Improvement of Thai Speech Emotion Recognition Using Face Feature Analysis

Igor Stankovic

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Computer Engineering

Prince of Songkla University

2012

Improvement of Thai Speech Emotion Recognition Using Face Feature Analysis

Igor Stankovic

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Computer Engineering

Prince of Songkla University

2012

Thesis Title	Improvement of Thai Speech Emotion Recognition Using Face Feature Analysis
Author	Mr. Igor Stankovic
Major Program	Computer Engineering

Major Advisor :

……..……..
(Assoc. Prof. Dr. Montri Karnjanadecha)

Co-advisor :

……..……..
(Assoc. Prof. Dr. Vlado Delic)

Examining Committee :

.....……Chairperson
(Dr. Anant Choksuriwong)

……..……..
(Assoc. Prof. Dr. Montri Karnjanadecha)

……..……..
(Assoc. Prof. Dr. Vlado Delic)

……..……..
(Ret. Prof. Dr. Miodrag Zlokolica)

The Graduate School, Prince of Songkla University, has approved this thesis as partial fulfillment of the requirements for the Doctor of Philosophy Degree in Computer Engineering.

……..……
(Prof. Dr. Amornrat Phongdara)
Dean of Graduate School

This is to certify that the work here submitted is the result of the candidate's own investigations. Due acknowledgement has been made of any assistance received.

_____ Signature

(Assoc. Prof. Dr. Montri Karnjanadecha)

Major Advisor

_____ Signature

(Mr. Igor Stankovic)

Candidate

I hereby certify that this work has not already been accepted in substance for any degree, and is not being concurrently submitted in candidature for any degree.

_____ Signature

(Mr. Igor Stankovic)

Candidate

Thesis Title Improvement of Thai Speech Emotion Recognition Using Face Feature Analysis
Author Mr. Igor Stankovic
Major Program Computer Engineering
Academic Year 2012

ABSTRACT

Emotions are not usually expressed in the Thai language, because emotional stress would interfere with the speaker's meaning, which makes emotion recognition using Thai speech difficult. Humans use audio cues to recognize fear, anger and disgust, while happiness and surprise seem to be strong "visual" emotions. Our experiments prove this assumption and we present a simple way to combine speech and facial features. The proposed Thai emotion recognition system augments the speech emotion recognition process with face feature analysis, via an audiovisual Thai emotion database. Our speech emotion recognizer is based on calculating mel-frequency cepstral coefficients, zero crossing rate, and energy from short-time speech signals. On the other hand, our face feature analysis, composed of several newly proposed approaches, such as the use of a reference point and middle frames in image sequences, and fully automatic facial landmark point extraction, reaches very good facial expression recognition results and improves the accuracy of the overall system and reduces errors. The combination of speech and facial features show that both vision and hearing play an important part in expressing and recognizing emotions in Thai.

Keywords: Thai speech emotion recognition, facial expression recognition, Thai audiovisual emotion database

ACKNOWLEDGEMENT

Firstly, I would like to thank my advisor, Assoc. Prof. Montri Karnjanadecha, PhD, from the Prince of Songkla University (PSU), for his tireless guidance, attention, patience and understanding during my PhD study at PSU. Also, I am very much grateful to my co-advisor, Assoc. Prof. Vlado Delic, PhD, from the University of Novi Sad (UNS), for always providing me with great ideas and support.

I would like to express my special gratitude to ret. Prof. Miodrag Zlokolica, PhD, from UNS, for introducing me to this amazing country and University, and for his constant support. Without Prof. Zlokolica's help, my dream of studying PhD at PSU would not come true. Also, my thanks go to former Dean Prof. Ilija Cosic, PhD, from the Faculty of Technical Sciences at UNS, former Dean Assoc. Prof. Chusak Limsakul, PhD, from the Faculty of Engineering at PSU, and ret. Prof. Miodrag Zlokolica, PhD, for cherishing PSU-UNS collaboration.

I am very much thankful to PSU and the Thai government for financially supporting my PhD research. I am grateful to Dr. Andrew Davison and Mr. Karlo Poljakovic for help polishing the language of my published papers and my final thesis. Also, my sincere gratitude goes to lecturer Mr. Thanit Phreagathra and his group of students from the Suan Sunandha Rajabhat University in Bangkok, for their help in recording the Thai emotion audiovisual database.

Furthermore, I would like to thank Ms. Bongkot Pruksapong and Ms. Kingkarn Tonnayopas for their unselfish help during my stay in Hat Yai. I am very much grateful to all professors at the Department of Computer Engineering, as well as to all of Thailand and its people for making me feel like home. Their amazing culture, never-ending hospitality and smiles have always given me strength to push further.

I saved my greatest thanks for my family. My parents' and sister's love and guidance through my life are things that I will forever be grateful for. Finally, from the bottom of my heart I thank my lovely wife and son for their patience, love and support, for following me and making my life complete.

Igor Stankovic

CONTENTS

CONTENTS (continued)

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF ABBREVIATIONS

AAM	Active Appearance Model
AU	Action Unit
CK+	The extended Cohn-Kanade database
FACS	Facial Action Coding System
HCI	Human-Computer Interaction
MFCC	Mel-Frequency Cepstral Coefficient
mRMR	minimum-Redundancy & Maximum-Relevance
NN	Neural Network
ROI	Region Of Interest
SVM	Support Vector Machine
ZCR	Zero Crossing Rate

CHAPTER 1

INTRODUCTION

1.1 Background and Rationale

In recent years, the field of Human-Computer Interaction (HCI) has drawn much attention of researchers all over the world. Technology is so advanced these days that the next step is to communicate with machines.

Humans take emotion expression and recognition for granted, but it is actually a complex process that everybody learns from the day they were born. Communication through emotions presents a huge part of everyday communication between people, and emotions are present in almost any interaction. In the near future, it will be impossible to examine any speech recognition or a speech understanding system, or build a facial tracking system without analyzing one of the key elements of communication – emotions.

The field of emotion recognition has shown tremendous potential in many areas, such as the commercial use of emotion recognition in voices in call center queuing systems (Petrushin, 2000). The use of emotion recognition technology has recently been brought under the spotlight in terms of its potential to support countering terrorism with technology. Ball (2011) discusses enhancing border security with automatic emotion recognition. Another possible use of emotion recognition is as an aid to speech understanding (Nicholson *et al*., 1999). They stress that emotion in speech understanding is traditionally treated as "noise", but that a better approach would be to subtract emotions from speech and improve the performance of speech understanding systems. A number of further applications have been proposed, which might benefit from emotion recognition components (Hone and Bhadal, 2004), such as intelligent tutors which change the pace or content of a computer-based tutorial based on sensing the level of interest or puzzlement of the user (Lisetti and Schiano, 2000; Picard, 1997), entertainment applications such as games or interactive movies, where the action changes based on the emotional response of the user (Nakatsu *et al*.,

1999), help systems which detect frustration or confusion and offer appropriate user feedback (Klein *et al*., 2002), and so on.

However, even though there is much work on facial expression recognition, speech recognition and understanding, it seems that emotion recognition, both from speech and facial expressions, is still an unsolved field (without any universal method) of HCI. The lack of a standard, a universally agreed method for emotion recognition perhaps lies in the fact that expressing and recognizing emotions comes so naturally to us, humans, thus being particularly difficult for us to tell what distinguishes one emotion from another. A bimodal system (see Figure 1) for emotion recognition in Thai is presented in this work.

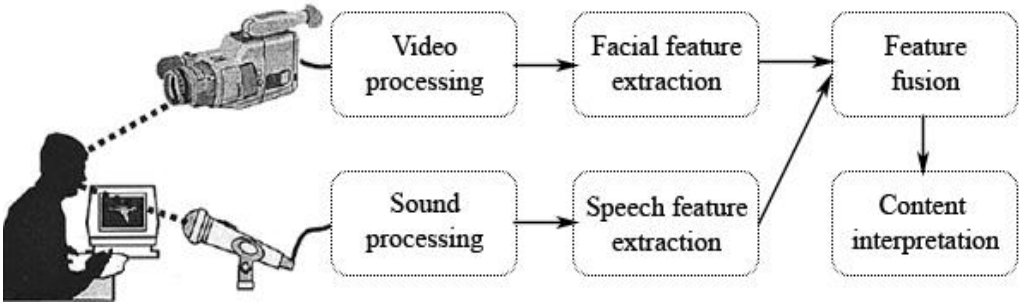

Figure 1. Architecture of bimodal systems.

It is hard to tell what speech characteristics will be useful for recognizing emotions in any language. Language is something that we learn, so it depends on many factors, such as the language family, culture, education etc. Thai is particularly difficult for emotion recognition, both from speech and facial expressions, because Thai people do not stress words as, for instance, Indo-Europeans do. In a way, it is due to the Thai culture, which contains a strong censure against public displays of negative emotions. Also, Thai is a tonal language, it includes five different tones, and so the use of emotion in Thai speech could not only express certain emotional state, but also change the meaning of a word/sentence.

