussion will be provided in the Appendix AIII.

6.1 VA Approach based on Affective Space Model Understanding

The general architecture of the proposed speech emotion recognition system is described in Figure 3.1. It comprised of three modules; namely: pre-processing, feature extraction and classification. This proposed generalized approach is refined in Chapter 5 using the discrete-class classification system.

In the discrete-class approach, the classifier consists of four networks to identify the three basic emotions; of: anger, happiness and sadness, as well as neutral (serving as emotionless state). The output is derived from the maximum value of the respective emotion network. Detailed description of the approach has been presented in Chapter 5, see Figure 5.1. The two-stage cultural detection coupled with the discrete-class speech emotion recognition approach is only suitable for supervised learning where the target class is known a priori. The dynamic changes in speech emotion cannot be considered in such approach thus limits the analysis. Hence, the VA analysis is proposed from the psychologists' understanding that emotion can be represented using Valence (V) and Arousal (A) values in an Affective Space Model. Figure 6.1 illustrates the block diagram of the two different classification approaches adopted in this thesis; namely: discrete-class and VA analysis classification approach.

Figure 6.1 Block Diagram of Different Classification Approach

A two-dimensional data-driven affective space model will be generated based on these two outputs and a particular emotion can be discriminated using its axis location in the generated affective state space model. *Valence* refers to the impact of the emotion to oneself ranging from a positive to a negative effect and is denoted as the *x-axis*. In addition, *Arousal* ranges from calm to excited in order to describe the emotion activation and is represented in the *y-axis*. Figure 6.2 depicts the block diagram of the transformation process of speech emotion signal to Valence and Arousal values to be used in the VA approach. The Valence and Arousal values for these emotions are transformed from the MFCC features vector and classifiers. Details of the MFCC feature extraction method is described in Chapter 3; see Section 3.2.1. Subsequently the affective space model is generated from the V and A outputs, as shown by the block diagram of Figure 6.2. Neutral is then recalibrated to derive the new origin for the affective space model, thus generating the emotion quadrant of the affective space model. Figure 6.3 depicts the generic quadrant labelling and the emotions represented within the quadrant of the generated affective space model. The dashed line illustrating the boundary of the

generated affective space model. Detailed explanation of the convergence of the valence and arousal values from MFCC vectors are provided next.

Figure 6.2: The Transformation Process of Speech Emotion Signal to VA Analysis

Consider a speech signal after the features had been extracted using the MFCC feature extraction method and the system is to be trained with the respective basic emotions. In this case $(V,A) = (0,0)$ is set for neutral or emotionless state, $(V,A) = (+1,+1)$ is set for happiness, $(V,A) = (-1,+1)$ is set for anger, $(V,A) = (-1,-1)$ is set for sadness and $(V,A) = (+1,-1)$ is set for calm. Figure 6.4 shows the block diagram of how training can be carried out based on the minimum mean square error (MMSE) technique that can transform the MFCC features into Valence and Arousal values respectively.

Figure 6.3 The Quadrant Labelling and Its Occupied Emotion

Figure 6.4 Block Diagram of the VA Approach

Assuming that all the captured speech emotion signals are wide-sense stationary and features from the speech emotion signal were extracted where it is first transformed

to frequency domain using the short time Fourier transform (STFT). Thus the input speech emotion signals can be represented by Equation (6.1):

$$\mathbf{Y}(k) = H(k_1) \cdot X(k_1) + H(k_2) \cdot X(k_2) + \dots + H(k_n) \cdot X(k_n) \quad (6.1)$$

Where $X(k_n)$ is the transformed original speech signal into features (in this case, the MFCC features), and

$H(k_n)$ is the pre-processing filter.

In Figure 6.4, the microphone used is assumed linear and close enough to the speaker in such a way that the environmental noise can be considered as negligible. Two linear discrete filters (or Wiener filters) are then tuned to minimize the outputs error ($eV(t)$ or $eA(t)$) for either the Valence (V) or Arousal (A) output respectively. These training of the Wiener filters for the V or A outputs enables the Wiener filters to learn from the reference speech emotion database that are previously collected. From Figure 6.4, the output error will learn from the reference speech emotion database where switch SW_V and SW_A will be switched in accordance with the respective emotion based on the psychologists' emotion understanding. Thus, the following VA switch values are set; namely:

switch (SW_V, SW_A) = (0, 0)	for neutral serving as emotionless state,
switch (SW_V, SW_A) = (+1, +1)	for happiness emotion,
switch (SW_V, SW_A) = (-1, +1)	for anger emotion,
switch (SW_V, SW_A) = (-1, -1)	for sadness emotion, and
switch (SW_V, SW_A) = (+1, -1)	for calm emotion.

These switch values are selected to represent the four different quadrants in the affective space model with neutral centroid acting as the origin segregating the quadrants as illustrated in Figure 6.3.

The valence (V) output error signal of Figure 6.4 is defined as Equation (6.2)

$$\begin{aligned} e_v(k) &= V'(k) - \mathbf{W}_v(k) \mathbf{Y}(k) \\ e_v^2(k) &= V'(k)^2 - 2V'(k) \mathbf{W}_v(k)^T \mathbf{Y}(k) + \mathbf{W}_v(k)^T \mathbf{Y}(k) \mathbf{Y}(k)^T \mathbf{W}_v(k) \end{aligned} \quad (6.2)$$

Where $V'(k) = \{-1, 0, +1\}$ is the expected valence output to be trained based on the input reference emotion,

$\mathbf{Y}(k) = f_1, f_2, \dots, f_n$ is the input vector to the Wiener filter. The $f_1, f_2, \dots, f_n$ are the features after the extraction process in frequency domain,

$\mathbf{W}_v(k) = [W_v(k-n-1), W_v(k-n-2), \dots, W_v(k)]$ is the Wiener filter weights (taps) at sampling time k of tap length n and

It is assumed that the signal $\mathbf{Y}(k)$ are wide-sense stationary.

The expected value in Equation (6.3) becomes:

$$E[e_v^2(k)] = E[V'(k)] - 2E[V'(k) \mathbf{Y}(k)] \mathbf{W}_v(k)^T + \mathbf{W}_v(k)^T E[\mathbf{Y}(k) \mathbf{Y}(k)^T] \mathbf{W}_v(k) \quad (6.3)$$

Note that the expected value of any sum is the sum of the expected values, but the expected value of a product is not necessarily the product of the expected values unless the variables are statistically independent.

Consider the square matrix for input $\mathbf{Y}(k)$ such that:

$$\begin{aligned} \mathbf{Q} &= E[\mathbf{Y}(k)\mathbf{Y}(k)^T] \\ &= E \begin{bmatrix} y(k-n-1)^2 & y(k-n-1)y(k-n-2) & L & y(k-n-1)y(k) \\ y(k-n-2)y(k-n-1) & y(k-n-2)^2 & L & y(k-n-2)y(k) \\ M & M & & M \\ y(k)y(k-n-1) & y(k)y(k-n-2) & L & y(k)^2 \end{bmatrix} \end{aligned} \quad (6.4)$$

The Mean Square Error (MSE) component of Equation (6.4) is also known as the input correlation matrix. The main diagonal terms are the mean squares of the input components and the cross terms are the respective cross correlations among the input components. Similarly a cross correlation matrix between the scalar input $V'(k)$ and the vector input $Y(k)$ is given as Equation 6.5:

$$\begin{aligned} \mathbf{P} &= E[V'(k)\mathbf{Y}(k)] \\ &= E[V'(k)y(k-n-1) \quad V'(k)y(k-n-2) \quad L \quad V'(k)y(k)]^T \end{aligned} \quad (6.5)$$

Element of $\mathbf{P}$ is all constant second order statistics when $V'(k)$ and $Y(k)$ are stationary. The mean square error components of Equation (6.5) can now be defined as Equation (6.6).

$$\text{MSE} = \varepsilon = E[e_v^2(k)] = E[V'^2(k)] + \mathbf{W}_v(k)^T \mathbf{Q} \mathbf{W}_v(k) - 2\mathbf{P}^T \mathbf{W}_v(k) \quad (6.6)$$

Clearly from the above expression the mean square error ε is precisely a quadratic function of the components of weight vector $\mathbf{W}_v(k)$ when both the input components are stationary stochastic variables. This implies the elements of $\mathbf{W}_v(k)$ will only appear as first and second degree when Equation (6.7) is expanded. Let consider the gradient of the MSE given by Equation (6.7) by differentiating with respect to $\mathbf{W}_v(k)$ such that:

$$\nabla = \frac{\partial \varepsilon}{\partial \mathbf{W}_v(k)} = 2\mathbf{Q}\mathbf{W}_v(k) - 2\mathbf{P} \quad (6.7)$$

Taking the minimum mean-square error of Equation (6.7), the weight vector $\mathbf{W}_V(k)$ is set at its optimal value where the gradient is zero, such that:

$$\nabla = 0 = 2\mathbf{Q}\mathbf{W}_{V(\text{opt})} - 2\mathbf{P} \quad (6.8)$$

$\mathbf{W}_{V(\text{opt})}$ is sometimes called the Weiner weight vector and is given by:

$$\mathbf{W}_{V(\text{opt})} = \mathbf{Q}^{-1}\mathbf{P} \quad (6.9)$$

provided $\mathbf{Q}$ is nonsingular. The above expression is the well-known Weiner-Hopf equation in matrix form. Thus the minimum error of Equation (6.8) now becomes:

$$\begin{aligned} \mathcal{E}_{\min} &= E[V'^2(k)] + \mathbf{W}_{V(\text{opt})}^T \mathbf{Q} \mathbf{W}_{V(\text{opt})} - 2\mathbf{P}^T \mathbf{W}_{V(\text{opt})} \\ &= E[V'^2(k)] + [\mathbf{Q}^{-1}\mathbf{P}]^T \mathbf{Q} \mathbf{Q}^{-1}\mathbf{P} - 2\mathbf{P}^T \mathbf{Q}^{-1}\mathbf{P} \\ &= E[V'^2(k)] + \mathbf{P}^T \mathbf{Q}^{-1}\mathbf{P} - 2\mathbf{P}^T \mathbf{Q}^{-1}\mathbf{P} \\ &= E[V'^2(k)] - \mathbf{P}^T \mathbf{Q}^{-1}\mathbf{P} \end{aligned} \quad (6.10)$$

Thus it can be seen from Equations (6.4), (6.5), (6.9) and (6.10) that the Weiner filter can be trained to tune to the expected valence (V') at time k . Similarly with the Arousal (A) Weiner filter $\mathbf{W}_A(k)$ it will be used to train the Arousal portion of the VA analysis.

Testing the input speech emotion signals can be easily carried out by first having the input speech via the microphone be preprocessed and features extracted. Once this is accomplished, the output will be tapped from the V and A output as shown by Figure 6.4. Since the Weiner filters were already tuned to the respective Valence (V) and Arousal (A) of the basic emotions, the output V and A will produce the respective Valence and Arousal of the basic emotions. Any classifiers that allow learning to replace the Linear Discrete Filter $\mathbf{W}_V(k)$ and $\mathbf{W}_A(k)$ should be able to produce the V and A output respectively. Based on this notion, the selected three classifiers of MLP, ANFIS and GenSoFNN can replace the digital filters hence are able to derive the V and A values.

Based on this approach, a good approximation of an emotion region can then be generated. This would allow not only the trained basic emotions to be detected but also the other emotions that are excluded in training. In this case, the calm emotion can be detected using the VA approach although the classifier is only trained using anger, happiness and sadness as well as neutral acting as the emotionless state. Hence, this approach enables the of unknown emotion territory as well as providing a visualization tool for the emotion profiles. This method of emotion profiling also allows better understanding of the cultural effect on the respective datasets, which can be further used to analyse and explain the inaccuracy performance of most speech emotion recognition system based on the discrete-class approach. Further explanation of the data-driven affective space model generation will be discussed in the following section.

6.1.1 Generation of the Affective Space Model Origin and Quadrant Based on Neutral Recalibration

Most psychologists agree that neutral is represented by the coordinate of (0,0) in the affective space model. This is because neutral refers to an emotionless state, which has no inclination towards positive or negative value of the Valence (x-axis) and Arousal (y-axis) dimensions. However, to obtain a ‘pure’ neutral data that can be located in the (0,0) coordinate in the affective space model is almost impossible. This is because of other parameters influencing the speech. Culture for instance can introduce biasness to the Valence and Arousal values. Hence, based on the hypothesis that neutral is the base of all emotion and its Valence and Arousal value is supposed to be in (0,0) coordinate, the centroid of neutral in the culturally-influenced speech emotion dataset is proposed to be set as the origin of the affective space model. The four quadrants of the generated

affective space model are then defined as hard boundary derived from the origin as a straight line separating positive and negative values of the Valence and Arousal (V,A) axes.

In order to find the emotion and neutral centroid, k-means clustering method [Jain10, Lloy82, Lloy57] is adopted. K-means clustering treats each instance as a point with its own coordinate. The point coordinate in this case refers to its Valence (x-axis) and Arousal (y-axis) values. K-means determines the centroid by iteratively compute the distance between the points and the cluster centre so as to minimize the within-cluster sum of squares. Thus, it can be said that the centroid of the cluster is the point to which the sum of distances from all points in that cluster is minimized. However, instead of using Euclidean distance as a tool to compute the sum of absolute differences, Manhattan distance technique (also known as the L_1 distance) is employed. L_1 distance is selected for the emotion centroid determination due to its capability to get component-wise median of the points in the particular emotion cluster. This is very important to ensure that the median value of Valence and Arousal of that selected emotion is retained in such a way that it represents the central point of the emotion cluster distribution.

Consider Manhattan distance, d_1 between 2 vectors of p and q in a 2-dimensional affective space model in a fixed Cartesian coordinate system. Centroid location can be calculated using Equation (6.11) where the result is the sum of the lengths of the projections of the line segment between the points onto the coordinate axes.

$$d_1(p,q) = \|p - q\|_1 = \sum_{i=1}^2 |p_i - q_i| \quad (6.11)$$

where $p = (p_1, p_2, \dots, p_n)$ and $q = (q_1, q_2, \dots, q_n)$ are vectors. For instance, the L_1 distance between coordinate (p_1, p_2) and (q_1, q_2) is $|p_1 - q_1| + |p_2 - q_2|$. Figure 6.5 depicts the emotion and its centroid for NTU_American dataset using ANFIS classifier. The x-axis represents Valence and the y-axis represents Arousal. It is noted that neutral distribution is concentrated in the center compared to anger that bias towards Quadrant 1, happiness bias towards Quadrant 2 and sadness bias towards Quadrant 3. This finding is consistent to the affective space model description in Chapter 2. Based on Figure 6.5, it is observed that neutral did not occupy exactly the coordinate (0,0) although the location is approximately near to the origin. Hence, the neutral centroid is proposed as the origin for the generated affective space model in determining the segregation of positive and negative values of the two V and A axes. The neutral centroid is suitable to act as the origin based on the assumption that Valence and Arousal values for neutral are supposed to be (0,0). Such assumption is derived from psychologists' understanding of affective space model [Grim07, Schl54].