We address this complex problem by augmenting voice analysis with face feature analysis. Our system shows that both vision and hearing play an important part in emotion recognition.

1.2 Review of Literature

There are many papers in the field of emotion recognition. This subchapter presents papers that are the most relevant to our research.

1.2.1 Emotion speech recognition

Williams and Stevens (1972) studied the spectrograms of real emotional speech and compared them with acted speech. They found similarities which suggest the use of acted data. Murray and Arnott (1993) reviewed findings on human vocal emotions. They also constructed a synthesis-by-rule system to incorporate emotions in synthetic speech. However, to date, most works have concentrated on the analysis of human vocal emotions.

Traditional as well as most recent studies have used prosodic information, information related to the rhythmic characteristic of language, such as the pitch, duration, and intensity of the utterance, for recognizing emotion in speech. There are many research projects, differing mostly in the set of features used. For instance, Vogt and André (2009) use pitch, energy, Mel-Frequency Cepstral Coefficients (MFCC), the short-term frequency spectrum, and the harmonics-to-noise ratio for classification. Zhang and Jay Kuo (2001) and Zhang and Zhou (2003) utilize an energy function, the Zero Crossing Rate (ZCR), and the fundamental frequency (F_0).

1.2.2 Facial expression recognition

Much research has also been done on facial expression recognition. For several decades, the field of facial expression recognition has been an important research area, especially in HCI. Ekman and Friesen (1971) discussed six emotions: happiness, sadness, surprise, anger, fear, and disgust, which became the “basic” emotions, used in much related research since.

In one of their later studies, Ekman and Friesen (1977) defined the Facial Action Coding System (FACS) by closely examining facial movements. They concluded that every emotion facial expression is a combination of the movements of several facial muscles. Each basic facial movement is coded as an Action Unit (AU), so that every facial expression can be represented by a group of several AUs.

Facial expression recognition research can be divided roughly into three parts: 1) facial feature extraction, 2) the examination of the changes of those extracted features, and 3) the classification of the gathered information.

Tian *et al*. (2001) used permanent features, such as optical flow, Gabor wavelets, and multi-state models, together with Canny edge detection as transient features. Dornaika and Davoine (2008) chose a candidate face model to track features, while Lucey *et al*. (2010) presented their baseline results in facial feature extraction by utilizing Active Appearance Models (AAMs).

Facial landmarks were extracted by Michel (2003) by employing an *Eyematic FaceTracker* application. Expressions were classified by calculating displacement vectors for each landmark between the first and peak frames in all expression sequences. In his paper, and in the study by Lucey *et al*. (2010), Support Vector Machines (SVMs) were used as classifiers, giving excellent results.

Cohen *et al*. (2003) proposed a new multilevel architecture of hidden Markov models (HMM) for automatic segmentation and recognition of human facial expressions from video sequences. They conducted their research with several classifiers, such as naive Bayes (NB), tree-augmented naive Bayes (TAN), HMM, multilevel HMM, and stochastic structure search (SSS), reaching a maximum accuracy of 83.3% with the TAN classifier and the Cohn-Kanade database. Sebe *et al*. (2007) utilized Bayesian nets, SVM, and decision trees for classification, and reached the accuracy of 93.6%.

There are several facial expression databases, including the MMI Facial Expression database (Pantic *et al*., 2005), the Japanese Female Facial Expression (JAFFE) database, the Cohn-Kanade database (also known as the CMU Pittsburgh database) (Kanade *et al*., 2000), and the improved Cohn-Kanade (CK+) database (Lucey *et al*., 2010).

Many studies, such as Lucey *et al.* (2010), show excellent results for recognizing happiness and surprise, while the other four basic emotions (sadness, anger, fear, and disgust) have much lower results. This is probably due to the more extreme facial deformation and movements used to express happiness and surprise, making them easier to recognize. This recognition gap was stressed by many

researchers, such as Bettadapura (2009), who call for more work towards recognizing all expressions with equal or similar accuracy.

1.2.3 Bimodal systems

There is some work similar to our research, using both audio (speech) and video (facial expressions) to classify emotions. In the well-known study by Pantic and Rothkrantz (2003), all previous work in the field of HCI was surveyed, and a set of recommendations put forward for further HCI development. Chen (2000) combined speech and face features for Spanish and Sinhala (a language spoken in Sri Lanka), with significantly better results when using joint audio-video information.

However, current bimodal recognition systems lack a standard audiovisual database, which makes it hard to compare different systems.

1.3 Objectives

There were several goals that we were focused on during our research in the field of emotion recognition:

A. Collect audiovisual Thai emotion database - this project is the first of its kind for Thai and no emotion audiovisual database for Thai is publicly available, so an audiovisual Thai emotion database had to be recorded prior to our experiments.

B. Develop high accuracy Thai speech emotion recognizer - our next research step was to discover what speech characteristics are most important for distinguishing Thai speech between emotional classes.

C. Develop high accuracy facial expression recognizer - much related work on facial expression recognition show unequal recognition rates of basic emotions. In this thesis we will focus on reducing those recognition gaps.

D. Improve the accuracy of speech emotion recognition using facial expressions - the last goal of this thesis is to improve the results obtained from speech emotion recognizer using information from facial expression part, and to investigate the best way of combining this bimodal information in order to reach as highest final results as possible.

1.4 Scope of the Thesis

The bimodal system for emotion recognition in Thai language is presented in this paper. Because it is a tonal language, emotions are usually not stressed in Thai as much as in any Indo-European language, mostly because any strong emotional emphasis could possibly interfere and change the meaning of the word itself. This fact makes this project complex and unique.

Our idea is that both audio and vision play crucial and inseparable roles in recognizing emotions. Accordingly, in order to successfully recognize emotions, not only voice and its hidden information, but also facial expressions of a speaker were analyzed. This is what this project is based on – building a speech emotion recognition system and improving it with face feature analysis.

Emotion recognition systems that utilize only speech or facial expressions do not represent a realistic way of communication and expressing emotions. Instead of recognizing emotions with the use of all possible sources of emotional state, like humans do, those systems are focused only on specific information, neglecting other sources. Our research presents a more realistic, natural and intuitive way of emotion recognition, using a bimodal system that employs audio and vision for recognition, proving that both hearing and vision are important in recognizing emotions.

In our speech emotion experiments (see subchapter 3.2), after calculating many features and performing feature selection, MFCC+ΔMFCC+ΔΔMFCC, ZCR, and energy proved as a dominant feature set for recognizing emotions in Thai speech. On the other hand, our research on facial expression recognition (see subchapter 3.3) focuses on increasing the recognition rate by tracking facial landmarks more closely. Our system calculates the displacement of each facial landmark to a frame's base-

point in the first and peak frames. The septum, the skin that separates the two nostrils, was chosen as the base-point. As the result, our system is more resistant to head movement errors, thereby increasing the recognition accuracy. Furthermore, emotions are expressed differently over time (Batty and Taylor, 2003) so, our system also utilizes the frames at one third and two thirds along each time sequence. Also, we propose methods for fully automatic facial landmark extraction of several face points relevant to our research. Finally, our proposed technique of combining audio and video information is presented (see chapter 4), and our results show that face feature analysis can greatly improve the speech recognition system, proving that audio and video cannot go separately in emotion recognition systems.

1.5 Organization of the Thesis

This thesis is organized as follows:

Chapter 1 presents an introduction to the topic, a brief literature review, the thesis objective, and the list of materials and equipment used during our research.

In chapter 2, the complete theoretical background is presented. Also, it overviews the used methods, and gives a more detailed explanation of emotions in speech and in facial expressions, and classification techniques.

Our proposed approaches and experimental results are presented step-by-step in chapter 3. Both speech emotion and facial expression recognition results are shown in this chapter.

Chapter 4 contains details on human performance, speech-face result fusion, classification, and final bimodal results.

The conclusion is drawn in chapter 5, with discussion on the thesis in general and its contributions. A short glimpse at our future work is also presented at the end of chapter 5.

CHAPTER 2

THEORETICAL BACKGROUND

This chapter gives a closer look at the theoretical background of the topic, including definition of emotions in speech, facial expressions, well-known speech and image analysis techniques, and methods of classification used in our research.

2.1 Basic Emotions

In their study, Ekman and Friesen (1971) discussed six emotions: happiness, sadness, surprise, anger, fear, and disgust, which became known as the "basic" emotions (Ekman *et al*., 2002), used in much related research since in both speech emotion and facial expression recognition.

A. Happiness

Figure 2. Expression of happiness.