Figure 6.5 NTU_American Dataset Emotions and its Centroid using ANFIS Classifier

Figure 6.6(a) and Figure 6.6(b) depict the effect of using coordinate (0,0) and neutral centroid as origin for the quadrant determination using the ANFIS classifier for

the Berlin dataset respectively. It is observed in Figure 6.6(b) that the usage of the neutral centroid as the new origin shifts and recalibrates the centroid of other emotions in the correct quadrant as opposed to the usage of original origin of coordinate (0,0) in Figure 6.6(a). It is noted that happiness is now in Quadrant 1 instead of Quadrant 2 in Figure 6.6(a). Hence, this finding empirically demonstrated the use of neutral centroid is able to recalibrate the origin and is able to cater to the dynamic characteristic of a culturally-influenced speech emotion dataset. Subsequently, a homogeneous speech emotion dataset would have the origin at (0,0) coordinate without the need of neutral recalibration. However, since collecting a speech emotion dataset without a cultural bias is a challenge, recalibration of the neutral is important to ensure cultural bias is considered because the culture effect can shift the origin in the affective space model.

Figure 6.6: Generation of Affective Space Model Origin for Berlin Dataset Discriminated by ANFIS Classifier using Different Origin

6.1.2 Generation of the Affective Space Model Neutral-Emotion Boundary in a Specified Quadrant

Once the origin and the quadrant of the affective space model has been established, the boundary between emotion and neutral has to be determined in each quadrant. This is to ensure that the testing instances belong to the right class of either neutral or the specified emotion in a particular quadrant. It is known that neutral is located basically around the origin, irrespective of the quadrants in the affective space model, as shown by Figure 6.5. Thus, it is not easy to exclusively establish the neutral region. Even to this date, the neutral region is still debatable among psychologists and researchers due to its presence in all emotions. However, in this thesis, the focus is given to segregate emotion and neutral in the quadrant rather than to establish a neutral region per se in the newly neutral recalibrated affective space model.

Figure 6.7: Emotion Segregation Example using Soft-Boundary Approach

Consider instance i of the speech emotion that would be translated using the selected classifier into (x_i, y_i) value where x_i represents the Valence and y_i represents the Arousal values respectively. The centroid (x_e, y_e) for a given set of emotion can be derived using the K-means approach discussed in section 6.1.1, where $e = S$ denotes for sadness, $e = H$ denotes for happiness, $e = A$ denotes for angry and $e = N$ denotes for neutral. The boundary between emotions is defined as hard boundary which is determined

by the centroid of neutral (x_n, y_n) . This neutral centroid is also selected as the the new (recalibrated) origin for the Valence and Arousal (V, A) axis respectively.

The neutral-emotion boundary refers to the soft boundaries segregerating neutral and emotion in its respective quadrant. Consider the points in Figure 6.7 where the testing instance i distance to emotion centroid $d_{(i,e)}$ or neutral centroid $d_{(i,n)}$ can be computed by Equation (6.12) of Euclidean distance calculation. It is selected to get the nearest distance from one point to the neutral and emotion centroid.

$$\begin{aligned} d_{(i,e)} &= \sqrt{(x_i - x_e)^2 + (y_i - y_e)^2} \\ d_{(i,n)} &= \sqrt{(x_i - x_n)^2 + (y_i - y_n)^2} \end{aligned} \quad (6.12)$$

where x_i and y_i are the Valence and Arousal values derived from the classifier outputs,

x_n and y_n are the Valence and Arousal values of the neutral centroid, and

x_e and y_e are the valence and arousal values of the respective emotion centroid.

Based on Figure 6.7, the distance from point i to the newly recalibrated origin/neutral centroid (x_n, y_n) is denotes as $d_{(i,n)}$, the distance from point i to the anger emotion centroid (x_a, y_a) in Quadrant 1 is denoted as $d_{(i,a)}$ and the distance from point i to the happiness centroid (x_h, y_h) in Quadrant 2 is denotes as $d_{(i,h)}$. The value of these respective distances are assumed positive and it can be ascendingly arranged that $d_{(i,h)} < d_{(i,a)} < d_{(i,n)}$. Based on the scenario illustrated in Figure 6.7, the instance i will never be considered as happiness although the distance is the nearest compared to distance to anger centroid, $d_{(i,a)}$ and neutral centroid $d_{(i,n)}$. This is because the instance i is located in Quadrant 1. Hence, the possible class for instance i is either anger or neutral. Anger or neutral need to be discriminated since an instance can only be referred to one emotion.

Thus for any instance i of the affective space model, the emotion class determination of any emotion can be represented by the specified emotion quadrant or bounded by the neutral boundary. The soft-boundary of neutral to emotion can thus be developed. Such understanding is based on the assumption that the testing instant belongs to the class with nearest distance of either emotion centroid or neutral centroid in its respective emotion quadrant. Thus the instance i is more likely to be a member of a particular class if it has the nearest distance between the neutral and respective emotion centroid as depicted by Equation (6.13), where the boundary between neutral and the respective emotion can be derived.

$$B = \frac{d_{i,e}}{d_{i,N}} = \begin{cases} > 1 & \text{instance } i \text{ is neutral} \\ = 1 & \text{this is the boundary and instance } i \text{ is taken as neutral} \\ < 1 & \text{instance } i \text{ is the respective emotion} \end{cases} \quad (6.13)$$

From Equation (6.13), for $B \geq 1$ means that the VA value of the instance is nearer towards neutral centroid, thus the instance is classified as neutral. On the other hand, for $B < 1$ refers to instance that is in favor of respective emotion class as it is nearer to the centroid of that emotion. It is important to bear in mind that this neutral-emotion boundary (B) is only applicable for the respective emotion of a particular quadrant where instance i is located.

Figure 6.8 shows the NTU_American dataset with the neutral-emotion boundary for each quadrant with exception of Quadrant 4. This is because the calm centroid is not available since emotion calm is not included in the speech data for analysis. The soft boundary approach allows different size of neutral centroid to be used for segregating different emotion in the different quadrant. However, for simplification, neutral region in Quadrant 4 is assumed to have similar value as Valence of Quadrant 2 and Arousal of

Quadrant 3. Red dashed line in Figure 6.8 shows the projection of neutral boundary for calm region in Quadrant 4.

Figure 6.8 NTU_American Dataset with Neutral-Emotion Boundary Using ANFIS

6.1.3 Experimental Set-Up and Data Structures for VA Approach

To ensure consistency and repeatability of the experiment, the training and testing data must first be normalized and the number of data used for each basic emotion must also be consistent. In this study, the number of instances for the emotion classes is limited to 1000 instances to ensure that each class has a fair chance of training and the classifier is not bias towards any class that has a larger number of instances. To determine the global accuracy, a 5-fold cross-validation technique is implemented so that 80% of the data are used for training and the remaining 20% are used for testing. These training-testing data pairs are iteratively used until all pairs have been trained and tested. The overall accuracy is obtained from the mean accuracy of each training-testing experimental result. For speech emotion recognition, the dataset is divided into two types of data pool; namely: homogeneous data and heterogeneous data. Different homogeneity experiments are

conducted to investigate cultural impact on the proposed system performance based on the different cultural dependency.

6.1.4 Speech Emotion Identification Result and Discussion using the Proposed VA Approach

Based on the discussion in Chapter 5, although it is clear that the discrete-class classification approach is able to segregate the different emotion, the VA approach is introduced to allow a detailed performance analysis for the proposed system. Based on the hypothesis that emotion can be represented in a 2D manner using the affective space model, the classifiers; namely: MLP, ANFIS and GenSoFNN are trained to reflect the Valence and Arousal values of the tested instances. Each Valence and Arousal value is mapped onto the x and y axes of the affective space model so that identification of an emotion for a particular instance can be made based on its location within the neutral recalibrated affective space model.

Figure 6.9 and Figure 6.10 illustrate the location of different centroids derived from different homogeneous and heterogeneous data pool using the different classifiers. It is consistently observed that the anger centroids are biased towards Quadrant 1, happiness centroids are biased towards Quadrant 2 and sadness centroids are biased towards Quadrant 3. In addition, it is interesting to observe that the neutral recalibration of the affective space model can represent the cultural effect of that particular culturally-influenced speech emotion dataset. Based on this observation, it is clear that neutral cannot be regarded as inconsequential in speech emotion recognition task as neutral can be used as an indicator of culture influence based on its centroids biasness.

Figure 6.9 shows the different emotions centroids for homogeneous intra-cultural experiment using the NTU_American, Berlin and NTU_Asian dataset respectively using three different classifiers; namely: MLP, ANFIS and GenSoFNN are presented. The blue triangle in Figure 6.9 represents the neutral origin hence marks the origin of the recalibrated affective space model. The red diamond represents anger centroid, the orange square represents the centroid of happiness and the green circle represents the centroid of sadness. It is noted that the neutral centroids for all different datasets using various classifier yielded consistent result with minimal variance except for Berlin dataset using the GenSoFNN classifier, see Figure 6.9(f). In the NTU_American dataset, neutral centroid is distributed in the middle of the generated affective space model, which is near to (0,0) coordinate. Conversely, the Berlin dataset neutral centroid is more biased towards Quadrant 1 that favours the anger emotion and the NTU_Asian origin tended to be in Quadrant 3 that representing the sadness emotion. Based on this observation, it can be summarized that although neutral is regarded as emotionless state and hypothesized to be the same across all dataset it seems to vary slightly and culture does influence the neutral distribution. The result in Figure 6.9 also empirically verified that coordinate (0,0) might not be the best origin for a particular dataset, hence justified the use of the neutral centroid as the new origin for affective space model.

Based on the result of Figure 6.9, the different emotion quadrants size in the newly recalibrated affective space model can also be observed. The red rectangular background represents Quadrant 1 (anger), the yellow rectangular background represents Quadrant 2 (happiness), the green rectangular background represents Quadrant 3 (sadness) and the blue rectangular background represents Quadrant 4 (calm). The small

blue triangle in Figure 6.9 represents the origin of the respective culturally-influenced datasets.

Figure 6.9: The Different Emotion Centroid of Homogeneous Intra-Cultural Experiment

It can be observed that the size of all quadrants in NTU_American dataset is almost equal for the four quadrants. This result can be correlated to the experimental result in Chapter 5 that NTU_American dataset always yielded the best accuracy

compared to Berlin and NTU_Asian dataset. The Berlin experimental result in Chapter 5 is always biased towards sadness. This situation can be explained since Berlin neutral centroid is located in Quadrant 1. Such scenario results in the change of sizes for Quadrant 3 (sadness) and 4 (calm) to be bigger than the sizes of Quadrant 1 (anger) and 2 (happiness). Since calm emotion is not included as a training class in discrete-class approach, the result is biased towards sadness. Consistent deduction is also applicable for the NTU_Asian dataset. The experimental result for the homogeneous experiment using the NTU_Asian dataset reported in Chapter 5 shows that happiness is highly recognizable when compared against the other emotions. As depicted in Figure 6.9(g-i), the size of Quadrant 2 (happiness) is bigger than the other quadrant. Hence, based on the result in Figure 6.9, it is noted that the VA approach is able to complement the discrete-class approach by providing an in-depth understanding of the result. The VA approach is able to offer further explanation of the discrete-class approach, which facilitates a better understanding of speech emotion. Table 6.1 summarizes the different quadrant bias yielded using recalibrated neutral affective space model.

Table 6.1. Summary for Different Quadrant Bias Based on Neutral Centroid

Valence value	Arousal value	Location of neutral centroid	Quadrant bias	Emotion bias
Positive (+)	Positive (+)	Quadrant 2	Quadrant 3	Sadness
Positive (+)	Negative (-)	Quadrant 4	Quadrant 1	Anger
Negative (-)	Positive (+)	Quadrant 1	Quadrant 4	Calm
Negative (-)	Negative (-)	Quadrant 3	Quadrant 2	Happiness

From Table 6.1, if the neutral centroid has positive Valence and Arousal values, it will be located in Quadrant 2. Based on this, the size of the four quadrants can be re-arranged in

such a way that: size of Quadrant 2 < size of Quadrant 1 < size of Quadrant 4 < Size of Quadrant 3. Hence, the emotion identified will bias towards sadness since Quadrant 3 is the largest. Consider the situation when the neutral centroid has positive value of Valence and negative value of Arousal. The origin of the recalibrated affective space model will be located at Quadrant 4 representing calm emotion. The size of the quadrants can be arranged such that: size of Quadrant 4 < size of Quadrant 3 < size of Quadrant 2 < size of Quadrant 1. Based on this situation, the emotion identified will be biased towards anger emotion, which has negative Valence and positive Arousal value. The third situation to consider is when neutral centroid has negative Valence value and positive Arousal value. The origin of the newly recalibrated affective space model will be located in Quadrant 1, which is associated with anger emotion. This will result to: the size of Quadrant 1 < size of Quadrant 2 < size of Quadrant 3 < size of Quadrant 4. The emotion identified will be biased towards calm emotion since the size of Quadrant 4 is more than the other three quadrants. Subsequently, if the neutral centroid has negative value for both Valence and Arousal, the origin of the recalibrated affective space model will falls in Quadrant 3. The relationship of the different sizes of Quadrant can be written as: the size of Quadrant 3 < size of Quadrant 4 < size of Quadrant 1 < size of Quadrant 2. Hence, the emotion identified will be biased towards happiness.

Further analysis is conducted using the heterogeneous intra-culture experiment where two or three culturally-influenced datasets are combined as a data pool for training and testing. The rationale of such an approach is to suppress the influence of a particular culture and enhance the classifier's ability to be more general. The higher the number of culturally-influenced data added to the mixture, the higher the non-linear elements need to be handled by the classifier resulting in reduced accuracy.

Figure 6.10 illustrated the different emotion centroids derived from the heterogeneous datasets. The small red diamond in Quadrant 1 represents anger centroid, the small yellow square in Quadrant 2 represents happiness centroid, the small green circle in Quadrant 3 represents sadness centroid and the small blue triangle represents neutral centroid or the origin of the newly recalibrated affective space model. In addition, the red rectangular background represents anger region (Quadrant 1), the yellow rectangular background represents happiness region (Quadrant 2), the green rectangular background represents sadness region (Quadrant 3) and the blue rectangular background rectangle calm region (Quadrant 4) of the affective space model.

It is observed in Figure 6.10 that the centroid of neutral for the heterogeneous intra-cultural experiment seems to fall in between the centroid of neutral for homogeneous intra-cultural outcome of the respective dataset, see Figure 6.9. The heterogeneous neutral centroid seemed to be closer to the (0,0) coordinate in comparison to the homogeneous neutral centroid. It is also noted that MLP and ANFIS yielded consistent neutral centroid with minimal distance difference compared to the GenSoFNN results as in Figure 6.10(c), Figure 6.10(f), Figure 6.10(i) and Figure 6.10(l). Hence, it can be said that the combination of the culturally-influenced dataset is able to minimize the inter-cultural effect. This scenario caused the neutral centroid to be closer to the origin (0,0) coordinate of the recalibrated affective space model.

Figure 6.10: The Different Emotion Centroids in Heterogeneous Intra-Cultural Experiment Using Different Classifiers

From the above Figure 6.10, it can be observed the way the origin of the data-driven affective space model is recalibrated using the neutral centroid determination based on Manhattan distance. This is to ensure that the neutral centroid is a component-wise median of the neutral points in the neutral emotion cluster. Once the affective space model has been recalibrated, the different classes of the testing instances can be segregated based on its Valence and Arousal values. These values will be mapped onto the newly generated affective space model and be discriminated based on its quadrant. To differentiate one instance from the emotion of its respective quadrant; namely: anger for Quadrant 1, happiness for Quadrant 2, sadness for Quadrant 3 and calm for Quadrant 4 with neutral cluster, the Euclidean distance of the instance to the origin and the emotion centroid in the quadrant it is located will be calculated, see Section 6.1.2. The accuracy can then be computed using Equation (6.14).

$$Accuracy = \frac{\text{Number of true positive}}{\text{Total number of emotion instances} - \text{neutral instances}} \quad (6.14)$$

Figure 6.11 shows the homogeneous intra-cultural speech emotion identification result using different classifiers of MLP, ANFIS and GenSoFNN based on the VA approach. From the result in Figure 6.11, ANFIS performed marginally better with 71.7% mean accuracy when compared against the MLP (69.1%) and GenSoFNN (68.1%) performance. All three classifiers yielded similar result pattern for all the culturally-influenced datasets with exception of NTU_Asian dataset. This result is consistent with experimental result of discrete-class approach presented in Chapter 5.

Figure 6.11: The Performance for Homogeneous Intra-Cultural Experiment based on the VA Analysis

The result for NTU_Asian trained with ANFIS favours sadness with 75.3% accuracy as opposed to happiness by MLP (69.0%) and anger by GenSoFNN (77.3%). The result pattern of the Berlin dataset trained with all three classifiers are consistent with sadness obtained the highest performance with mean accuracy of 74.6% followed by anger (63.5%) and happiness (57.3%). The classifier results trained using NTU_American dataset also yielded similar pattern as the experimental result trained with the Berlin dataset. Sadness recorded the highest mean accuracy of 80.8% followed by anger and happiness with differences of 9.4% and 15.9% respectively.

It is noted that the best emotion identification result is produced when attempting to recognize sadness in the NTU_American dataset using the GenSoFNN classifier with an accuracy of 84.7% and the worst emotion identification result is produced by happiness identification in the Berlin dataset using the GenSoFNN classifier with an accuracy of 56.2%. This result illustrates the MLP capability in giving consistent result to differentiate the emotion with smallest mean variance of 2.2% as compared against ANFIS and GenSoFNN with 4.0% and 4.5% respectively. Such situation may result from minimal neutral recalibration in the affective space model; see Figure 6.9(b), Figure 6.9(e) and Figure 6.9(h). The NTU_American dataset leads the rank of mean accuracy with 72.4% and followed by NTU_Asian and Berlin dataset with degradation of 1.1% and 7.3% accuracy respectively. Such small variance in the result is because the affective space model is now tuned to the cultural biasness through neutral recalibration.

The heterogeneous intra-cultural speech emotion identification is also conducted to allow different culturally-influenced datasets to be mixed. Figure 6.12 shows the accuracy of the heterogeneous intra-cultural speech emotion identification using different classifiers. The result is consistent with the homogeneous intra-cultural speech emotion identification experiment although it is clear that heterogeneous experiment yielded lower accuracy than the homogeneous intra-cultural experimental result. This is because the mixture of the different culturally-influenced datasets in the training and testing data pool can have negative correlation from some features of differing culture.

Figure 6.12 The Accuracy for Heterogeneous Intra-Cultural Experiment based on VA Analysis

Experimental result in Figure 6.12 shows that the ANFIS obtained marginally better performance with a mean accuracy of 66.5% followed by MLP and GenSoFNN with performance difference of 0.2% and 1.5% respectively. It is observed that ANFIS and MLP classifiers share similar performance pattern of all heterogeneous intra-cultural speech emotion identification with sadness producing the highest accurate emotion that can be segregated with ratio between 71.7% and 78.1%. Anger and happiness trained using MLP and ANFIS obtained almost similar performance with mean accuracy of 62.6% and 62.3% respectively with only 0.3% variation. GenSoFNN however yielded a different pattern of accuracy for the heterogeneous intra-culture experiments with sadness

recorded the highest mean accuracy of 67.8% followed with 66.6% and 60.7% for happiness and anger respectively. This scenario may result from the different neutral centroid location as depicted in Figure 6.10 earlier. From the result in Figure 6.12, it is interesting to note that all classifier obtained almost equal overall accuracy. This result has empirically demonstrated that no classifier can claim superiority among others. This result also illustrated the ability of the newly data-driven recalibrated neutral affective space model to cater the culture influence of the speech emotion dataset.

Subsequently, the result also confirms the hypothesis that higher number of culturally-influenced data added in the mixed data pool may affect the accuracy of the proposed system. From the result in Figure 6.12, it is observed that the 2 different cultures mixture is better than 3 different cultures mixture with accuracy difference ratio between 2.7% to 5.0% for ANFIS, 0.5% to 3.4% for MLP and 1.2% to 8.0% for the GenSoFNN classifier (with exception of Berlin+NTU_Asian dataset). ANFIS recorded the most consistent result for mean variance of different emotion with 1.6% followed by 2.1% and 3.2% for MLP and GenSoFNN respectively. The results indicate that there are certain similar acoustic characteristics that are shared within the same culture. However, once the fusion of different culturally-influenced mixture dataset is formed, the distinction between emotions becomes unclear. This notion is empirically supported by the result of mean overall accuracy for 2 cultures combined, which seems to be slightly better than the mean overall accuracy for 3 cultures combined. This result can be correlated to the initial assumption that a higher number of cultures combined in a mixed dataset may result in higher non-linear elements thus complicates the classifier to correctly identify the emotion. Hence it can be concluded that culture does influence the performance of the proposed speech emotion recognition system.

6.1.5 Summary of VA Approach Applied on Intra-cultural and Inter-cultural Speech Emotion Recognition Task

Experimental result for the VA approach in section 6.1.4 shows that it is feasible to identify emotion regardless of the culturally-influenced dataset used for both homogeneous and heterogeneous intra-cultural experiments. The homogeneous intra-cultural experiment refers to the segregation of the emotion class from one culturally-influenced dataset. The system is required to recognize the three emotions; namely: anger, happiness and sadness from a single cultural bias. Subsequently, the heterogeneous intra-cultural experiment addresses the task of segregating the emotion class from a mixture of two or three different culturally-influenced speech emotion data in the training and testing data pool. The system has to generalize over a greater number of cultural biases, which introduces a higher probability that some features may have negative correlation due to culture influence. Such situation resulted in lower accuracy when compared against the homogeneous intra-cultural experiment result.

The implementation of the VA approach also provides more in-depth understanding of the result obtained in the discrete-class approach presented in Chapter 5. The VA approach is inspired from the affective space model understanding indicating that emotion can be represented using two-dimensional axes of valence and arousal. Valence refers to the degree of the particular emotion effect ranging from positive to negative. On the other hands, Arousal applies to the degree of activation ranging from excited to calm. These two axes can generate an affective space model, which can be used to explain the emotion characteristics in a more comprehensive manner. Neutral is found to be the foundation of the generation of the affective space model. Such notion is based on psychologists' agreement that neutral does not bias to either positive or negative value of

the valence and arousal axis. It is supposed to be located in the (0,0) coordinate of the affective space model. However, due to the effect of cultural influence, the neutral centroid is shifted according to the cultural bias. Hence, it is deemed appropriate to recalibrate the origin by re-assigning neutral as the new origin of the affective space model.

Figure 6.13 The Summary Accuracy for VA Approach Experiment

Figure 6.13 depicts the summary for the VA approach performance using different culturally-influenced datasets coupled with different classifiers. It can be observed that the disparity of the accuracy recorded for the homogeneous and heterogeneous experiment empirically demonstrates a far better performance as compared to that of the discrete class classification method shown by Figure 5.9, where the differences between heterogeneous and homogeneous can be greater than 30%. There are common acoustical characteristics in the emotion that are universally accepted and yet there are unique attributes that are only shared by similar culture. The complexity of the task that the classifier needs to handle will be increased as more culturally-influenced datasets are introduced to the data pool for the heterogeneous intra-culture experiment. Such situation may result in accuracy degradation.

Based on the result in Figure 6.13, the NTU_American produced the best homogeneous intra-cultural experiment performance with mean accuracy of 72.4% followed with 71.3% for NTU_Asian and 65.1% for Berlin dataset. The accuracy for heterogeneous with two cultures mixture is ranging between 63.8% (Berlin+NTU_Asian) and 69.8% (NTU_American+NTU_Asian). The lowest mean accuracy of 63.7% is recorded by the three culturally-influenced datasets mixture of NTU_American+Berlin+NTU_Asian. The ANFIS produced slightly higher accuracy of 68.7% when compared against MLP (67.5%) and GenSoFNN (66.3%). The worst performance is recorded when attempting to recognize Berlin+NTU_Asian heterogeneous experiment using GenSoFNN classifier (60.6%) and on the contrary, the best performance is produced when recognizing emotion in homogeneous intra-cultural experiment using NTU_American dataset coupled with ANFIS classifier (75.6%). It is interesting to observe that the performances of the three classifiers in the three cultures mixture of heterogeneous intra-cultural experiment are almost identical. Such situations are the result of the recalibrated origin of the affective space model that explicitly caters the culture influence of the speech emotion data.

Subsequently, an extension of the VA approach to re-profile the different driver behaviour state (DBS) based on the understanding of the newly recalibrated affective space model is conducted by using the Real-time Speech Driving (RtSD) dataset. The detailed analysis will be presented in Appendix AIII.

CHAPTER 7

SUMMARY, CONCLUSION AND FUTURE WORK

Many researchers of multi-disciplinary fields; namely: engineering, psychology, sociology, neuroscience, computer science, linguistic and many others combine their efforts in trying to understand the intricate concept of human emotion. To date, mixed findings were reported justifying the need for more intensive and well coordinated collaboration between researchers from these fields to facilitate better understanding of emotion. This thesis has presented different approaches in identifying and categorizing emotions ~ from a simple discrete-class approach to a more complex VA approach. Both approaches are complementary to each other and show potential to recognize, understand and analyze speech emotion.

Based on the work conducted, there are three main contributions of this thesis to the speech emotion community. These contributions are:

Introduction of new method to to analyze speech emotion in more detail from two different perspectives; namely: emotion as a discrete-class and emotion represented in affective space model.

Emotion is a complex psychophysical process that arises spontaneously and involuntarily without explicit conscious effort. It plays a central role in displaying and regulating patterns of behavior that is realized in many human activities. It is propagated through speech towards the convergence and optimization of the interactive process. Hence, understanding speech emotion is the key to boost the effectiveness of the human computer interaction (HCI). The first main contribution of this thesis is the introduction

of new method to enable researchers, engineers, psychologists, medical practitioners as well as an intelligent system developers to analyze speech emotion in more detail from two different perspectives; namely: emotion as a discrete-class and emotion represented in the affective space model. Such methods are able to encompass culture influence understanding that is currently missing from most speech emotion researches.

The discrete-class speech emotion recognition system allows speech emotion to be segregated provided that the training speech emotion class is known prior to the testing exercise (supervised learning). Such system is particularly effective for a limited number of emotion to be identified. To date, most of the state-of-the-art speech emotion recognition systems are employing such method. However, culture influences the performance of the speech emotion recognition system. Most researchers in the speech emotion field tend to ignore culture due to its complexity. In this thesis, a two-stage culture-sensitive speech emotion recognition approach is introduced to utilize the intra-cultural performance of such system. It is based on the hypothesis that if the culture of a particular speech emotion instance is known, emotion can be identified based on the trained network using similar culturally-influenced speech emotion data. Such understanding is derived from the observation that the accuracy of intra-cultural speech emotion recognition system can always provide better accuracy than that of the inter-cultural speech emotion recognition system. Experimental result shows potential of such approach to be incorporated in real applications.

Subsequently, a Valence-Arousal (VA) approach is introduced to complement the discrete-class approach. The VA approach is derived to overcome the drawback of the discrete-class approach that the analysis from this approach is limited to the trained class so that further analysis may not be possible. The VA approach is inspired from the

psychologists' hypothesis based on the affective space model understanding that emotion can be represented in two different axes of the Valence (V) and Arousal (A). These VA axes can be illustrated graphically in such a way that similarity and dissimilarity measurement of the emotion pattern can be expressed and visualized as Euclidean distance between points in this space. The difference values of such axes can be used to automatically segregate one emotion to another.

Based on the VA approach, the shift of neutral centroid is found to provide an insight to the characteristic of a particular culture and their bias towards certain emotion. Neutral is always regarded as inconsequential and trivial to speech emotion researchers as it represents emotionless state. Psychologists hypothesized that neutral is located in the (0,0) coordinate of the affective space model that has zero values for Valence and Arousal. However, culture influence speech emotion resulting in the shift of the affective space model origin. Recalibration of the affective space model based on the neutral centroid of the trained culturally-influenced speech emotion database provides an intuitive and practical approach. Experimental results in this thesis empirically proved that neutral is indeed very important and contradict the common belief that neutral is insignificant.

The VA approach also allows the projection of emotion that is not available during the training exercise. A soft-boundary segregating emotion and neutral in a specific quadrant can be used to allocate the region of a specific emotion. For instance, calm region in Quadrant 4 of affective space model can be estimated by assuming the boundary of neutral region for Valence is equal to the Happiness Valence (Quadrant 2) and the Arousal is equivalent to the sadness Arousal (Quadrant 3). Based on this soft boundary approach, a correlation between driver behavior state and emotion primitives' values can

be re-profiled and analyzed. The experimental result in Appendix AIII is able to show hidden relationship that cannot be visualized using the discrete-class approach. Hence, it can be summarized that VA approach opens new dimension to speech emotion research as an extension to the discrete class classification approach where detail analysis of the results can be interpreted through visualization.

Empirical demonstration of the cultural effect on speech emotion.