Expression of happiness (see Figure 2) is universally and easily recognized, and is interpreted as enjoyment, pleasure, and friendliness. Examples of happy expressions are the easiest of all emotions to find in photographs, and are produced by people on demand in the absence of any emotion. Happiness is often a rehearsed

expression because it is used so often to hide other emotions and deceive or manipulate other people, thus distinguishing between a real and an acted smile is a new developing topic in the field of facial expression recognition.

B. Sadness

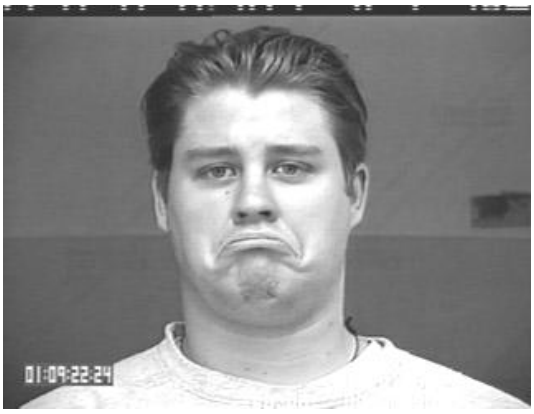

Figure 3. Expression of sadness.

Sadness (see Figure 3) is interpreted as an emotion opposite to that of happiness, but this view is too simple. Sad expressions send messages related to loss, discomfort, pain, helplessness, etc. Many cultures contain a strong censure against public displays of sadness, which makes this facial expression more difficult to express, especially for men. By many psychologists, sad emotion faces are only lower intensity forms of crying faces, developed when we were newly-born.

C. Surprise

Figure 4. Expression of surprise.

The expression of surprise (see Figure 4) is difficult to detect or record in real time. It occurs in response to events that are unexpected, sudden, novel, or amazing. The brief surprise expression is often followed by other expressions that reveal emotion in response to the surprise feeling – emotions such as happiness or fear. For example, most of us have been surprised, perhaps intentionally, by people who appear suddenly or do something unexpected, and elicit surprise, but if the person is a friend, a typical after-emotion is happiness; in case of a stranger, it is fear. A surprise seems to act like a reset switch that shifts our attention.

D. Anger

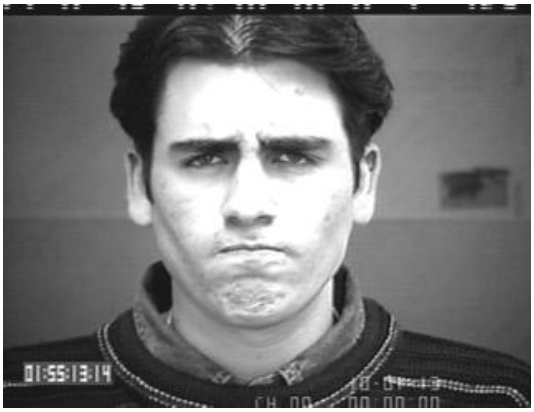

Figure 5. Expression of anger.

The expression of anger (see Figure 5) is found increasingly often in modern society, as daily stresses and frustrations have become a part of everyday life. Anger is a primary concomitant of interpersonal aggression, and its expression means hostility, opposition, and potential attack. A cultural prohibition on expression of anger by women created a distribution of anger expressions that differed between the sexes. Although frequently associated with violence and destruction, anger is probably the most socially constructive emotion as it often underlies the efforts of individuals to shape societies into better, more just environments, and to resist the imposition of injustice and tyranny.

E. Fear

Figure 6. Expression of fear.

The expression of fear (see Figure 6) is not often seen in societies where good personal security is typical. Fear expressions carry information about imminent danger, threat, or likelihood of bodily harm. The experience of fear has an extremely negative felt quality, and is reduced when the threat has been avoided or has passed. Organization of behavior and cognitive functions are adversely affected during fear, as escape becomes the peremptory goal.

F. Disgust

Figure 7. Expression of disgust.

Disgust (see Figure 7) is part of the body's responses to objects that are revolting and unhealthy, such as rotting flesh, faecal matter and insects in food, or other offensive materials that are rejected as unsuitable to eat. Obnoxious smells are

effective in eliciting disgust reactions. Disgust expressions are often displayed in public as a commentary reaction, but these acted reactions have nothing to do with the primal origin of disgust as a rejection of possible unhealthy food. That is why, even thought people feel that disgust expression is easy to express, it is extremely difficult to read disgust emotion.

2.2 Basic Emotions in Speech

Even though there is much research done on the speech recognition topic, we still lack a standard method for capturing emotions from speech. This is probably due to the fact that there are many factors that influence speech/language, such as culture, language group, education etc. Hence, in different languages, different speech characteristics show different importance to recognize and "capture" emotions. By clicking (for hardcopy of this thesis please visit: *http://youtu.be/yZrv_vhSFhk*) on the icons in Table 1, you can hear the difference in expressing speech emotions (happiness, anger, and disgust) in two different languages (German and Thai).

Table 1. Comparison of emotional speech for different languages.

Emotion / Language	*Happiness*	*Anger*	*Disgust*
German			
Thai			

Notice how happiness, anger, and disgust sound much more similar in Thai than in German. This is, firstly, because of the cultural influence – Thai culture does not encourage strong emotional reactions. Secondly, strict tone rules in Thai language disable the usage of strong emotional emphasis, in order not to change the meaning of the words. Instead, Thai people use different lingual particles, such as "khrap" (ครับ),

“kha” (ค่ะ/คะ), “cha” (จ๊ะ/จ๋า), “wa” (วะ), “ha” (ฮะ), “na” (นะ) etc., at the end of sentences to emotionally color their statements. These facts make Thai extremely difficult for any kind of speech emotion recognition. Some well-known techniques could yield much lower recognition rates, while some other speech characteristics that are investigated in this paper, could prove useful in distinguishing Thai speech between emotions more preciously.

There are several methods, such as MFCC, that generally show good results in speech and emotion recognition. Also, some speech characteristics show similar “behavior” in most of the examined languages (see Table 2), and represented a starting point in our research.

Table 2. Characteristics of several speech emotion features for some emotions (Pantic and Rothkrantz, 2003).

	Happiness	*Sadness*	*Anger*	*Fear*
Pitch	**Increase in mean, range, variability**	**Decrease in mean, range**	**Increase in mean, range, variability**	**Increase in mean, range**
Intensity	**Increased**	**Decreased**	**Increased**	**Normal**
Duration (speech rate)	**Increased rate Slow tempo**	**Reduced rate**	**Increased rate Reduced rate**	**Increased rate Reduced rate**
Speech contour	**Descending line**	**Descending line, stressed syllables ascend frequently and rhythmically, irregular up & down inflection**	**Disintegration in pattern and great number of changes in the direction**	**Descending line**

2.2.1 Thai language

Thai (ภาษาไทย – *Phasa Thai*), more precisely Central Thai or Siamese is the national and official language of Thailand and the native language of the Thai people,

Thailand's dominant ethnic group. It is the language taught and used in schools, the one used by the media and for government affairs. According to the 1980 census, an estimated 80 per cent of Thailand population speaks Thai (Comrie, 1990). Outside Bangkok and the central plains, other dialects and languages of the same family coexist with the standard: Northern Thai, Southern Thai, and North-eastern Thai. In addition, Thailand has many minority groups who speak languages that do not belong to the same language family.

Thai belongs to the Tai language family, a subgroup of the Kadai or Kam-Tai family, and it is a tonal and analytic language.

In a language, tone is the use of pitch to distinguish lexical or grammatical meaning. All verbal languages use pitch to express emotional and other linguistic information, and to convey emphasis, contrast, and other features in intonation, but not all languages, like tonal languages, use tones to distinguish words. Tonal languages are extremely common in Africa and East Asia, but rare elsewhere in Asia and in Europe.

2.3 Basic Facial Expressions

As already mentioned in this thesis, Ekman and Friesen (1977) defined FACS by closely examining facial movements. Every emotion facial expression is just a combination of the movements of several facial muscles, and each basic facial

Figure 8. Six basic facial expressions from CK+.

movement is represented as AU. Thus, presence/absence of certain AUs can tell us a lot about an expressed emotion. Six basic facial expressions are presented in Figure 8.

Table 3 (Lucey *et al*., 2010) displays dependence of presence/absence of certain facial movements on facial expressions. There are certain facial movements (AUs) that are present in almost all expressions of one emotion, and absent from expression of all other emotions. For example, happy expression is almost always expressed by pulling lip corners – smile. However, facial expression also depends on many factors (culture, temperament etc.), so expressions differ from one subject to another.

Table 3. Facial expressions explained through facial movements.

	Criteria
Happiness	**Lip corners pulled**
Sadness	**Inner brows raised, brows lowered, and lip corners depressed; or cheek raised and lip corners depressed**
Surprise	**Inner brows raised and outer brows raised, or upper lip raised**
Anger	**Lips tightened and lips pressed**
Fear	**Inner brows raised, outer brows raised, and brows lowered**
Disgust	**Nose wrinkled or upper lip raised**

2.4 Basic Emotions in Bimodal Systems

Bimodal systems contain both speech emotions and facial expression, thus audio and video have their influences on one another. For example, subjects are unable to express surprise as they would express it if only expressing facial gestures without speaking (see Figure 9).

On the left side of Figure 9, the subject (from CK+) is in a surprised state without speaking, while on the right hand side (from our database) is the same expression but with the presence of speech. This influence of emotion speech on face movements makes it more difficult to recognize facial expressions in bimodal

systems, simply because those facial expressions are less acted, less expressive (with slighter deformation of face), hence less informative, and more difficult to recognize.

Figure 9. Facial expression of surprise in absence (left) and presence (right) of speech.

Notice how in Figure 9 on the left, the subject's mouth is open (displaced) much more than the subject's on the right. Of course, due to his speech activity, the subject on the right was not able to express such a strong facial expression as the subject on left. This fact makes the right frame more difficult to recognize, but it also represents a more realistic expression. In a real life situation, it is highly unlikely to find surprise expressed as strongly as presented in the left hand side of Figure 9.

Furthermore, bimodal systems still lack a standard database, so it is particularly difficult to compare those systems.

Examining our work, and many related papers, it is clear that in the near future more systems will focus on employing bimodal information (speech and vision), because it represents a more natural and realistic research environment, which is surely one of the main goals in the field of emotion recognition and in engineering in general.

2.5 Sound Analysis

This subchapter presents the basics of several well-known techniques used in speech and emotion speech processing that are part of our system. Our proposed approaches are explained in subchapter 3.2.

2.5.1 Mel-frequency cepstral coefficients

In sound processing, the Mel-Frequency Cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency. MFCCs are coefficients that collectively make up an MFC. The idea is usually referenced to Mermelstein (1976) and it represents one of the most powerful techniques in speech processing.

Calculation of MFCCs can be divided into several steps:

1. Apply window function.
2. Compute power spectrum (using Fast Fourier Transform – FFT).
3. Apply mel-filter bank.
4. Apply Discrete Cosine Transform (DCT).
5. The MFCCs are the amplitudes of the resulting spectrum.

MFCC are computed (Davis and Mermelstein, 1980) as:

$$MFCC_i = \sum_{k=1}^{N} X_k \cos\left[i(k - \frac{1}{2}) \frac{\pi}{N} \right], i = 1, 2, ..., M \quad (1)$$

where M is the number of cepstrum coefficients, X_k, k=1,2,…, N, represents the log-energy output of the k^{th} filter, and N is the number of triangular band pass filters (with standard value of around 20).

Furthermore, MFCCs also show good results in emotion recognition from speech, and were calculated in our experiments employing the MATLAB's *SpeechCore* toolbox (Omogbenigun, 2007), along with its first and second derivative.

2.5.2 Formants

Formants are defined by Fant (1960) as the spectral peaks of the sound spectrum of the voice, and they represent distinguishing frequencies of human speech. For example, the information that humans require to distinguish between vowels can be represented by frequency of the vowel sounds.

The formant with the lowest frequency is called f_1, the second f_2, and the third f_3. Most often the two first formants, f_1 and f_2, are enough to disambiguate the vowel in speech processing.

In our experiments, formant frequencies (*F)* are calculated by the following formula:

$$F = \frac{Fs}{2\pi} \arctan \frac{im(s)}{re(s)} \tag{2}$$

where *Fs*, *im*(*s*) and *re*(*s*) respectively represent the sampling frequency, imaginary and the real part of sound signal *s*.

The formants bandwidth, *Bw*, is represented by:

$$Bw = \frac{Fs}{\pi} \log(|s|) \tag{3}$$

where *Fs*, and *s* represent the sampling frequency and the sound signal.

Because formants show excellent results in speech processing, they have been tested in our system as speech emotion features, and were calculated using the MATLAB's toolbox for tracking formants (Ghosh, 2001), developed in the *SpeechLab* at the Boston University, USA.

2.5.3 Pitch

Along with duration, timbre, and loudness, pitch is a major property of tones. It is closely related to frequency, but represents a more subjective approach. For instance, as they oscillate, sound waves do not contain pitch, and can be measured only in frequencies. However, the subjective sensation in which a human listener assigns that tone to a position in the musical scale represents pitch of that sound wave.

Human perception of sound is logarithmic. Accordingly, many researchers tried to calculate pitch as a number. In our experiments, Hu and Zahorian's (2008) algorithm is used to track and calculate pitch.

2.5.4 Zero crossing rate

ZCR is the rate at which a sign in the signal changes from positive to negative, or vice versa, that is the rate at which a signal crosses/cuts the zero-line. This feature is often used in speech recognition and it is also suitable for emotion recognition

because its hidden information includes speech rate and the speed of changing tempo. These elements are useful since they differ between emotional speeches.

ZCR's formula is defined by Chen (1988) as follows:

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} II\{s_t s_{t-1} < 0\} \tag{4}$$

where s_t and s_{t-1} are signals at time t and $t-1$ respectively, and function II is 1 if the argument is true or 0 otherwise.

Figure 10 explains a simple approach to calculate ZCR over the whole sound signal.

```
Window_size = Sampling_freq * Window_length_in_ms / 1000

Movement    = Sampling_freq *(Window_length_in_ms –
                            - Overlapping_length_in_ms) / 1000

Iteration = (size(Sound) / Movement) – 2

Pos = 1

FOR i = 1 TO Iteration

    Sum = 0

    FOR j = Pos + 1 TO Pos + Window_size

        Sum = Sum + (Sound(j) * Sound(j-1) < 0)

    END FOR

    Mat(i) = Sum / (Window_size - 1)

    Pos = Pos + Movement
END FOR
```

Figure 10. Pseudo-code of calculating ZCR over a whole sound signal.

The final *Mat* vector represents ZCR of every 30 ms segment over the whole sound file (*Sound*), and was used as a speech emotion feature in our experiments.

2.5.5 Short-time energy

When an object vibrates, it basically moves air particles and produces sound. Sound energy is associated with the energy that is produced by those vibrations.

In our experiments, short-time energy of an audio signal, *E*, proven as a useful feature in speech emotion recognition, is calculated (Gustavsen, 2007) by the following formula:

$$E(i) = \sum_{n=0}^{N-1} s^2(n), (0 \le i \le I-1) \tag{5}$$

where *N* is the length of the sound-window, *I* is the number of windows, and *s* sound signal. The MATLAB code (Sharma, 2009) was used for calculating this feature.

2.5.6 Minimum-redundancy and maximum-relevance

The minimum-Redundancy and Maximum-Relevance (mRMR) is a feature selection technique proposed by Peng *et al*. (2005). Many studies on feature selection stress that selecting a good set of features is probably the most important step in reducing the overall error. Peng et al. (2005) researched how to select good features according to the maximal statistical dependency criterion based on mutual information. Their experiment shows good results on several datasets classified via SVMs, Naive Bayes, and linear discriminate analysis.

The given variables *x* and *y*, their mutual information, *I*, is defined as:

$$I(x;y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}\, dxdy \tag{6}$$

where *p(x)* and *p(y)* represent their marginal probabilities, and *p(x,y)* their joint probability distribution.

The idea of minimum redundancy is to find features that are maximally dissimilar, and making those features a more important representation of the whole database. The minimum redundancy condition, *W*, is:

$$\min \mathbf{W}_I, \mathbf{W}_I = \frac{1}{|\mathbf{s}|^2} \sum_{i,j \in \mathbf{s}} I(i,j) \tag{7}$$

where *I(i,j)* is the mutual information between features *i* and *j*, and **s** is a set of features.

A maximum relevance between classes, *V*, is measured:

$$\max \mathbf{V}_I, \mathbf{V}_I = \frac{1}{|\mathbf{s}|} \sum_{i \in \mathbf{s}} I(h,i) \tag{8}$$

where **s** is a set of features, and *I*(*h*,*i*) mutual information between target classes.

The mRMR feature set is obtained by optimizing the conditions in equation 7 and equation 8 simultaneously. Optimization of these two conditions requires combining them into a single criterion function, hence two simplest approaches of finding final criteria were calculated as:

$$\max(\mathbf{V}_I - \mathbf{W}_I) \tag{9}$$

$$\max\left(\frac{\mathbf{V}_I}{\mathbf{W}_I}\right) \tag{10}$$

These criteria are called mutual information difference (MID), and mutual information quotient (MIQ) techniques.

In their several publications (Peng *et al*., 2005; Ding and Peng, 2005; Ding and Peng, 2003; Zhou and Peng, 2007), authors have developed a MATLAB code for their proposed feature selection method. That code was employed in our system.