Culture is hypothesized as one of the factor that influenced the performance of speech emotion recognition system. Experimental findings reported that some emotions are universally accepted [Elfe02, Sche01] while other emotions show considerable diversity that are only shared by individual within the community [Dewa08, Jusl03, Bezo83]. Due to the complexity in understanding diverse speech emotion parameters, that characterize the speech emotion, culture effect is always regarded as insignificant and hence most researchers tend to ignore this parameter in developing speech emotion recognition system. Although, there are some attempts by psychologist to understand the intra and inter-cultural speech emotion effect on human conveyence and perception, such analyses are conducted based on ethnography method (qualitative approach) rather than quantitative method. In this thesis, the evidence is provided by quantifying the emotion through the translation of the psychologists' understanding of speech emotion coupled with engineering and neuro-computational method for a better speech emotion recognition system. Hetrogeneous and homogeneous training and testing of the speech emotion recognition system allow a detail analysis and understanding of failure in most speech emotion reognition system. These combinations of different approaches and testing from different fields of expertise are able to facilitate better understanding of speech emotion to the speech emotion community at large. To this day, this work is one of the very few

efforts recorded using computerized analysis in the attempt to understand cultural influence on speech emotion.

Introduction of the two culturally-influenced speech emotion datasets of NTU_American and NTU_Asian and a Real-time Speech Driving dataset.

The first two datasets are used as a yardstick to identify and segregate three basic emotions; namely: anger, happiness, sadness and neutral serving as emotionless state. The third dataset is employed to find correlation between driver behaviour state; namely: distracted, laughed, normal and sleepy to emotion primitives' values of Valence and Arousal. The collection of such data is served as an alternative annotated data for the speech emotion research community to analyze and peruse.

7.1 Future Work and Potential Applications

Recognizing human emotions through speech could have endless amount of application improvement ranging from understanding the caller's emotion in a call center to understanding driver behavior for comfort and safe driving. Figure 7.1 shows the block diagram of a generalized speech emotion system using the VA analysis. The culturally-influenced speech emotion database is required as a reference to create the necessary profile for prospective application. First, the appropriate features need to be extracted and pre-processed so that the relevant VA output can be generated. The centroid of neutral will be calculated to form the origin of the affective space model. Such recalibration is needed to cater the different cultural influence in the culturally-influenced speech emotion database. The emotion primitives' values of Valence and Arousal for the various emotions need to be computed to generate the different profiles for various applications.

This profiling method allows similarity and dissimilarity assessment to be visualized to facilitate in depth analysis of the result as a complementary to the state-of-the-art discrete-class approach.

Figure 7.1 Block Diagram of a Speech Emotion Generalized Application System.

Plotting the change of emotions on the VA plot can provide a dynamic information of the individual resulting in the trajectory of the emotion profile. Such information can facilitate emotion prediction analysis.

7.2 Conclusions

Based on the explanation and discussion provided throughout this thesis, it can be concluded that this thesis has met the research objectives as outlined in Chapter 1 and contributed to the body of research at large. Speech emotion is a worth-while field to be studied in detail as it offers many benefit especially in enhancing human computer interaction. Many potential work can be extended to consolidate the finding reported in this PhD thesis. There is no doubt that the VA analysis approach as well as the intra-

cultural and inter-cultural assessment provided in this thesis can enrich the understanding of speech emotion. It is hoped that such understanding can be applied in future applications with better accuracy and optimally benefited the mankind at large.

REFERENCES

- [Angk08] Angkititrakul, P., Hansen, J. H. L., Choi, S., Creek, T., Hayes, J., Kim, J., Kwak, D., Noecker, L. T. & Phan, A. (2008). UTDive: The Smart Vehicle Project. In K. Takeda, J. H. L. Hansen, H. Erdoğan, & H. Abut (Eds.), *In-Vehicle Corpus and Signal Processing for Driver Behavior*, (pp. 55-67), New York, USA: Springer Science+Business Media.
- [Arno60] Arnold, M. B. (1960). *Emotion and Personality*. New York: Columbia University Press.
- [Aver80] Averill, J.R. (1980). A Constructivist View of Emotion. In R. Plutchik & H. Kellerman (Ed.), *Emotion: Theory, Research and Experience* (Vol. 1, Theories of Emotion, pp. 305-339). New York: Academic Press. [Reprinted in A. G. Halberstadt & S. L. Ellyson (Eds.) (1990). *Social Psychology Readings: A Century of Research* (pp. 143-156) New York: McGraw-Hill.]
- [Bat11] Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Kessous, V. A. L. & Amir, N. (2010). Who-dunnit – Searching for the Most Important Feature Types Signaling Emotion-related User States in Speech, *Computer Speech and Language*, 25(1), 4-28.
- [Batt94] Battiti, R. (1994). Using Mutual Information for Selecting Features in Supervised Neural Net Learning, *IEEE Transactions on Neural Networks*, 5, 537 -550.
- [Beau05] Beaupre, M. G. & Hess, U. (2005). Cross-cultural Emotion Recognition Among Canadian Ethnic Groups. *Journal of Cross-cultural Psychology*, 36(3), 355 – 370.
- [Bech04] Bechara, A. (2004). The Role of Emotion in Decision-Making: Evidence from Neurological Patients with Orbitofrontal Damage. *Brain and Cognition*, 55, 30–40.
- [Bech05] Bechara, A. & Damasio, A. R. (2005). The Somatic Marker Hypothesis: A Neural Theory of Economic Decision Making. *Games and Economic Behavior*, 52(2), 336-372.
- [Bege06] Begeer, S., Rieffe, C., Terwogt, M. M. & Stockmann, L. (2006). Attention to Facial Emotion Expressions in Children with Autism, *Autism*, 10 (1), 137-151.
- [Bell09] Bellet, T., Bailly-Asuni, B., Mayenobe, P. & Banet, A. (2009). A Theoretical and Methodological Framework for Studying and Modeling Drivers' Mental Representations. *Safety Science*, 47(9), 1205 – 1221.

- [Berg08] Berg, B. L. (2008). *Qualitative Research Methods for the Social Sciences*, 7th Edition, Boston: Allyn and Bacon. pp. 3.
- [Bezo83] Bezooijen, R. V., Otto, S. A. & Heenan, T.A. (1983). Recognition of Vocal Expression of Emotion: A Three-Nation Study to Identify Universal Characteristics. *Journal of Cross-Cultural Psychology*, 14(4), 387-406.
- [Bish95] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*, New York, USA: Oxford University Press.
- [Brad94] Bradley, M. M. & Lang, P. J. (1994). Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49-59.
- [Brea03] Breazeal, C. (2003). Emotion and Sociable Humanoid Robots. *International Journal of Human-Computer Studies*, 59(1-2), 119-155.
- [Buck99] Buck, R. (1999). The Biological Affects: A Typology. *Psychological Review*, 106, 301-336.
- [Burk05] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A Database of German Emotional Speech. In: *Proceeding of INTERSPEECH'05*, Lisbon, Portugal, 1517-1520.
- [Call08] Callejas, Z. & Lopez-Cozar, R. (2008). Influence of Contextual Information in Emotion Annotation for Spoken Dialogue Systems. *Speech Communication*, 50 (5), 416 – 433.
- [Camp94] Campos, J. J. (1994, Spring). The New Functionalism in Emotion. *Society for Research in Child Development (SRCD) Newsletter*, 1-14.
- [Cann27] Cannon, W. B. (1927). The James-Lange Theory of Emotion: A Critical Examination and An Alternative Theory. *American Journal of Psychology*, 39, 10-124.
- [Cham83] Chambers, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983). *Graphical Methods for Data Analysis*. Belmont, CA: Wadsworth, 158 – 162.
- [Char98] Charles, D. (1998). Constrained PCA Techniques for the Identification of Common Factors in Data. *Neurocomputing*, 22, 145-156.
- [Chiu94] Chiu, S. L. (1994). Fuzzy Model Identification Based On Cluster Estimation. *Journal of Intelligent and Fuzzy Systems*, 2(3), 267-278.

-
- [Clor07] Clore, G. L. & Huntsinger, J.R. (2007). How Emotions Inform Judgment and Regulate Thought. *Trends in Cognitive Science*, 11(9), 393-399.
- [Corn96] Cornelius, R. R. (1996). *The Science of Emotion: Research and Tradition in the Psychology of Emotion*, Upper Saddle River, NJ: Prentice-Hall.
- [Cowi00] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M. & Schroder, M. (2000). 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time, In: *Proceeding of ISCA Workshop on Speech & Emotion: A Conceptual Framework for Research*, Belfast, Northern Ireland, 19-24.
- [Cowi03] Cowie, R. & Cornelius, R. R. (2003). Describing the Emotional States That Are Expressed in Speech. *Speech Communication-Special Issue on Speech and Emotion*, 40(1-2), 5-32.
- [DaFo09] Da Fonseca, D., Segulier, V., Santos, A., Poinso, F. & Deruelle, C. (2009). Emotion Understanding in Children with ADHD, *Child Psychiatry and Human Development*, 40 (1), 111-121.
- [Dama94] Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Grosset/Putnam.
- [Darw72] Darwin, C. R. (1872). *The Expression of Emotions in Man and Animals*. In: Murray, J. (Ed.): Reprinted: 1965), Chicago, IL: University of Chicago Press.
- [Davi80] Davis, S. B. & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transaction on Acoustic, Speech and Signal Processing*, 28 (4), 357 – 366.
- [Dell96] Dellaert, F., Polzin, T. & Waibel, A. (1996). Recognizing Emotion in Speech Processing. In: *Proceeding of the 4th International Conference on Spoken Language (ICSLP '96)*, 3, Philadelphia, USA, 1970-1973.
- [Demp77] Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1), 1-38.
- [Dern10] Derntl, B., Finkelmeyer, A., Eickhoff, S., Kellermann, T., Falkenberg D. I., Schneider, F. & Habel, U. (2010). Multidimensional Assessment of Empathic Abilities: Neural Correlates and Gender Differences, *Psychoneuroendocrinology*, 35, 67—82.
- [Dewa08] Dewaele, J. M. (2008). The Emotional Weight of I Love You in Multilinguals' Languages. *Journal of Pragmatics*, 40 (10), 1753-1780.
-

- [Duch09] Duchateau, J., Kong, Y. O., Cleuren, L., Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Verhelst, W. & Van Hamme, H. (2009). Developing a Reading Tutor: Design and Evaluation of Dedicated Speech Recognition and Synthesis Modules. *Speech Communication*, 51(10), 985–994.
- [Duda00] Duda, R. O., Hart, P. E. & Stork, D. G. (2000). *Pattern Classification*, 2nd edition, New York, USA: Wiley-Interscience.
- [Ekma69] Ekman, P., Sorenson, E. R. & Friesen, W. V. (1969). Pan-Cultural Elements in Facial Display of Emotions. *Science*, 164(3875), 86-88.
- [Ekma71a] Ekman, P. (1971). Universals and Cultural Differences in Facial Expressions of Emotions. In J. Cole (Ed.), *Nebraska Symposium on Motivation* (pp. 207-283). Lincoln, Nebraska: University of Nebraska Press.
- [Ekma71b] Ekman, P., Friesen, W. V. & Tomkins, S. S. (1971). Facial Affect Scoring Technique - A First Validity Study. *Semiotica*, 3, 37-58.
- [Ekma71c] Ekman, P. & Friesen, W. V. (1971). Constants Across Cultures in the Face and Emotion. *Journal of Personality and Social Psychology*, 17, 124-129.
- [Ekma88] Ekman, P. & Heider, K. G. (1988). The Universality of a Contempt Expression - A Replication. *Motivation and Emotion*, 12(4), 303-308.
- [Ekma07] Ekman, P. (2007). *Emotion Revealed: Recognizing Faces and Feeling to Improve Communication and Emotional Life*, 2nd Edition. New York, USA: Owl Books.
- [ElAy11] El Ayadi, M., Kamel, M. S. & Karray, F. (2011) Survey on Speech Emotion Recognition: Features, Classification Schemes and Databases. *Pattern Recognition*, 44, 572-587.
- [Elfe02] Elfenbein, H. A. & Ambady, N. (2002). On the Universality and Cultural Specificity of Emotion Recognition: A Meta-analysis. *Psychological Bulletin*, 128(2), 203-235.
- [Eske09] Eskenazi, M. (2009). An Overview of Spoken Language Technology for Education. *Speech Communication*, 51(10), pp. 832–844.
- [Fern03] Fernandez, R. & Picard, R. W. (2003). Modeling Driver's Speech under Stress. *Speech Communication*, 40, 145-159.
- [Figu02] Figueiredo, M. A. F. & Jain, A. K. (2002). Unsupervised Learning of Finite Mixture Models, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 24 (3), 381-396

- [Forb11] Forbes-Riley, K. & Litman, D. (2011). Designing and Evaluating a Wizarded Uncertainty-adaptive Spoken Dialogue Tutoring System. *Computer Speech and Language*. 25(1), 105-126.
- [Frij87] Frijda, N. H. (1987). Emotions, Cognitive Structure and Action Tendency. *Cognition and Emotion*, 1(2), 115-143.
- [Frij88] Frijda, N. H. (1988). The Laws of Emotion. *American Psychologist*, 43(5), 349-358.
- [Fu03] Fu, X. & Wang, L. (2003). Data Dimensionality Reduction with Application to Simplifying RBF Network Structure and Improving Classification Performance, *IEEE Transactions on Systems, Man, and Cybernetics – Part B: Cybernetics*, 33 (3), 399 – 409.
- [Ganc05] Ganchev, T., Fakotakis, N. & Kokkinakis, G. (2005). Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task. In: *Proceeding of the 10th International Conference on Speech and Computer (SPECOM 2005)*, 1, Patras, Greece, 191 – 194.
- [Germ08] Germanakos, P., Tsianos, N. Lekkas, Z., Mourlas C. & Samaras, G. (2008). Capturing Essential Intrinsic User Behavior Values for the Design of Comprehensive Web-based Personalized Environments. *Computers in Human Behavior*, 24(4), 1434-1451.
- [Gord89] Gordon, S. L. (1989). The Socialization of Children's Emotions: Emotional Culture, Competence and Exposure. In C. Saarni & P. L. Harris (Eds.), *Children's Understanding of Emotion* (pp. 319-349). Cambridge: Cambridge University Press.
- [Gord90] Gordon, S. L. (1990). Social Structural Effects on Emotion. In T.D. Kemper (Ed). *Research Agendas in the Sociology of Emotion* (pp. 145-179). Albany, NY: State University of New York Press.
- [Goud09] Goudbeek, M., Goldman, J. P. and Scherer, K. R. (2009). Emotion Dimensions and Formant Position, In: *Proceeding of the 10th Annual Conference of the International Speech Communication Association INTERSPEECH 2009*, Brighton, United Kingdom, 1575-1578.
- [Grim07] Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based Evaluation and Estimation of Emotion in Speech. *Speech Communication*, 49 (10-11), 787-800.
- [Guy03] Guyon, I. & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157-1182.