2.6 Image Analysis

Some well-known methods in image processing used in facial expression recognition, and in our system, are listed and briefly explained in this subchapter. Our proposed techniques and approaches on how to improve facial expression recognition are given in details in subchapter 3.3.

2.6.1 Active appearance model

The AAM technique was developed by Cootes *et al*. (1998), and it represents a model that yields photo-realistic objects. AAM is a fast, robust method of interpreting face images, and have become one of the most used methods as a part of facial expression recognition.

Figure 11. Example of face image (Cootes et al., 1998) labeled with 122 landmarks.

Appearance model is defined as a combination of shape and texture, where shape represents a set of locations in an image and the texture intensity patch of an image. Figure 11 displays an example of a labeled face image – facial landmarks represent the model's shape, and intensity of color in pixels of the model's texture.

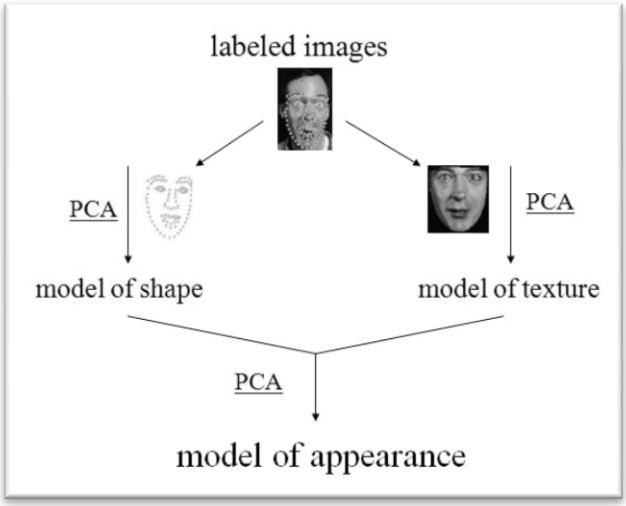

Figure 12. Modeling appearance with AAM.

If training set of shapes is $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$, than model of shape is calculated as:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P_s b_s} \tag{11}$$

where $\bar{\mathbf{x}}$, $\mathbf{P_s}$, and $\mathbf{b_s}$ represents the mean shape, matrix of eigenvectors that define the model, and a vector of model parameters respectively.

If the training set of textures is $\mathbf{g} = \{g_1, g_2, \ldots, g_n\}$, then the model of texture is calculated as:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g b_g} \tag{12}$$

where $\bar{\mathbf{g}}$, $\mathbf{P_g}$, and $\mathbf{b_g}$ represents the mean shape, matrix of eigenvectors that define the model, and a vector of model parameters respectively.

In combining two models, the joint parameter vector, **b**, is calculated as:

$$\mathbf{b} = \begin{pmatrix} \mathbf{W_s b_s} \\ \mathbf{b_g} \end{pmatrix} \tag{13}$$

where $\mathbf{W_s}$ is a diagonal matrix of weights for each shape parameter.

In the training set, each (x_i, g_i) pair is obtained:

$$\mathbf{b}_i = \begin{pmatrix} \mathbf{W_s b_{si}} \\ \mathbf{b_{gi}} \end{pmatrix} = \begin{pmatrix} \mathbf{W_s P_s}^T (\mathbf{x_i} - \bar{\mathbf{x}}) \\ \mathbf{P_s}^T (\mathbf{g_i} - \bar{\mathbf{g}}) \end{pmatrix} \tag{14}$$

and after applying principle component analysis (PCA) to the training set $\mathbf{b} = \{b_1, b_2, \ldots, b_n\}$, the model for the parameters is:

$$\mathbf{b} = \mathbf{P_c c}, \mathbf{P_c} = [\mathbf{P_{cs}} \mid \mathbf{P_{cg}}]^T \tag{15}$$

Finally, the combination model is calculated:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Q_s c}, \mathbf{g} = \bar{\mathbf{g}} + \mathbf{Q_g c} \tag{16}$$

where:

$$\mathbf{Q_s} = \mathbf{P_s W_s P_{cs}}, \mathbf{Q_g} = \mathbf{P_s P_{cg}} \tag{17}$$

2.6.2 Face detector

Face Detector is a MATLAB toolbox for tracking facial points developed by Aldrian *et al*. (2009) at the University of Leoben, Austria. The main focus of their project was eye and pupil detection, but it also tracks the mouth and nose region as presented in Figure 13.

Firstly, a face is detected by a modified OpenCV's Viola and Jones (2001) publicly available implementation (Krishna, 2008). OpenCV has an effective face detector function named *cvHaarDetectObjects* that is based on calculating Haar-like features (Papageorgiou *et al*., 1998; Viola and Jones, 2001). It finds rectangular regions in the given image that are likely to contain objects the cascade has been trained for and returns those regions as a sequence of rectangles. The function scans the image several times at different scales. After it has proceeded and collected the candidate's rectangles (regions that passed the classifier cascade), it groups them and returns a sequence of average rectangles for each suitably large group (Bradski *et al*., 2008).

Figure 13. Highlighted facial regions extracted using Face Detector (Aldrian et al., 2009).

In 2008, the *MEX*-file, which calls this OpenCV function from MATLAB, was published (Krishna, 2008), and used in experiments by Aldrian *et al*. (2009). In their research, after selecting a face, several facial regions are located (see Figure 13) using face geometry information where these regions should be on a selected face.

Final selected regions on a face were used as a starting point of our experiments. Our experimental setups and results are presented in chapter 3 and chapter 4.

2.6.3 Gabor filters

Gabor filters are linear filters used for edge detection, and their representation of frequency and orientation are similar to those of the human visual system. A 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave, and Gabor filters are self-similar, meaning that all filters can be generated from one mother wavelet by dilation and rotation (Daugman, 1985).

Gabor filters, *g*, are calculated by the following formula:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\cos\left(2\pi\frac{x'}{\lambda} + \psi\right) \tag{18}$$

where:

$$x' = x\cos\theta + y\sin\theta \tag{19}$$

$$y' = -x\sin\theta + y\cos\theta \tag{20}$$

In our system, Gabor filters were used to emphasize a stronger image contrast if a face was not successfully localized in a certain frame. Our experiments show that using Gabor filters, a face can be more easily localized. In the frames where the use of Gabor filter was needed, after finding face coordinates, the original image was loaded back and further processed in our proposed methods. Accordingly, Gabor filters in our system present only a part of the aid in finding Region Of Interest (ROI) for a face.

2.6.4 Canny edge detector

In 1986, John Canny presented a new edge detection algorithm (Canny, 1986) to detect a wide range of edges in images, but also produced a computational theory of edge detection explaining why his technique works. In his work, he focused on reaching three main goals:

- *good detection* – the algorithm should mark as many real edges in the image as possible.
- *good localization* – edges marked should be as close as possible to the edge in the real image.
- *minimal response* – a given edge in the image should only be marked once, and where possible, image noise should not create false edges.

The Canny algorithm comes with several parameters, of which its threshold is the most relevant for our research. A threshold set too high can miss important information, but on the other hand, a threshold set too low will falsely identify irrelevant information (such as noise) as important. It is difficult to give a generic threshold that works well on all images.

MATLAB's build-in function *edge*, with parameter *'canny'* and different thresholds, was used throughout our experiments.

2.7 Classification Methods

Two classification techniques are employed in our work. To compare our facial expression recognition results with the results in several related papers, SVM was utilized as the classifier in our experiments. On the other hand, speech segments, part of the facial expression classification, and our final results are classified via Neural Network (NN). Here are the short descriptions of both classification methods used in our system.

2.7.1 Neural networks

NNs, or more precisely artificial neural networks, are composed of interconnecting artificial neurons that are constructed in a way to mimic the properties of biological neurons. NN are often used to solve artificial intelligence problems without necessarily creating a model of a real biological system, which is highly complex.

There are three types of neuron layers in basic NN architecture: input, hidden, and output. In feed-forward networks, the signal flow is from input to output units, strictly in a feed-forward direction, with back-propagation correction of networks weights. Example of simple NN is presented in Figure 14, while work by Hopfield (1982) gives more detailed explanation of the idea behind NN.

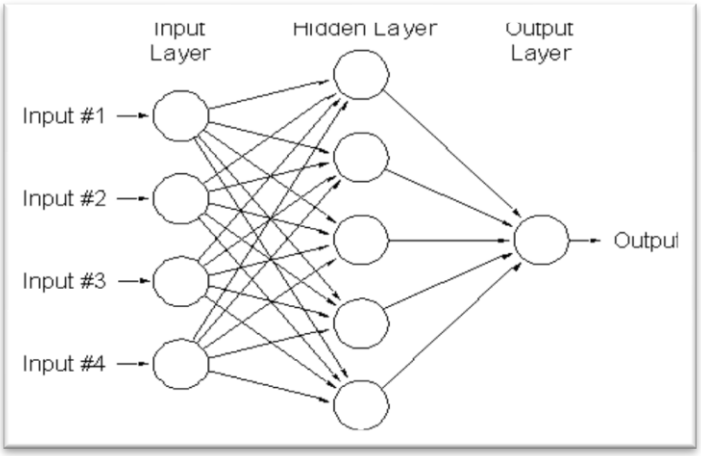

Figure 14. Simple neural network.