-
- [Hall00] Hall, J. A., Carter, J. & Horgan, T. (2000). Gender Differences in the Nonverbal Communication of Emotion. In A. Fischer (Ed.) *Gender and Emotion: Social Psychological Perspective* (pp. 97 – 117). Paris: Cambridge University Press.
- [Hall04] Hall, J. A. & Matsumoto, D. (2004). Gender Differences in Judgment of Multiple Emotions from Facial Expressions, *Emotion*, 4(2), 201 – 206.
- [Heal05] Healey, J. A. & Picard, R. W. (2005). Detecting Stress during Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2), 156-166.
- [Hear10] Hearing Lost Association of North Carolina (2010, November 28). *Cochlea* [Online]. Available: <http://www.nchearingloss.org/gloss.htm>
- [Hebb49] Hebb, D. O. (1949). The Organization of Behavior: A Neuropsychological Theory, New York: Wiley.
- [Hein08] Heinroth, T. & Buhler, D. (2008). Arrigator – Evaluation of a Speech-based Pedestrian Navigation System. In: *Proceeding of the 4th IET International Conference on Intelligent Environments (IE'08)*, Seattle, USA, 1-4.
- [Herb01] Herbrich, R., Graepel, T. and Campbell, C. (1999). Bayes Point Machines. *The Journal of Machine Learning Research*, 1, 245-279.
- [Hoff10] Hoffmann, H., Kessler, H., Eppel, T., Rukavina, S., Traue, H. C. (2010). Expression Intensity, Gender and Facial Emotion Recognition: Woman Recognize Only Subtle Facial Emotions Better Than Men, *Acta Psychologica*, 135, 278 – 283.
- [Holl92] Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*, Cambridge, England: MIT Press / A Bradford Books.
- [Hu07] Hu, H., Xu, M. & Wu, W. (2007). GMM Super Vector Based SVM with Spectral Features for Speech Emotion Recognition. In: *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, Honolulu, USA, IV-413 – IV-416.
- [Jain10] Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-means. *Pattern Recognition Letters*, 31 (8), 651-666.
- [Jang93] Jang, J. –S. R. (1993). ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Transaction on Systems, Man and Cybernetics*, 23 (3), 665-685.
- [John92] Johnson-Laird, P. N. & Oatley, K. (1992). Basic Emotions, Rationality and Folk-theory. *Cognition and Emotion*, 6, 201-223.
-

- [Jone10] Jones, C. R. G., Pickles, A., Falcato, M., Marsden, A. J. S., Happé, F., Scott, S. K., Sauter, D., Tregay, J., Phillips, R. J., Baird, G., Simonoff, E., & Charman, T. (2010). A Multimodal Approach to Emotion Recognition Ability in Autism Spectrum Disorders. *Journal of Child Psychology and Psychiatry*. Advance Online Publication. doi:10.1111/j.1469-7610.2010.02328.x.
- [Jusl03] Juslin, P. N. & Laukka, P. (2003). Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code? *Psychological Bulletin*, 129(5), 770-814.
- [Kawa04] Kawaguchi, N., Takeda, K. & Itakura, F. (2004). Multimedia Corpus of In-Car Speech Communication. *Journal of VLSI Signal Processing Systems*, 36 (2/3), 153-159.
- [Kemp04] Kemp, A. H., Silberstein, R. B., Armstrong, S. M. & Nathan, P. J. (2004). Gender Differences in the Cortical Electrophysiological Processing of Visual Emotional Stimuli. *NeuroImage*, 21(2) 632 – 646.
- [Klei81] Kleinginna, P. R. Jr. & Kleinginna, A. M. (1981). A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definition, *Motivation and Emotion*, 5 (4), 345 – 379.
- [Kost08] Kostoulas, T., Mporas, I., Ganchev, T. & Fakotakis, N. (2008). The Effect of Emotional Speech on a Smart-Home Application. In: *Proceeding of the 21st International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems: New Frontiers in Applied Artificial Intelligence (IEA/AIE 2008)*, Wroclaw, Poland, 305-310.
- [LaMa06] LaMar, M.D., Xin, J. & Qi, Y. (2006). Signal Processing of Acoustic Signals in the Time Domain with an Active Nonlinear Nonlocal Cochlear Model. *Signal Processing*, 86, 360 – 374.
- [Lang95] Lang, P. J. (1995). The Emotion Probe: Studies of Motivation and Attention. *American Psychologist*, 50(5), 372-385.
- [Laza91] Lazarus, R. S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.
- [Laza94] Lazarus, R. S. & Lazarus, B. N. (1994). *Passion and Reason: Making Sense of Our Emotions*. New York, USA: Oxford University Press.

- [Lee01] Lee, C.M., Narayanan, S. S. & Pieraccini, R. (2001). Recognition of Negative Emotions from the Speech Signal. *In: Proceeding of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2001)*, Trento, Italy, 240-243.
- [Lee02] Lee, T. M., Liu, H. L., Hoosain, R., Liao, W. T., Yuen, K. S., Chan, C. C., Fox, P. T. & Gao, J. H. (2002). Gender Differences in Neural Correlates of Recognition of Happy and Sad Faces in Humans Accessed by Functional Magnetic Resonance Imaging, *Neuroscience Letter*, 333(1), 13 -16.
- [Lee05] Lee C. M. & Narayanan, S. S. (2005). Toward Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2), 293 – 303.
- [Lewi83] Lewis, M. & Michaelson, L. (1983). *Children's Emotions and Moods: Developmental Theory and Measurement*. New York: Plenum Press.
- [Li09] Li, Q. (2009). An Auditory-based Transform for Audio Signal Processing. *In: Proceeding of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '09)*, New York, USA, 181 – 184.
- [Lisb08] Lisboa, P. J. G., Ellis, I. O., Green, A. R., Ambrogi, F. & Dias, M. B. (2008). Cluster-based Visualization with Scatter Matrices. *Pattern Recognition Letters*, 29, 1814-1823.
- [Litm06] Litman, D. J. & Forbes-Riley, K. (2006). Recognizing Student Emotions and Attitudes on the Basis of Utterances in Spoken Tutoring Dialogue with both Human and Computer Tutors. *Speech Communication*, 48(5), 559-590.
- [Live07] Livelybrooks, D. (2007). Human Hearing and Voice [On-line]. Available : <http://hendrix2.uoregon.edu/~dlivelyb/phys152/L9.html>
- [Lloy57] Lloyd, S. P. (1957). Least Square Quantization in PCM. *Bell Telephone Laboratories Paper*.
- [Lloy82] Lloyd, S. P. (1982). Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129-137.
- [Love97] Loveland, K. A., Tunali-Kotoski, B., Chen, Y. R., Ortegón, J., Pearson, D. A., Brelsford, K. A. & Gibbs, M. C. (1997). Emotion Recognition in Autism: Verbal and Nonverbal Information, *Development and Psychopathology*, 9, 579-593.
- [MacK92] MacKay, D. J. C. (1992). Bayesian Interpolation. *Neural Computation*, 4(3), 415-447.

-
- [MacK94] MacKay, D. J. C. (1994). Bayesian Non-linear Modelling for the Energy Prediction Competition. *ASHRAE Transactions*, 100(2), 1053-1062.
- [Mamd77] Mamdani, E. H. (1977). Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Systems. *IEEE Transactions on Computers*, 12, 1182-1191.
- [Mats01] Matsumoto, D. (2001). Culture and Emotion. In: Matsumoto, D. (Ed). *Handbook of Culture and Psychology*, (pp. 171-194), New York: Oxford University Press.
- [Miya10] Miyamoto, Y., Uchida, Y. & Ellsworth, P. C. (2010). Culture and Mixed Emotions: Co-occurrence of Positive and Negative Emotions in Japan and the United States. *Emotion*, 10(3), 404 – 415.
- [Moll06] Moller, S., Krebber, J. & Smeele, P. (2006). Evaluating the Speech Output Component of a Smart-home System. *Speech Communication*, 48(1), 1–27.
- [Mont05] Montagne, B., Kessels, R.P., Frigerio, E., de Haan, E.H., Perrett, D.I. (2005). Sex Differences in the Perception of Affective Facial Expressions: Do Men Really Lack Emotional Sensitivity? *Cognitive Processing*, 6(2), 136—141.
- [Morr07] Morrison, D., Wang, R. & De Silva, L. C. (2007). Ensemble Methods for Spoken Emotion Recognition in Call-Centres. *Speech Communication*, 49(2), 98–112.
- [Mowe09] Mower, E., Mataric, M. J. & Narayanan, S. S. (2009). Human Perception of Audio-visual Synthetic Character Emotion Expression in the Presence of Ambiguous and Conflicting Information, *IEEE Transaction on Multimedia*, 11(5), 843-855.
- [Mozz00] Mozziconacci, S. J. L. & Hermes, D. J. (2000). Expression of Emotion and Attitude Through Temporal Speech Variations. In: *Proceeding of International Conference on Spoken Language Processing (ICSLP '00)*, Beijing, China, 373-378.
- [Murr96] Murray, I. R., Baber, C., & South, A. (1996). Towards a Definition and Working Model of Stress and Its Effects on Speech. *Speech Communication*, 20 (1-2), 3-12.
- [Naqv06] Naqvi, N., Shiv, B. & Bechara, A. (2006). The Role of Emotion in Decision Making: A Cognitive Neuroscience Perspective. *Current Directions in Psychological Science*, 15(5), 260-264.
- [Nass05] Nass, C., Jonsson, I-M., Harris, H., Reaves, B., Endo, J., Brave, S. & Takayama, L. (2005). Improving Automotive Safety by Pairing Driver Emotion and Car Voice Emotion. In: *Extended Abstracts on Human Factors in Computing Systems (CHI'05)*, Portland, USA, 1973-1976.
- [Neal96] Neal, R. M. (1996). Bayesian Learning for Neural Networks. *Lecture Notes in Statistics*, 118. New York, NY: Springer-Verlag.
-

- [Neal98] Neal R. M. (1998). Assessing Relevance Determination Methods Using DELVE. In C. M. Bishop (Ed.), *Neural Networks and Machine Learning*, (pp. 97-129), New York, NY: Springer-Verlag.
- [Nich00] Nicholson, J., Takahashi, K. & Nakatsu, R. (2000). Emotion Recognition in Speech Using Neural Networks. *Neural Computing and Applications*, 9 (4), 290-296.
- [Nwe03] Nwe, T. L., Foo, S. W. & De Silva, L. C. (2003). Speech Emotion Recognition Using Hidden Markov Models. *Speech Communication*, 41 (4), 603-623.
- [Oent06] Oentaryo, R. J. & Pasquier, M. (2006). GenSoFNN-Yager: A Novel Hippocampus-like Learning Memory System Realizing Yager Inference. In: *Proceeding of the International Joint Conference on Neural Networks (IJCNN'06)*, Vancouver, Canada, 705 – 712.
- [Oent08] Oentaryo, R. J., Pasquier, M. & Quek, C. (2008). GenSoFNN-Yager: A Novel Brain-inspired Generic Self-organizing Neuro-fuzzy System Realizing Yager Inference, *Expert System with Applications*, 35(4), 1825 – 1840.
- [Pank05] Panksepp, J. (2005). Affective Consciousness : Core Emotional Feelings in Animals and Humans, *Consciousness and Cognition*, 14, 30-80.
- [Patt96] Patterson, D. (1996). *Artificial Neural Network*. Singapore: Prentice Hall.
- [Patt02] Patton, M. Q. (2002). *Qualitative Evaluation and Research Methods*, 3rd Edition, Thousand Oak, CA: Sage Publications, pp.39.
- [Pell08] Pell, M. D. & Monetta, L. (2008). How Parkinson's Disease Affects Nonverbal Communication and Language Processing, *Language and Linguistics Compass*, 2 (5), 739-759.
- [Pell09a] Pell, M. D., Paulmann, S., Dara, C., Alasseri, A. & Kotz, S. A. (2009). Factors in the Recognition of Vocally Expressed Emotion: A Comparison of Four Languages. *Journal of Phonetics*, 37, 417-435.
- [Pell09b] Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing Emotions in Foreign Language. *Journal of Nonverbal Behavior*, 33 (2), 107-120.
- [Petr99] Petrushin, V. A. (1999). Emotion in Speech: Recognition and Application to Call Centers. In: *Proceeding of Artificial Neural Networks in Engineering (ANNIE '99)*, St. Louise, Missouri, 7-10.
- [Plut02] Plutchik, R. (2002). *Emotions and Life: Perspective from Psychology, Biology and Evolution*. Washington, DC: American Psychological Association.

-
- [Prin04] Prinz, J. (2004). Embodied Emotions. In R. C. Solomon (Ed), *Thinking About Feeling: Contemporary Philosophers on Emotions (Series in Affective Science)*, (pp.44-60), New York, NY: Oxford University Press.
- [Qi04] Qi, Y., Minka, T.P., Picard, R.W. & Ghahrami Z. (2004). Predictive Automatic Relevance Determination by Expectation Propagation. In: *Proceeding of the 21st International Conference on Machine Learning (ICML-04)*, Banff, Canada.
- [Quah07] Quah, K. H. & Quek, C. (2007). MCES: A Novel Monte Carlo Evaluative Selection Approach for Objective Feature Selections, *IEEE Transaction on Neural Networks*, 18 (2), 431 – 448.
- [Rabi93] Rabiner, L. R. & Juang, B. H. (1993). *Fundamentals of Speech Emotion Recognition*. Englewood Cliffs, NJ : Prentice-Hall.
- [Raym09] Reymond, J. (2009). Interaction of Attention, Emotion and Motivation. *Progress in Brain Research*, 176, 293-308.
- [Reyn00] Reynolds, D. A., Quatieri, T. F. & Dunn, R. B. (2000) Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3), 19-41.
- [Reyn95a] Reynolds, D. A. & Rose, R. C. (1995). Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*, 3(1), 72-83.
- [Reyn95b] Reynolds, D. A. (1995). Speaker Identification and Verification using Gaussian Mixture Speaker Models. *Speech Communication*, 17, 91 – 108.
- [Rip196] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, U.K: Cambridge University Press.
- [Robe98] Roberts, S. J., Husmeier, D., Penny, W. & Rezek, I. (1998). Bayesian Approaches to Gaussian Mixture Modeling, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20 (11), 1133 – 1142.
- [Rose62] Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanism*, Washington D.C: Spartan Books
- [Rume86] Rumelhart, D. E., Geoffrey, E. H. & Williams, R. J. (1986). Learning Internal Representations by Error Propagation, In D. E. Rumelhart, J. L. McClelland and the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Volume 1: Foundation, Cambridge, MA, USA MIT Press
-

- [Russ77] Russell, J.A., & Mehrabian, A. (1977). Evidence for a Three-factor Theory of Emotions. *Journal of Research in Personality*, 11(3), 273-294.
- [Russ79] Russell, J. A. (1979). Affective Space is Bipolar. *Journal of Personality and Social Psychology*, 37(3), 345-356.
- [Russ85] Russell, J. A. & Bullock, M. (1985). Multidimensional Scaling of Emotional Facial Expressions: Similarity from Preschoolers to Adults. *Journal of Personality and Social Psychology*, 48(5), 1290-1298.
- [Sche78] Scherer, K. R. (1978). Personality Inference from Voice Quality: The Loud Voice of Extroversion. *European Journal of Social Psychology*, 8, 467-487.
- [Sche86] Scherer, K. R. (1986). Vocal Affect Expression: A Review and a Model for Future Research. *Psychological Bulletin*, 99, 143-165.
- [Sche89] Scherer, K. R. (1989). Emotion Psychology can Contribute to Psychiatric Work on Affect Disorder: A Review. *Journal of the Royal Society of Medicine*. 82, 545-547.
- [Sche01a] Scherer, K. R., Banse, R. & Wallbott, H. G. (2001). Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *Journal of Cross-Cultural Psychology*, 32 (1), 76-92.
- [Sche01b] Scherer, K. R., Shorr, A., & Johnstone, T. (Ed.) (2001). *Appraisal Processes in Emotion: Theory, Methods, Research*. Canary, NC: Oxford University Press.
- [Sche03] Scherer, K. R. (2003). Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication*, 40, 227-256.
- [Schl54] Schlosberg, H. (1954). Three Dimensions of Emotion. *Psychological Review*, 61(2), 81-88.
- [Schu06] Schuller, B., Lang M. & Rigoll. G. (2006). Recognition of Spontaneous Emotion by Speech within Automotive Environment. In: *Proceeding of 32nd Deutsche Jahrestagung fur Akustic (DEGA '06)*, Braunschweig, Germany, 57-58.
- [Schu09] Schuller, B., Steidl, S. & Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. In: *Proceeding of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, Brighton, UK, 312-315.
- [Self10] Self, D. (2010). *Audio Engineering Explained – Professional Audio Recording*, Burlington, MA, USA: Focal Press.