In our system, feed-forward back-propagation NNs were used, with different number of nodes in a hidden layer. Our experimental setups and results are presented in chapter 3 and chapter 4.

2.7.2 Support vector machines

The concept of SVM is associated with Vladimir Vapnik, but the complete idea was presented in the work of Cortes and Vapnik (1995).

SVM represents a method for analyzing data, recognizing patterns, and classification. In our experiment, linear SVM was employed, which predicts for each

input to the two classes of which it belongs. A linear SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear linear gap that is as wide as possible (see Figure 15). New points from input data are then mapped into that same space and classified based on which side of the gap they fall on.

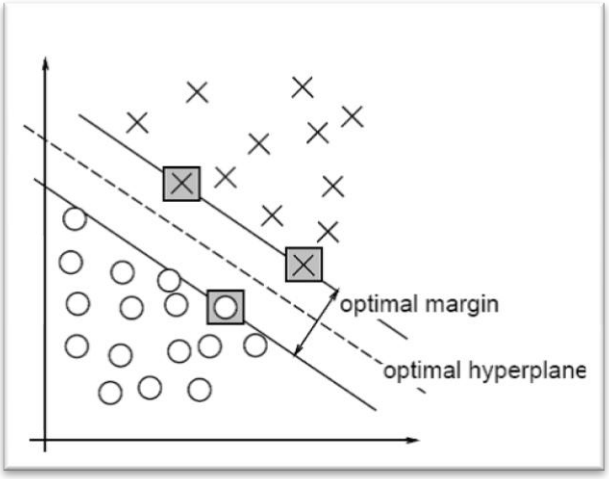

Figure 15. SVM process of dividing and finding a maximum gap between two categories.

In our experiments, a MATLAB's *libsvm* toolbox (Chang and Lin, 2011) was used.

CHAPTER 3

PROPOSED METHODOLOGY AND EXPERIMENTAL RESULTS

This chapter gives detailed descriptions of our proposed methods and approaches utilized in our system.

Firstly, our newly recorded Thai emotion audiovisual database is presented (see subchapter 3.1). Then approach to speech emotion recognition in Thai is shown, followed by speech emotion recognition results on our database (see subchapter 3.2). Several proposed methods are presented step-by-step in subchapter 3.3, with results on CK+ and our database.

Our system's architecture is displayed in Figure 16.

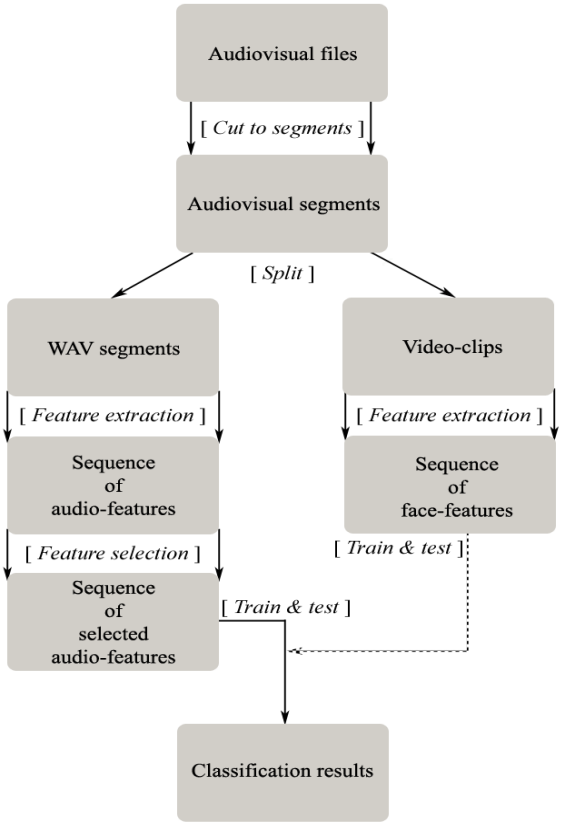

Figure 16. Our audiovisual emotion recognition process.

3.1 Thai Emotion Audiovisual Database

There is no publicly available emotion audiovisual database for Thai, so we recorded one, structurally similar to an existing database for Serbian (Jovičić *et al*., 2004).

There are three main types of database for emotion recognition utilizing either acted emotions, neutral spontaneous emotions, or elicited emotions. Using actors has always been a popular approach, but has been criticized because the emotions expressed by the actors seem to differ from real-life emotions. Recent studies show that the problem with acted databases is not the use of actors and acted emotions, but rather the elicitation method employed in the recording processes. Aside from this concern, the best results are usually obtained from acted emotion databases because they contain strong emotional expressions.

Our acted emotions database utilized six drama students and was recorded in a professional recording studio at the Suan Sunandha Rajabhat University (มหาวิทยาลัยราชภัฏสวนสุนันทา), Bangkok, Thailand.

Being the first work of its kind for Thai, only six basic emotions (Ekman and Friesen, 1971) were examined – happiness, sadness, surprise, anger, fear, and disgust.

In most languages only 800-1,000 words are used in basic everyday conversation (Lewis, 2009), so the 1,000 most frequently used Thai words were used as the recording corpus of our database. One of the possible future works based on this research is to improve Thai speech understanding, so using most common Thai words will certainly play a big role in it.

There are many lists of most commonly used Thai words, such as Haas (1964), NECTEC's Linguistics and Knowledge Science Laboratory – LINKS and Chulalongkorn University's Orchid Corpus (Sornlertlamvanich *et al*., 1997), and a Barrow's Thai-English-Thai CD dictionary (Barrow, 2010). Table 4 shows the first 20 most frequently used words in Thai compiled from those sources. Our list of the most frequent words was selected and recorded from the largest corpus, the LINKS database with corpus of 416,000 Thai words. Our final list of words was cut into 20 groups, resulting with around 50 words in each group. This number of words is best for emotion expression in Thai – less number of words in group would be too short

for a subject to reach a desirable emotional state, and more words per group would possibly lead to boredom. These 20 groups were then recorded by six subjects in six emotions, and every session was saved into a separate video clip, with a resolution of 320 x 240 pixels and audio sampling rate of 16 kHz.

Table 4. Twenty most frequently used Thai words.[*]

Source name (corpus size) / **Number**	*Haas* *(27,000 words)*	***LINKS*** ***(416,000 words)***	*Orchid* *(275,000 words)*
1	เป็น /pen/	**การ /kan/**	ที่ /thi/
2	ใจ /chai/	**และ /lae/**	เป็น /pen/
3	ไม่ /mai/	**ใน /nai/**	จะ /cha/
4	การ /kan/	**ที่ /thi/**	การ /kan/
5	น้ำ /nam/	**มี /mi/**	ไม่ /mai/
6	หน้า /na/	**ของ /khong/**	มี /mi/
7	ตัว /tua/	**เป็น /pen/**	ใน /nai/
8	ตา /ta/	**จะ /cha/**	ของ /khong/
9	ที่ /thi/	**ได้ /dai/**	และ /lae/
10	ไป /pai/	**ให้ /hai/**	ได้ /dai/
11	ลูก /luk/	**ความ /khwam/**	ไป /pai/
12	กัน /kan/	**ไม่ /mai/**	ให้ /hai/
13	มี /mi/	**ว่า /wa/**	ว่า /wa/
14	ผู้ /phu/	**พัฒนา /patthana/**	มา /ma/
15	พระ /phra/	**ใช้ /chai/**	ก็ /go/
16	ทาง /thang/	**ก็ /go/**	คน /khon/
17	ความ /khwam/	**เรา /rao/**	แล้ว /laeo/
18	ไม้ /mai/	**นี้ /ni/**	ความ /khwam/
19	หัว /hua/	**หรือ /rue/**	กับ /kap/
20	ขึ้น /khuen/	**กับ /kap/**	อยู่ /yu/

[*] *Thai words transcribed in the Latin alphabet via the official Royal Thai General System of Transcription (RTGS)*

3.1.1 Database processing

Firstly, the pre-processing stage of setting up the database involved deleting 28 repeated words out of the selected 1,000 most commonly used Thai words, resulting in a list of 972 words that were recorded by six speakers in six emotions. Final list can be found in the appendix part at the end of this thesis.

Secondly, after the recordings were done, as part of post-processing, several poorly recorded sessions and sessions where a subject was interrupted were deleted.