- [Skow02] Skowronski, M. D. & Harris, J. G. (2002). Increased MFCC Filter Bandwidth for Noise-robust Phoneme Recognition. *In: Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1, Florida, USA*, I-801 – I-804.
- [Skow04] Skowronski, M. D. & Harris, J. G. (2004). Exploiting Independent Filter Bandwidth of Human Factor Cepstral Coefficient in Automatic Speech Recognition. *Journal of Acoustical Society of America*, 116 (3), 1774-1780.
- [Slan98] Slaney, M. (1998). Auditory Toolbox. (Ver. 2). Technical Report #1998-010, Interval Research Corporation. (2007, June 13). [Online]. Available: <http://cobweb.ecn.purdue.edu/~malcolm/interval/1998-010/>
- [Somo10] Somol, P., Novovicova, J. & Pudil, P. (2010). Efficient Feature Subset Selection and Subset Size Optimization, *Pattern Recognition Recent Advances*, 75 – 97.
- [Ste09] Steidl, S. (2009). Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech, PhD Thesis, Logos Verlag, Berlin, Germany.
- [Step06] Stephens, A. N. & Groeger, J.A. (2006). Do Emotional Appraisals of Traffic Situations Influence Driver Behavior? *In: Paper Presented at the Behavioral Studies Seminar*, Bath, UK.
- [Stor97] Storn, R. & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces, *Journal of Global Optimization*, 11, 341 -359.
- [Tung02a] Tung, W. L. & Quek, C. (2002). GenSoFNN: A Generic Self-Organizing Fuzzy Neural Network. *IEEE Transactions of Neural Networks*, 13(5), 1075-1086.
- [Tung02b] Tung, W. L. & Quek, C. (2002). DIC: A Novel Discrete Incremental Clustering Technique for the Derivation of Fuzzy Membership Functions. In M. Ishizuka & A. Sattar (Eds.). *PRICAI 2002: Trends in Artificial Intelligence. Lecture Notes in Computer Science, LNAI 2417*, (pp.485-491). Berlin Heidelberg: Springer-Verlag.
- [Verv03] Ververidis, D. & Kotropoulos, C. (2003). A State of the Art Review on Emotional Speech Databases. *In: Proceeding of the First Richmedia Conference*, Lausanne, Switzerland, 109-119.
- [Verv06] Ververidis D. & Kotropoulos, C. (2006). Emotional Speech Recognition: Resources, Features and Methods, *Speech Communication*, 48(9), 1162-1181.

-
- [Vogt07] Vogt, T., Andre E. & Wagner, J. (2007). Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization. In: C. Peter & R. Beale (Eds.) *Affect and Emotion in Human-Computer Interaction. Lecture Notes in Computer Science (LNCS)*, 4868, Heidelberg, Germany: Springer.
- [Waha07] Wahab, A., Tan, C. K., Abut, H. & Takeda, K. (2007). Driver Recognition System using FNN and Statistical Methods. In H. Abut, J. H. L. Hansen & K. Takeda (Eds.), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards*, Chapter 2, (pp. 11 – 23). New York, USA: Springer Science+Business Media.
- [Waha09] Wahab, A., Quek, C., Keong, T. C. & Takeda, K. (2009). Driving Profile Modeling and Recognition Based on Soft Computing Approach. *IEEE Transactions on Neural Networks*, 20(4), 563-582.
- [Wang08] Wang, Y. & Guan, L. (2008). Recognizing Human Emotional State from Audiovisual Signals. *IEEE Transactions on Multimedia*, 10(5), 936 – 946.
- [Weis06] Wieser, M. J., Muhlberger, A., Alpers, G. W., Macht, M., Ellgring, H. & Pauli, P. (2006). Emotion Processing in Parkinson's Disease: Dissociation Between Early Neuronal Processing and Explicit Ratings. *Clinical Neurophysiology*, 117, 92-102.
- [Witt99] Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G. & Cunningham, S. J. (1999) Weka: Practical Machine Learning Tools and Techniques with Java Implementations. In: N. Kasabov & K. Ko (Eds.). *Proceedings of the ICONIP/ANZIIS/ANNES'99 International Workshop on Emerging Knowledge in Engineering and Connectionist-Based Information Systems*. Dunedin, New Zealand, 192-196.
- [Woll09] Wollmer, M., Eyben, F. Schuller, B., Douglas-Cowie, E. & Cowie, R. (2009). Data-driven Clustering in Emotional Space for Affect Recognition Using Discriminatively Trained LSTM Networks, In: *Proceedings of the 10th Annual Conference of the International Speech Communication Association INTERSPEECH 2009*, Brighton, United Kingdom, 1595-1598.
- [Woma99] Womack, B. D. & Hansen, J. H. L. (1999). N-channel Hidden Markov Models for Combined Stressed Speech Classification and Recognition. *IEEE Transactions on Speech and Audio Processing*, 7(6), 668-677.

-
- [Wu10] Wu, D., Parsons, T. D. & Narayanan, S. S. (2010). Acoustic Feature Analysis in Speech Emotion Primitives Estimation, *In Proceeding of the 11th Annual Conference of the International Speech Communication Association INTERSPEECH 2010*, Chiba, Japan, 785-788.
- [Yage94] Yager, R. & Filev, D. (1994). Generation of Fuzzy Rules by Mountain Clustering. *Journal of Intelligent and Fuzzy System*, 2(3), 209-219.
- [Yild11] Yildiz, O. T. (2011). Mapping Classifiers and Datasets. *Expert Systems with Application*, 38, 3697-3702.
- [Yama06] Yamaguchi M., Wakasugi J. & Sakakima J. (2006). Evaluation of Driver Stress Using Biomarker in Motor-Vehicle Driving Simulator. *In: Proceedings of the 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '06)*, New York City, USA, 1834-1837.
- [Yang92] Yang, X., Wang, K. & Shamma, S. (1992). Auditory Representations of Acoustic Signals, *IEEE Transaction on Information Theory*, 38(2), part 2, 824 – 839
- [Youn95] Young, S. J., Jansen, J., Odell, J. J., Ollason, D. G. & Woodland, P. C. (1995). *The HTK Book Manual* (Ver. 2.1). Cambridge University Engineering Department and Entropic Cambridge Research Laboratory, Cambridge, UK.
- [Zhou01] Zhou, G., Hansen, J H L. & Kaiser, J. F. (2001). Non-Linear Feature-based Classification of Speech under Stress. *IEEE Transactions on Speech and Audio Processing*, 9(3), 201-216.

LIST OF PUBLICATIONS

JOURNAL

- [1] Wahab, A., **Kamaruddin, N.**, Palaniappan, L. K., Li, M. & Khosrowabadi, R. (2010). EEG Signals for Emotion Recognition. *Journal of Computational Methods in Science and Engineering (JCMSE)*, 10(Supplement 1), S1-S11.
- [2] **Kamaruddin N.** & Wahab A. (2009). Features Extraction for Speech Emotion. *Journal of Computational Methods in Science and Engineering (JCMSE)*, 9(Supplement 1), S1 – S12.
- [3] **Kamaruddin, N.**, Wahab, A. & Chai, Q. (2011). Cultural Dependency Analysis for Understanding Speech Emotion. *Journal of Expert System with Application (ESWA)*, 39(5), 5115-5133.

BOOK CHAPTER

- [1] **Kamaruddin, N.**, Wahab, A. and Abut, H. (2011). DSP (Digital Signal Processing) for In-Vehicle Systems and Safety, Chapter 2: Driver Emotion Profiling from Speech, J. H. L. Hansen, P. Boyraz, K. Takeda, H. Abut (Eds.), In Press, Springer, New York, 2011.

CONFERENCES

2012

- [1] **Kamaruddin, N.** and Wahab, A. (2012). Human Behavior State Profile Mapping Based on Recalibrated Speech Affective Space Model *In: Proceeding of The 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '12)*, 28 August – 1 September, 2012, San Diego, USA
- [2] Yaacob, H., Karim, I., Wahab, A. and **Kamaruddin, N.** Two Dimensional Affective State Distribution of the Brain under Emotion Stimuli, *In: Proceeding of The 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '12)*, 28 August – 1 September, 2012, San Diego, USA

2011

- [3] **Kamaruddin, N.** & Wahab, A. (2011). Heterogeneous Driver Behavior State Recognition using Speech Signal *In: Proceeding of the 10th WSEAS International Conference on System Science and Simulation in Engineering (ICOSSSE '11)*, 3 – 5 October, Penang, Malaysia, 207-212.
- [4] Nor, M. N., Wahab, A., Majid, H. & **Kamaruddin, N.** (2011). Pre-Post Accident Analysis Relates to Pre-cursor Emotion for Driver Behavior Understanding *In: Proceeding of the 11th WSEAS International Conference on Applied Computer Science (ACS '11)*, 3 – 5 October, Penang, Malaysia, 152-157.
- [5] Nor, N. M., Wahab, A., **Kamaruddin, N.** & Abut, H. (2011). Valence-Arousal Analysis for Understanding Driver Behavior *In: Proceeding of the 5th Biennial Workshop on Digital Signal Processing for In-Vehicle Systems*, 4 -7 September 2011, Kiel, Germany.
- [6] Nor, N. M., Wahab, A., **Kamaruddin, N.** & Majid, H. (2011). Post Accident Analysis of Driver Affection *In: Proceeding of the 2011 IEEE 15th International Symposium on Consumer Electronics (ISCE '2011)*, 14-17 June 2011, Singapore, 278-283.
- [7] Rahnuma, K. S., Wahab, A., **Kamaruddin, N.** and Majid, H. EEG Analysis for Understanding Stress Based on Affective Model Basis Function *In: Proceeding of the 2011 IEEE 15th International Symposium on Consumer Electronics (ISCE '2011)*, 14-17 June 2011, Singapore, 592-597,

2010

- [8] **Kamaruddin, N.** & Wahab, A. (2010). Driver Behavior Analysis through Speech Emotion Understanding. *In: Proceeding of the 2010 IEEE Intelligent Vehicle Symposium (IV 2010)*, 21-24 June 2010, San Diego, California, USA, 238-243.
- [9] **Kamaruddin, N.** & Wahab, A. (2010). Cultural Influence on Speech Emotion Identification. *In: Proceeding of the 25th International Conference on Computer and Its Application (CATA 2010)*, 22-25 March 2010, Honolulu, Hawaii, USA, 264-269.

2009

- [10] **Kamaruddin, N.** & Wahab, A. (2009). CMAC for Speech Emotion Profiling. *In: Proceeding of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009)*, 6-10 Sep 2009, Brighton, England, UK, 1991-1994. ~ **Awarded Student Travel Grant from INTERSPEECH 2009**
- [11] **Kamaruddin, N.**, Wahab, A. & Abut, H. (2009). Driver Emotion Profiling from Speech. *In: Proceeding of the 4th Biennial Workshop on DSP for In-Vehicle Systems and Safety*, 24-27 Jun 2009, Dallas, Texas, USA
- [12] **Kamaruddin, N.** & Wahab, A. (2009). Speech Emotion Profiling for Inter and Intra Cultural Assessment. *In: Proceeding of the 18th International Conference on Software Engineering and Data Engineering (SEDE 2009)*, 22-24 June 2009, Las Vegas, Nevada, USA, 275-280.
- [13] Wahab, A., **Kamaruddin, N** & Palaniappan, L. (2009). EEG Signals for Emotion Recognition. *In: Proceeding of the 18th International Conference on Software Engineering and Data Engineering (SEDE 2009)*, 22-24 June 2009, Las Vegas, Nevada, USA, 289-294.

2008

- [14] **Kamaruddin, N.** & Wahab, A. (2008). Speech Emotion Verification and Profiling. *In: Proceeding of the 21st International Conference on Computer Applications in Industry and Engineering (CAINE 2008)*. 12-14 Nov 2008, Honolulu, Hawaii, USA, 44-49.
- [15] Wahab, A., **Kamaruddin, N.** & Khalid, M. (2008). CMAC for Driver Behavior Profiling Based on Real-Time Driving Data. *In: Proceeding of 21st International Conference on Computer Applications in Industry and Engineering (CAINE 2008)*, 12-14 Nov 2008, Honolulu, Hawaii, USA, 215-220.
- [16] **Kamaruddin, N.** & Wahab, A. (2008). Speech Emotion Verification System (SEVS) Based On MFCC For Real Time Applications. *In: Proceeding of the 4th International Conference on Intelligent Environments (IE '08)*, 21-22 Jul 2008. Seattle, Washington, USA, 1-7.

- [17] **Kamaruddin, N** & Wahab, A. (2008). Speaker Independent Emotion Recognition System – SIERS. *In: Proceeding of the 2008 International Conference on Artificial Intelligence (ICAI '08)*, 14-17 Jul 2008, Las Vegas, Nevada, USA, 10-15.
- [18] Khalid, M., Wahab, A. & **Kamaruddin, N.** (2008). Real Time Driving Data Collection and Driver Verification Using CMAC-MFCC. *In: Proceeding of the 2008 International Conference on Artificial Intelligence (ICAI '08)*, 14-17 Jul 2008, Las Vegas, Nevada, USA, 219-224.
- [19] Wahab, A., Keat, H. W. & **Kamaruddin, N.** (2008). CMAC-GMM Driver Verification. *In: Proceeding of the 2008 International Conference on Artificial Intelligence (ICAI '08)*, 14-17 Jul 2008, Las Vegas, Nevada, USA, 269-274.
- [20] **Kamaruddin, N.** & Wahab, A. (2008). Feature Extraction for Speech Emotion. *In: Proceeding of the 17th International Conference on Software Engineering and Data Engineering (SEDE 2008)*, 30 June – 2 July 2008. Los Angeles, California, USA, 120-125. ~Nominated for best paper award

2007

- [21] Wahab, A., Wen T. G. & **Kamaruddin, N.** (2007). Understanding Driver Behavior Using Multi-dimensional CMAC. *In: Proceeding of the 6th International Conference on Information, Communication & Signal Processing. (ICICS 2007)*, 10-13 Dec 2007, Singapore, 1-5.