Finally, following related database work, such as the Berlin database (Burkhardt *et al*., 2005), the audio of each recording was played to five Thai native speakers. If their recognition was above a certain threshold, a recording was kept, otherwise it was deleted. The idea is that if humans cannot correctly recognize emotion from a played file, a computer cannot do it either, and that file should be

Table 5. Inventory of our final dataset.

	Subject (from word no. – to word no.)
Happiness	Subject no. 2 (689 – 972) Subject no. 5 (1 – 405)
Sadness	Subject no. 1 (1 – 972) Subject no. 3 (86 – 90) Subject no. 5 (401 – 710)
Surprise	Subject no. 2 (1 – 400) Subject no. 4 (463 – 972)
Anger	Subject no. 2 (1 – 972) Subject no. 3 (1 – 972)
Fear	Subject no. 1 (1 – 62) Subject no. 5 (1 – 972)
Disgust	Subject no. 1 (1 – 62) Subject no. 2 (1 – 972)
Total	*6898 words*

discarded because it can only be misleading and confusing to any classifier. With the threshold of 0.6 (60%), the final dataset of 6,898 words (~75 min) was selected. Table 5 shows the final dataset inventory used in our system.

In our final selected dataset, 60% of dataset was used for training, 20% for validation, and 20% for testing, without any data set overlapping.

3.2 Emotion Speech Recognition

The full feature set for our Thai speech emotion system contains 180 features, including features such as median, minimum and maximum of a signal, variance, F_0, formants, pitch, ZCR, energy, speed (Δ) and acceleration ($\Delta\Delta$) of changing those features, as well as techniques such as MFCC with the first derivative of MFCC (ΔMFCC), and the second derivative of MFCC ($\Delta\Delta$MFCC), but our following step was to reduce that number and select only several features that are the most important for categorizing emotions in Thai.

3.2.1 Feature selection

Much related work stresses that proper feature selection is probably the most important in reducing the final error in any classification process. Doing calculations with a huge number of features requires a lot of processing, and many of these features will be irrelevant and even misleading for a specific problem. Consequently, only the most important features, which clearly distinguish between emotion groups, should be selected. Three techniques for selecting the features were combined – our ranking method, and two methods from mRMR – MID and MIQ.

Firstly, all features are ranked by dependence among files in each emotion group. A feature increases its rank if all files that belong to the same emotion are similar in that feature range, and different in the same feature range from files in other emotion groups. This means that features which separate files into emotion groups will be ranked higher.

If a certain feature shows different range and behavior of files in the same emotional group, it means that this feature is probably not good for distinguishing

files into different groups. If, however, that feature shows similar range in most of the files (say more than 60% of the files) of that emotion, this feature is selected. When this process is done on all features, all selected features are then further tested between every pair of emotions. For example, all files in which subjects expressed happiness and all files with sadness are tested feature by feature (see Figure 17). If they show similarities in range, that feature is ranked low for distinguishing between, in this case, happiness and sadness; if they show ranges that overlap, the feature is ranked depending on the percentage of overlapping (the lower the percentage, the higher the ranking); finally, if they show completely different ranges, without overlapping, that feature is ranked top. It means that selected top ranking features show range similarities in 60% (threshold that showed best performance in our system) of files inside the same emotional group, but different ranges comparing to files in at least one of other emotional groups.

```
FOR fea = 1 TO Num_of_features

      IF min(Em_group1(fea)) > max(Em_group2(fea)) OR
         min(Em_group2(fea)) > max(Em_group1(fea))

            feature_rank = ['HIGH RANK']
            RETURN

      ELSE
      Mutual_info = (max(Em_group1(fea)) - min(Em_group2(fea))) /
                  / (max(Em_group2(fea)) - min(Em_group1(fea)))

            IF Mutual_info > 1
                  Mutual_info = 1 / Mutual_info

      IF Mutual_info > Threshold

            feature_rank = ['MIDDLE RANK']
            RETURN

      ELSE
            feature_rank = ['LOW RANK']

      END IF
END FOR
```

Figure 17. Pseudo-code of our feature ranking process.

Our second selection approach was the mRMR (see subchapter 2.5.6) feature selection method (Peng *et al*., 2005), utilizing MID and MIQ. Using only the first technique, the selected features often give lower accuracy results than expected, because the features are often correlated. This means that some features are not really relevant, but show high relevance because they are dependent on another "real" feature. By introducing mRMR, this problem was solved.

Our proposed feature ranking, MID and MIQ techniques using equation 9 and equation 10 respectively, were performed, and mean value of features ranking position from all three techniques were calculated. Finally, top ranking features were chosen as selected features.

All our experiments in feature selection, combining these three techniques, show similar output, resulting in the selection of ZCR and energy as the most relevant features for our Thai audiovisual database. Also, feature selection was performed only on speech features, in order to reduce the huge number of features to the several most important ones. Face features were classified separately, and were used only if the classification of a segment with a speech feature was under a certain threshold. The joint classification process is explained in subchapter 4.2.

3.2.2 Experimental results

After computing 180 speech features, our feature selection techniques resulted in choosing ZCR and energy as the most important features in distinguishing Thai speech signals into emotional groups.

The audio signals are cut into 30 ms segments with 20 ms overlap, so each new segment only advances 10 ms from its previous segment, resulting in 100 segments per second. Our speech emotion feature extraction is then performed on each sound-segment.

All speech emotion recognition experiments were performed on the whole selected dataset (6,898 words) using 2-layer (different number of input nodes) feed-forward back-propagation NN classifier with one hidden layer (different number of hidden nodes) and six outputs (six basic emotions), implemented using MATLAB's

NN toolbox. Furthermore, classification decisions were made based on: a) a single frame, and b) a group's most frequent emotion.

In classification based on a single frame, classification results represent an NN output of all sound-segments. However, this classification decision alone would not be enough, as it does not represent a natural way of recognizing emotional states, thus our second classification decision was based on the most frequent emotion in a group.

Already classified segments are grouped together into groups of various lengths (from 2.6 s to 15 s). Groups are determined depending on recording files (see subchapter 3.1) from our database – maximum length of group is set to be 15 s, but if the file ends, so does the group. File ending means change of subject, emotion, and/or word group. Inside one determined group, classification results of all sound-segments in that group were analyzed and placed into a 6 x 1 matrix. Each field of the matrix

```
Happiness = 0
Sadness   = 0
Surprise  = 0
Anger     = 0
Fear      = 0
Disgust   = 0

FOR i = 1 TO Number_of_sound_segments_in_group

      IF      Emotion(Group_element(i)) == 1
                  Happiness = Happiness + 1
      ELSE IF Emotion(Group_element(i)) == 2
                  Sadness   = Sadness   + 1
      ELSE IF Emotion(Group_element(i)) == 3
                  Surprise  = Surprise  + 1
      ELSE IF Emotion(Group_element(i)) == 4
                  Anger     = Anger     + 1
      ELSE IF Emotion(Group_element(i)) == 5
                  Fear      = Fear      + 1
      ELSE IF Emotion(Group_element(i)) == 6
                  Disgust   = Disgust   + 1
      END IF

      Final_vector = [Happiness, Sadness, Surprise, Anger, Fear,
                      Disgust] / Number_of_sound_segments_in_group

      Winning_group_emotion = Index_max(Final_vector)

END FOR
```

Figure 18. Pseudo-code of finding a group's top emotion.

represents the frequency of recognized emotions from segments, where fields 1, 2, 3, 4, 5, and 6 correspond to emotions of happiness, sadness, surprise, anger, fear, and disgust respectively. For example, if one group contains 300 sound-segments, and if there are 100, 50, 50, 50, 50, and 0 segments that were recognized by NN as happiness, sadness, surprise, anger, fear, and disgust respectively, then this group will be classified into happiness, as happiness is the most frequent emotion in it.

Finally, each group matrix was scaled so that it sums to 1, making it very easy to combine with other classification results, such as the facial expression recognition results. Referring to our previous example, the 6-value-string [100 50 50 50 50 0] would be divided by the sum of those elements (in this case 300), resulting in a final recognition string for that group – [0.33 0.16 0.16 0.16 0.16 0]. The whole process is presented in the MATLAB oriented pseudo-code in Figure 18. The final number of groups was 305.

Because MFCCs generally show great performance in speech and emotion speech processing, our first experiment on speech emotion recognition was based on using MFCCs as speech emotion features, with different number of coefficients, together with first and second MFCC derivatives. Accordingly, the test dataset consisted of different number of features (depending on the number of MFCCs) with 448,413 rows (~75 min), after cutting the whole dataset into 30 ms sound-segments. Testing was performed on the whole database using NN. Table 6 presents the overall speech emotion recognition results depending on the number of MFCCs, with classification decisions based on both single frames and the groups' top emotion. As Table 6 shows, best performance on our database was achieved using 13 MFCCs.