SURVEY ON EMOTIONAL PERCEPTIONS**Personal Information** (Please tick (v) wherever appropriate)

Gender	☐ Male	☐ Female
Age	☐ Less than 20 years	☐ 20 to 29 years
	☐ 30 to 39 years	☐ More than 40 years
Education level	☐ Post graduate	☐ Graduate
	☐ Non-graduate	

Listen carefully the audio clip. Please tick whichever appropriate to what you associate with the feeling you feel when you listen to the clips. If necessary, repeat the audio clip to verify your emotion perception feedback.

	Happy	Angry	Disgust	Surprise	Sad	Neutral
File1						
File2						
File3						
File4						
File5						
File6						
File7						
File8						
File9						
File10						
File11						
File12						
File13						
File14						
File15						
File16						
File17						
File18						
File19						
File20						

Thank you for your cooperation.

ARD Ranking of Emotion versus Neutral

Rank #	Angry vs neutral		Happy vs neutral		Sad vs neutral	
	Feature #	ARD weight	Feature #	ARD weight	Feature #	ARD weight
1	12	8.8745	12	8.8745	9	5.4327
2	22	5.0711	22	5.0711	34	4.2355
3	33	3.6337	33	3.6337	18	2.9892
4	25	3.5043	25	3.5043	38	2.5623
5	4	3.2901	4	3.2901	21	2.4489
6	35	3.2898	35	3.2898	37	2.2234
7	5	3.2843	5	3.2843	22	2.0771
8	15	2.9060	15	2.906	25	1.9681
9	24	2.3584	24	2.3584	36	1.8649
10	31	2.2606	31	2.2606	7	1.8487
11	10	2.0664	10	2.0664	10	1.8063
12	21	2.0428	21	2.0428	23	1.7822
13	32	2.0379	32	2.0379	26	1.7663
14	30	1.9937	30	1.9937	1	1.6968
15	34	1.9515	34	1.9515	8	1.6615
16	28	1.7703	28	1.7703	20	1.5345
17	7	1.3872	7	1.3872	12	1.4628
18	19	1.3423	19	1.3423	33	1.37
19	2	1.3333	2	1.3333	2	1.2452
20	40	1.2934	40	1.2934	40	1.2362
21	8	1.1908	8	1.1908	32	1.1257
22	1	1.1335	1	1.1335	24	1.0411
23	14	1.1207	14	1.1207	16	1
24	16	1.0611	16	1.0611	14	0.9807
25	37	0.9804	37	0.9804	15	0.9529
26	6	0.9631	6	0.9631	5	0.9425
27	17	0.9088	17	0.9088	31	0.7914
28	13	0.8725	13	0.8725	27	0.7744
29	38	0.8549	38	0.8549	29	0.7439
30	9	0.8297	9	0.8297	17	0.7277
31	29	0.6704	29	0.6704	35	0.704
32	11	0.5553	11	0.5553	28	0.6856
33	23	0.5007	23	0.5007	11	0.5254
34	18	0.4350	18	0.435	3	0.3978
35	3	0.3178	3	0.3178	39	0.3708
36	36	0.3005	36	0.3005	13	0.3645
37	20	0.2628	20	0.2628	6	0.364
38	27	0.2367	27	0.2367	30	0.2287
39	26	0.2032	26	0.2032	19	0.1535
40	39	0.2011	39	0.2011	4	0.1141

APPENDIX A111**Re-profiling of the Driver Behaviour State: A practical example of using the VA approach for analysis**

Grimm in his finding [Grim07] has shown that emotion can be identified based on the Valence, Arousal and Dominance emotion primitives' values of the speech signals. From the VA approach and experimental results reported in Section 6.1, a similar set of observations are reached that emotion primitives' values of Valence and Arousal are able to generate a dynamic data-driven affective space model. This newly recalibrated affective space model is heavily biased to cultural influence. Such finding is also supported by many hypothesis that culture does affect the speech emotion recognition accuracy performance. Hence, the idea of the VA approach is extended to study the driver behaviour state (DBS). This approach is proposed due to the fact that both inputs are derived from human speech as a common communication medium and behavior is known as the manifestation of emotion. Such notion is supported on the basis of the hypothesis that emotion influences the driver behaviour and emotion can be quantified through speech. Since emotion is conditioned by culture, culturally-influenced speech datasets are employed as a training data for reference purposes. The three basic emotions of anger, happiness and sadness are used as a foundation for deriving the affective space model for the behavior profiling of the driver. In addition, neutral is used as an origin for the newly recalibrated data-driven affective space model. The idea of using the basic emotions as a basis function enables the correlation between the speech emotion and the DBS to be observed. This may provide better understanding of the effect of emotion to the driver while they are driving. In addition, a reasonable profile of the driver behavior with respect

to the Valence and Arousal emotion primitives' values can also be generated for further analysis.

Driver behaviour state (DBS) is dynamic and it can gradually change over a short period of time based on the driver interaction with the vehicle and the environment. Due to its complexity, most researchers focus on recognizing static behaviour. Such approach does not adequately explain the underlying drivers' state since it can be erratic and impulsive. In this section, four different DBS; namely: distracted (talking on the mobile phone), laughing, feeling sleepy and driver having normal conversation while driving are studied. These speech data are extracted from the Real-time Speech Driving (RtSD) dataset and the detailed explanation of the dataset is provided in the next sub-section. These DBS data are then re-profiled using the newly recalibrated data-driven affective space model based on the different culturally-influenced speech emotion datasets. The three datasets employed are: the Berlin, NTU_American and NTU_Asian datasets, which represent the European, American and Asian cultures respectively. The techniques employed for the generation of the affective space model using these culturally-influenced datasets has been described in Section 6.1. Based on the hypothesis that culture influences speech emotion recognition performance, both homogeneous and heterogeneous intra-cultural data mixture arrangements are used to give different perspective of cultural dependency. The rationale of the VA approach implementation in understanding DBS is to comprehensively study the underlying reciprocal relationship of the DBS and basic emotion by mapping the DBS Valence and Arousal emotion primitives' values on the VA space. Such detailed study facilitates better understanding of the emotion distribution and its relation to the driver's behaviour. A brief introduction on the driving activities and DBS, the RtSD dataset employed and its mapping to the

affective space model, will be presented in the next sub-section.

Brief Introduction of Driving Activities and Driver Behaviour State

Human judgment in reacting to changes and the environment are influenced by their emotion and perception. In addition, the adverse effects of the roused emotion can bias one's judgment and even change the priority of any action that are taken. Some psychologists proposed that emotion directly influences judgment by serving as experiential and bodily information regarding how one feels about the object that affect judgment [Clor07]. Such experiential information sometimes are processed faster than rational thought resulting in an irrational driving situation. Emotions can affect drivers' concentration, invoking recklessness and disabling appropriate self-control. There are many instances where drivers may experience overwhelmed emotion while driving. The examples could be when someone is notified of the death of a loved one, or passed through a street that raises melancholic memories, or when they encountered an obscene gesture or an unreasonable and offensive driver behaviour from another driver. Such a scenario may influence the driver to respond with 'rage driving' which displays a total lack of control and leads to aggressive driving. This observation has been supported by the finding of Stephens and Groeger [Step06] that most drivers would feel angry and frustrated when they have to reduce speed because of traffic and subsequently retaliated by accelerating after the impending event.

Driving activity in general can be viewed as a complex cognitive task that constitutes attention, information processing and decision making task which depends on the vehicle performance and environment itself. The current road condition, behaviour of the other road users, the vehicle state and resource information such as petrol and duration

of excursion are some examples of important inputs that are needed by the driver to make the necessary choice. Bellet et al. [Bell09] formulated a driver mental representation as an interaction between the long term memory (LTM) with the working memory (WM) and its environment as depicted in Figure A111.1.

Figure A111.1: Elementary Architecture of the Driver Cognitive System

The WM is a structure that hosts active knowledge instantiated through mental representations to serve the activity in progress that focuses on the driver's current goal. The LTM permanently stores the driving knowledge in the latent state from the learning process. The driving knowledge would then be reactivated later on the basis of needs. These driving knowledge couples with information from the environment, which is gathered through perceptive process, constructs the mental representation of the driving situation. Moreover, the construction of such mental representations is complemented with the aid of emotionally influenced reasoning process; which includes: decision making, anticipation and action planning. The output is then manifested through the realization of the driver's behaviour. But in some cases, the cognitive resources may be

influenced by some pre-emotion reaction or pre-cursor emotion, which will affect the output. From this understanding, it can be hypothesized that even the smallest disturbance can potentially lead to severe repercussions [Nass05].

Effect from illicit drug and alcohol consumption, fatigue, stress and excessive uncontrolled emotion exhibited formidable distraction level to the drivers. Due to these distractions, the driver's ability from observing the road condition such as an abrupt slowing of traffic or debris on the road or the other drivers' changing lanes, slowing or moving faster and maneuvering, may be seriously impaired. The failure to recognize such situations will harm not only the driver himself/herself but also the other road users and pedestrians. If such anomalous behaviors can be identified, a warning system can be made available to the driver to prevent any unforeseen situation from occurring. For instance, if it is discovered that the driver has shown any sign of emerging sleepiness (boredom), loss of concentration (distraction) or diminishing ability to drive, the driver is then urged to take a break. A driver emotion indicator application based on speech can be incorporated into an intelligent motor vehicle system as a preventive measure from any unwanted incidents. This is based on the hypothesis that emotion can influence the way the driver behaves and emotional profiles can be quantified using the driver's speech based on the VA approach of the affective space model. Any abnormalities in the driver's behavior can be identified so that an appropriate warning mechanism can be developed. Such a system can provide means to monitor the drivers' behaviour in complex traffic situation with practical implementation.

Figure A11.2 shows a simplified block diagram of the proposed driver behavior state analysis system where the learning process uses the basic emotion including the neutral serving as the emotionless state to generate the newly recalibrated

affective space model. The testing process is then uses the rules generated from this learning system as part of the mechanism to profile and hence determine the value of the emotion primitives correlating to the driver behavior state. The generation of the newly recalibrated data-driven affective space model has been discussed and analyzed in detailed in Section 6.1.

Figure A111.2: Block Diagram of the Extension of the Basic Speech Emotion Recognition System for Driver Behaviour State Analysis

The basic emotion identification using the VA approach illustrated on the left-hand side of the Figure A111.2 provides the profiling rules derived from the culturally-influenced speech emotion as a training data. Such learning is based on a supervised learning where the emotions' primitives' values of Valence and Arousal are labelled prior

to the testing exercise. Different DBS speech data are then tested with this set of rules to determine the influential emotion primitives' values of Valence (V) and Arousal (A) that have effect on the driver while he/she is in that particular state. Input features from drivers speech signal are extracted using the MFCC features extraction method. These information are then used as testing data to be mapped on the affective space model that is previously generated using the culturally-influenced dataset. From the result, correlation between the DBS and the emotion primitives' values can be visualized for further analysis and discussion.

Real-time Speech Driving (RtSD) Dataset

The Real-time Speech Driving dataset (RtSD) is collected to validate the proposed speech emotion recognition system. The RtSD dataset is used to analyze the driver behaviour state (DBS) from the perspective of speech emotion understanding and to determine the correlation between DBS and the V and A as the emotion primitives. The dataset is collected from real-time natural speeches of the drivers while they are driving a motor vehicle. Such an approach allows freely-expressed emotion to be captured in an actual driving condition.

Numerous data collection for real-time driving data has been collected over the years [Angk08, Kawa04]. Such large corpora collections enable researchers to dissect and analyze the relevant signals to understand further the driver behaviour state from many perspectives. For instance, some of the studies are conducted based on brake and gas pedal pressure signal [Waha09], heart rate and skin conductivity [Heal05], biomarkers in saliva [Yama06] as well as speech [Fern03]. As discussed in Chapter 2, Fernandez and Picard [Fern03] stated that a reasonable accuracy can be obtained in recognizing the

drivers' emotion. Based on this finding, it gives an indication that by using speech alone, driver behaviour changes can be detected. However, in this study, the focus is on profiling the different DBS based on the affective space model generated on the emotions primitives' values of V and A. Such approach facilitates better understanding of the correlation between the DBS and basic emotion.

Participants

The subjects are 11 adult drivers (3 female and 8 male drivers) who are a regular motorist. The drivers voluntarily participated and have given the unprimed consent indicating that they understood the risks and due responsibility if any harm actually happen during the driving exercise. The subjects are mostly university students and professors of Nanyang Technological University, Singapore. In preparation before performing the driving exercise, the subjects must complete the real-life driving survey questionnaires to determine their driving skill, years of driving experience, the hours they drive in a week and their initial emotion at the time of starting the driving exercise. Information pertaining to the experience and frequency of the driving of each driver indicates the driving knowledge stored in Long Term Memory. Moreover, information about initial emotion is crucial so that any change of emotion can be detected and measured. The driver also had to answer questions describing their health condition, medication taken and their last 15 minutes activity before the driving exercise. Any potential subjects that were deemed affirmatively unwell are excluded from the study to avoid any unwanted and unnecessarily risk.

All of the selected participants described themselves as a normal driver without any drug or alcohol misused and experiencing neutral emotion before they start to drive.

The years of driving experience ranging from 2 to 25 years and the drivers have at least 10 hours of driving per week. Methodology of this data collection is based on fair assumption that normal and distracted driver would behave differently and it is reflected prominently by their driving behaviour. Moreover, it is hypothesized that the drivers' emotion will be elicited given sufficient distractions/stimulus so that the drivers would behave differently. This hypothesis is rooted from Cannon-Bard theory of emotion [Cann27] stated that emotion is the result of the events. In this thesis, emphasis is given to the varying DBS re-profiling on the recalibrated data-driven affective space model generated by the emotion primitives' values. Such values are derived from the training of the different culturally-influence speech emotion datasets.

Equipment

Figure A111.3: Equipments Placement for the In-car Data Acquisition System

The driving exercise for this study is conducted in a Nissan Sunny Super Saloon car. The driver emotional speech is recorded with 2 mono microphones into a digital audio tape

recorder (Sony Digital Audio Tape Recorder) and two other microphones are used to record background and engine noises. These microphones are placed at the desktop of the vehicle, back seat, driver's collar and leg compartment as depicted in Figure 6.16. The additional microphones are important to enable the noise to be measured for the pre-processing stage. The recorded speech is subsequently re-digitized using Audacity software with 8 kHz sampling rate to allow the playback of the actual speech while driving to be listened. The speech data are then saved in Wave sound format (.wav).

Procedure

The in-car data acquisition system is installed and tested to ensure that it is in a good working condition. Briefing is provided prior to the driving exercise and is focused on safety, route condition, the tasks required to be performed and some reminder to the subject that they need to exhibit his/her normal driving behaviour to ensure that the data collected reflects the correct data. Synchronization and mileage initialization were recorded before the driving exercise commenced.

At each synchronization station, the driver needs to stop for a five-minute break so that the in-car data acquisition system equipment can be checked and re-synchronized. These stops allowed the driver to unwind and yet retain his/her driving momentum. The driver would iteratively do the drive and stop exercise until they reached the end point. At the end of the driving exercise, the subject is required to park the vehicle and data collection is stopped. The drivers are required to complete the questionnaire concerning their experience and emotion elicited during the driving exercise.

Speech sample

The speech samples are collected using the driver's speech while he/she has to communicate and at the same time they are required to complete the given task. There are four different scenarios under which the data were selected; namely: 1) driver talking through the cell phone (distracted driving), 2) laughing, 3) the driver feeling sleepy and 4) driver in the initial driving exercise where he/she is talking while driving normally. These four DBS are denoted as distracted, laughed, sleepy and normal states respectively. All of these DBS are recorded during the driving exercise in real time using normal driving road condition. The distracted DBS is selected based on the assumption that the driver's attention will be diverted because he/she had to multitask and divide the attention between concentrating on his/her driving and providing appropriate response to the caller on a cell phone. Such DBS inclusion in the analysis allows the detail profiling of the emotion primitives' values associated when the driver is interrupted by phone call during driving. Based on Murray et al. study [Murr96], this kind of distraction is grouped as a Level 4 Stressor to the speaker when the external stimuli is able to affect his/her speech processing.

The vocal expression of emotions is not limited to words and speeches. It also includes sobs, screams, growls and a spectrum of laughter types ranging from demure titters to snorting belly laughs. These non-linguistic repertoires complement the human semantic interaction. Thus for this study the laughed DBS is also included. Subjective tests reveal that it is difficult to discern happy emotions in speech purely on the basis of intonation differences. While it is possible to have happy speech without laughter in face-to-face interactions, the other non-verbal cues such as facial expressions help the interlocutor to understand that the underlying emotion is positive or happiness. However,

laughter is often used as a cue to manifest happiness for speech-only situations, hence making it suitable for positive Valence analysis. Thus, the laughed DBS is selected as one of the state to be studied in detail.

Sleepy DBS is incorporated in the analysis is because to further investigate the discrepancy of the Valence and Arousal values of the ‘pure’ sleepy and sleepy experienced by the drivers. Russell in his paper [Russ79] described sleepy as a state that has positive valence and negative arousal values (Quadrant 4 of affective space model) together with calmness. However, the driver may not be experiencing positive effect when feeling sleeping during driving. This is because he/she needs to maintain his level of arousal by forcing himself/herself to stay awake hence resulting in a negative effect of the Valence. This negative Valence values correspond to the negative emotion such as frustration or irritation. In addition, although sleepy is not considered as one of the Cornelius basic emotion [Corn96], it is deemed substantial to include sleepy detection in the proposed driver emotion indicator because sleepy is known as one of the factor that contributes to the high number of road accidents. Finally, the normal DBS is included to serve as a reference to the other DBS analyses. At this stage, the driver is experiencing neutral emotion at the initialization period prior to the commencement of the driving exercise. This DBS is important to gauge and measure the driver’s emotion differences before and after he/she completed the given tasks and responses during the driving exercise. The speech data are then labelled by matching each experimental condition with its distinctive DBS label; denoted namely as: distracted, laughed, normal and sleepy DBS for re-profiling purpose.

The Real-time Speech Dataset (RtSD) Validation

In order to validate the data collected in the RtSD dataset, a discrete-class DBS identification experiments are conducted. The rationale of this experiment is to investigate the ability of the different DBS data; namely: distracted, laughed, normal and sleepy to be discriminated. Such approach is suitable because the target class is known prior to the training exercise (supervised learning). In this experiment, the ability of the three selected classifiers; namely: the Adaptive Network-based Fuzzy Inference System (ANFIS), Multi Layer Perceptron (MLP) and Generic Self-organizing Fuzzy Neural Network (GenSoFNN) are studied in detail. In this experiment, the RtSD dataset is exclusively used for both training and testing exercise. A 5-fold validation technique is employed to ensure that the classifier is able to produce generalization accuracy performance. 80% of the data is used for training and the remaining 20% are used for testing. This cross-validation process is then repeated 5 times with each of the training-testing data pair used exactly once as the validation data. The result from each fold can be averaged to produce a single estimation of the DBS identification accuracy. Figure 6.17 shows the different classifiers performance of the DBS identification.

Figure A111.4: Performance Accuracy of Different Classifiers for DBS State Identification.

The supervised training classification result illustrated in Figure A111.4 shows that both the distracted and laughed DBS can be identified with ratio accuracy performance between 52.3% and 62.6%. The normal DBS is the most highly segregated DBS with accuracy performance reaching up to 80.6% for ANFIS, 76.8% for MLP and 65.4% for GenSoFNN classifier. This performance is closely followed by the sleepy DBS with the accuracy performance declining of 4.7% for ANFIS, 2.3% for MLP and 0.4% for GenSoFNN. It is interesting to observe that the implementation of the proposed DBS identification system is able to produce at least 65% accuracy in detecting sleepy driver. Subsequently, both the MLP and ANFIS classifiers yielded a fairly consistent result for the DBS identification experiments with ANFIS performing marginally better than the MLP with an improvement of 2.1% accuracy. However, GenSoFNN only managed to produce mean accuracy performance of 59.1% in comparison against ANFIS with mean overall performance of 70.0% and MLP with mean overall performance of 68.0%. However, this performance is two times higher than chance guessing with 25% probability of 4-class which empirically demonstrated the validity of the RtSD dataset.

The drawback of the discrete-class DBS identification approach is that the DBS can only be fairly accurately identified and segregated if the DBS are known prior to the training exercise. Such situation is not practical and the data collection process is a challenge because the DBS are very subjective and cannot be discretely defined. Hence, it is more practical to study the correlation between the DBS and the emotion primitives' values of Valence and Arousal based on the re-profiling of the recalibrated data-driven affective space model.

Experimental Result and Discussion for the DBS Re-profiling on the Recalibrated Affective Space Model Using the VA Approach

The VA approach as depicted by the block diagram of Figure 6.1 focuses on the implementation of affective space model as a determinant of a particular emotion. It consists of two important axes that describe the values of Valence and Arousal hence enables further analysis to be conducted. Unlike the discrete-class classification system which restricts the analysis to the 6 basic emotions list [Corn96], the VA approach is more generalized because it has the capability to project other emotion that is not available during training, see Section 6.1.2. In this example emotion primitives' values for calm emotion, which is not available during the training is projected in Quadrant 4. In addition, the employment of the affective space model also provides better explanation for cultural influence. The neutral centroid is proposed as the origin of the recalibrated affective space model that is shifted based on the culture effect on the emotion primitives' values of Valence and Arousal. Such data-driven approach is exclusively tailored to the available speech emotion data for training.

In order to examine the correlation between the DBS and emotion primitives' values of Valence and Arousal, the classifiers are subsequently trained with the culturally-influenced speech emotion datasets; namely: the Berlin, NTU_American and NTU_Asian to represent different cultures of European, American and Asian respectively. The homogeneous and heterogeneous intra-cultural training dataset arrangement are provided to empirically demonstrate the correlation of the DBS and the emotion primitives' values based on the different degree of cultural influence. Homogeneous intra-cultural correlation experiment is referred to the training of single culturally-influenced dataset whereas the heterogeneous intra-cultural correlation experiment is referred to the

combination of two or three dataset in the training data pool to generate the rules to derive the affective space model.

In this analysis, the recalibrated affective space model is generated based on the respective training dataset employed for training. Then, based on the rules derived by the classifier as illustrated in Figure A111.2, the RtSD dataset is re-profiled and the result is mapped onto the data-driven recalibrated affective space model. The result of this RtSD re-profiling is presented in the form of the scatter plot with an ellipse to emphasize the profile distribution for the correlation analysis for better data visualization. Since the centroid (x_e, y_e) for a given set of emotion has been derived using K-means approach as described in Section 6.1.1 and also used to derive the neutral recalibrated region, see Section 6.1.2, this centroid can now be used to analyze the scattered data derived from the DBS based on the Principle Component Analysis (PCA), so that an ellipse can be formed to cover majority of the emotion data [Yild11, Lisb08, Char98].

Consider the following situation, where centroid $u_e = (x_e, y_e)$ for each set of emotion is given by $u_a = (x_a, y_a)$ for anger emotion centroid, $u_h = (x_h, y_h)$ for happiness emotion centroid and $u_s = (x_s, y_s)$ for sad emotion centroid. Assuming a vector of instances that belongs to a set of emotion in the VA affective space model given by I then the centroid subtracted data of I will be:

$$B = I - \mu_e H \quad (6.15)$$

where H is a unity row matrix. Next, the Eigenvector and Eigenvalues of the co-variance matrix for B is derived, where the co-variance matrix B is just the outer product of B with itself, $E(B \cdot B^T)$, since B is always real and is an M x M matrix where M is the total

number of instances for the particular set of emotion. The Eigenvalues and Eigenvectors can be easily calculated using MATLAB function (eig). Ordering the Eigenvalues based on the diagonal of the eigenvectors is required to derive the principal components. Scaling the eigenvectors to a unit circle will help the ellipse plot that represents the probability distribution of the scattered data for a particular set of emotion. Detailed analysis of the DBS correlation and the emotion primitives' values of Valence and Arousal based on the homogeneous intra-cultural data arrangement are presented in Figure A111.5.

(a) ANFIS (NTU_American)

(b) MLP (NTU_American)

(c) ANFIS (Berlin)

(d) MLP (Berlin)

(e) ANFIS (NTU_Asian)

(f) MLP (NTU_Asian)

Figure A111.5 The Homogeneous Intra-Cultural DBS Correlation Experiment based on Different Classifiers and Datasets.

It can be observed from Figure A111.5 that both classifiers produced quite a consistent result in term of the quadrant bias for all DBS with exception of the density of the DBS distribution for each DBS. The size of the each ellipse generated is directly proportional to the eigenvectors calculated for the specific DBS distribution. Hence, explaining the different sizes of the ellipse produced in the result of Figure A111.5.

It is noted that for normal DBS in all of the results in Figure A111.5(a-f) are consistently located in the centre of the affective space model. Such result can be empirically validated the assumption that the driver in normal DBS experiencing neutral emotion which is located at the centre of the recalibrated affective space model. The sleepy DBS is biased towards negative Arousal value especially in NTU_Asian depicted in Figure A111.5(e) and A111.5(f). It can be corresponded to the passive state of the driver. Interestingly, the Valence value of the sleepy state is not exclusively located in the positive region signifying the negative effect of the sleepy DBS to the driver. Such result

indicates the possibility to identify sleepy driver who exhibit the tendency for dangerous situation if he/she is in negative Valence and Arousal values. Once the driver reached certain threshold of negative Valence and Arousal value, an indicator or some warning system can be given to the driver to stop the vehicle and take a break. The result also empirically demonstrates that although sleepy is hypothesized to be located in Quadrant 4 by psychologists' understanding, it may not be welcomed in the case of driving as illustrated in the Figure A111.5. In addition, both distracted and laughed DBS are located in the positive Arousal region showing that the drivers are alert and active in these two states. However, the distribution of the Valence values is more intense in the negative region for distracted DBS and conversely in positive region for laughed DBS. Such result can be explained as the driver feels slightly apprehensive in distracted DBS as he/she needs to switch his/her attention to appropriately respond to the caller and focusing to drive safely while in laughed DBS the drivers are experiencing pleasant feeling which is realized by laughing it out aloud. This result is in-line with psychologists' hypothesis that laughter is often cue for happiness especially in verbal communication without the aid of the face-to-face interaction.

Similar experiments are also carried out with the heterogeneous intra-cultural data arrangement to observe if the results are consistent. Figure A111.6 and A111.7 illustrate the correlation result yielded from mapping the DBS profiling onto the recalibrated affective space model using two culturally-influenced datasets and three culturally-influenced datasets respectively. From the both results, it can be observed that normal DBS consistently occupied the central region of the affective space model. This situation can be clearly observed in Figure A111.7 where three culturally-influenced datasets are combined as the training data pool in comparison against results in Figure A111.6.

(a) ANFIS (Berlin+NTU_American)

(b) MLP (Berlin+NTU_American)

(c) ANFIS (NTU_American+NTU_Asian)

(d) MLP (NTU_American+NTU_Asian)

(e) ANFIS (Berlin+NTU_Asian)

(f) MLP (Berlin +NTU_Asian)

Figure A111.6 The 2-culture Heterogeneous Intra-Cultural DBS Correlation Experiment based on Different Classifiers and Datasets.

Figure A111.7 The 3-culture Heterogeneous Intra-Cultural DBS Correlation Experiment
Based on Different Classifiers (Berlin+NTU_American+NTU_Asian).

Such result indicates that the combination of cultures manages to suppress the intra-cultural effect in such a way that it is more generalized so that it is able to better correlate the DBS and the emotion primitives' values. The sleepy DBS Arousal values are always lower than the distracted and laughed DBS and its negative Valence values can be related to the negative feeling experienced by the driver to stay awake. It can be induced that the lower the Valence and Arousal values of a particular driver, the more dangerous the situation would be as the driver may respond by driving faster and more reckless which can compromise his/her safety. The distracted DBS in the heterogeneous intra-cultural experiments shows that it is distributed in all four quadrants signifying that distraction can give positive as well as negative effect to the driver based on individual perception. This is because driver may welcome slight distraction in order to stay focus but not disturbed for a long period of time as it may affect his/her attention on the driving activity.

Subsequently, the laughed DBS is biased towards Quadrant 2. Such situation can be translated that laughing gives pleasant feeling to the interlocutor hence explaining the positive Valence values. However, laughing is not necessarily deemed as happiness all the time since human use variation type of laughter as sarcasm, polite response or hiding sad emotion based on the individual etiquette and cultural convention which is demonstrated by the different spectrum of Arousal values. It can be summarized that the result shows a strong correlation between DBS and emotion primitives' value in such a way that it can reflect the underlying information that the discrete-class DBS identification system may not be able to show. By employing the VA approach, one can expand the analysis in a more thorough manner and are able to empirically justify the understanding of DBS based on the emotion primitives' values.

Finally, it is important to note that the GenSoFNN classifier is not considered in both homogeneous and heterogeneous intra-cultural correlation experiment analysis due to its low accuracy performance yielded from the VA approach speech emotion recognition system, see Section 6.1.4.