Table 6. Speech emotion recognition depending on the number of MFCCs (+ΔMFCC +ΔΔMFCC), with decisions based on single frames and the groups' top emotion.

	Overall accuracy [%]					
No. of MFCCs / **Decision**	*2*	*5*	*10*	*12*	***13***	*15*
Single frame	40.2	43.3	45.5	44.8	**46.6**	45
Group	49.7	55.9	62.2	61.8	**64.5**	61.8

However, due to the enormous size of our database, the final dataset with 448,413 rows was too large to be trained and tested in one NN, constantly running into MATLAB's "out of memory" problem. In a correspondence with a few MATLAB developers, we have decided to split our dataset into several blocks, train them separately, and calculate final results as a mean value of all block-results, where block size was determined as a maximum size that can be trained in a single NN on our notebook. This approach yields results most similar to those gained using one NN.

Table 7. Speech emotion recognition using MFCC (13 coefficients) +ΔMFCC+ΔΔMFCC based on single frames.

Accuracy [%]	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**8.7**	9.3	14.7	37.1	15.4	14.9
Sad.	0.5	**61.1**	1.9	8.2	13	15.5
Sur.	1.5	14.4	**22.6**	34.9	2.2	24.5
Ang.	1.6	10.8	8.5	**65.3**	3.1	10.7
Fear	2.6	7.8	3.4	16	**63.5**	6.7
Dis.	0.9	17.1	13	27.4	2.6	**39**

Table 7 presents detailed results using MFCC (13 coefficients), ΔMFCC and ΔΔMFCC as a feature set, with classification decisions based on a single frame, and Table 8 using the same feature set but based on a groups' most frequent emotion. The

Table 8. Speech emotion recognition using MFCC (13 coefficients) +ΔMFCC+ΔΔMFCC based on the groups' top emotion.

Accuracy [%]	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**0**	0	0	83.9	16.1	0
Sad.	0	**72.9**	0	0	27.1	0
Sur.	0	0	**1.9**	80.8	0	17.3
Ang.	0	0	0	**100**	0	0
Fear	0	0	0	0	**100**	0
Dis.	0	3.5	0	19.3	0	**77.2**

main diagonal represents correct classification (happiness classified as happiness, sadness classified as sadness, etc.), while all other fields are part of a recognition error. Each row sums up to 100%, which means that rows represent target emotions, and columns do so with recognized emotions.

The overall accuracies for these two experiments are 46.6% (single frame) and 64.5% (group), as already presented in Table 6.

Our final experiments were conducted using MFCC (13 coefficients), ΔMFCC (13 coefficients), ΔΔMFCC (13 coefficients), and two features that were selected as the most relevant in classifying Thai speech emotions – short-time ZCR

Table 9. Speech emotion recognition using MFCC (13 coefficients) +ΔMFCC+ΔΔMFCC, ZCR, and energy, based on single frames.

Accuracy [%]	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**18.7**	4.5	21.2	26.5	16.4	12.7
Sad.	2.5	**57.1**	1.4	6.5	13.2	19.3
Sur.	2.5	4.3	**36.6**	27.1	3.2	26.3
Ang.	3.1	4.2	13.3	**63.4**	3	13
Fear	3.2	6.9	4.6	12.2	**69.2**	4
Dis.	2.4	9.9	17.2	15.8	3.9	**50.8**

Table 10. Speech emotion recognition using MFCC (13 coefficients) +ΔMFCC+ΔΔMFCC, ZCR, and energy, based on the groups' top emotion.

Accuracy [%]	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**12.9**	0	12.9	38.7	19.4	16.3
Sad.	0	**75**	0	0	25	0
Sur.	0	3.9	**42.3**	17.3	0	36.5
Ang.	0	0	0	**97.5**	0	2.5
Fear	0	0	0	0	**100**	0
Dis.	0	3.5	7	12.3	0	**77.2**

and short-time energy. Table 9 and Table 10 give final results with decisions based on single frames and the groups' most frequent emotion.

After utilizing ZCR and short-time energy, overall accuracy in groups increased from 64.5% to 72.4%, hence the error was reduced by 22% comparing to results obtained using only MFCC+ΔMFCC+ΔΔMFCC.

If Table 10 is analyzed more closely, it is clear that recognition for sadness and disgust yielded good results, almost perfect results for anger and fear, but very low recognition rates for happiness and surprise. This makes sense because happiness and surprise are considered as strong "visual" emotions that are very easily recognized from facial expressions. Subchapter 3.3.1.4 presents results from several papers on facial expression recognition, and all of them show excellent results in recognizing happiness and surprise, while yielding lower results for other basic emotions. On the other hand, our final results on speech emotion recognition shows completely opposite results. Fear, that is considered one of the most difficult facial expressions to recognize, yielded perfect results in our speech emotion experiments.

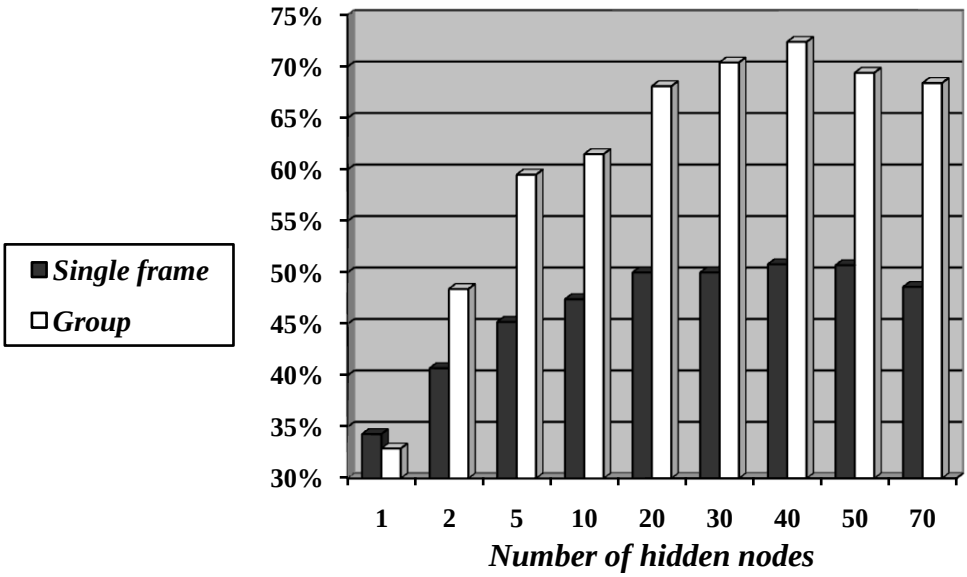

Figure 19. Emotion speech recognition rate depending on the number of hidden nodes in a hidden layer, obtained using 13 MFCCs (+ΔMFCC+ΔΔMFCC), ZCR, and energy, with decisions based on single frames and the groups' top emotion.

This set of results justifies our approach and proves that both hearing and vision play an important role in recognizing emotions.

Figure 19 displays the recognition rates, based on both single frames and the groups' top emotion decisions, depending on the number of hidden nodes in a hidden layer, and calculated using final 41-feature-set (1 feature – short-time ZCR, 1 feature – short-time energy, 39 features – MFCC+ΔMFCC+ΔΔMFCC). The highest accuracy of 72.4%, displayed in details in Table 10, was achieved using 40 hidden nodes, while chapter 4 presents this set of speech emotion results compared to human performance.

Our next step was to build a facial expression recognizer that will improve our speech emotion results, especially for happiness and surprise.

3.3 Facial Expression Recognition

The Cohn-Kanade (Kanade *et al*., 2000) database was released in 2000, and soon became one of the most frequently used databases for face recognition algorithm development and evaluation. To address a few concerns, the authors released the CK+ database in 2010 (Lucey *et al*., 2010). The number of sequences was increased by 22% and the number of subjects by 27%. The database was tested using AAMs and a linear SVM classifier (using a leave-one-subject-out cross-validation method) for both AU and emotion detection. The resulting emotion and AU labels, together with the extended image data and tracked facial landmarks, were made publicly available.

In this thesis we present several approaches for facial feature extraction that showed excellent performances on both CK+ and our database. Due to the fact that our database is audiovisual, experimental setups on CK+, i.e. the facial expressions database, and our database were a bit different. However, all image processing steps are identical.

Our proposed methods for facial expression recognition include:

- *Head movements correction*
- *Neurophysiologic approach*
- *Movements along the X- and Y-axes*
- *Automatic middle eyebrows point, the septum, and lip-corners detection*

3.3.1 Introducing a reference point and middle frames

3.3.1.1 Head movements correction

In the CK+ database, using AAMs, 68 facial landmarks were extracted (Lucey *et al*., 2010), and utilized as the starting point for our experiments. Our calculations

Figure 20. Example of an image sequence from CK+.

were based on tracking the displacements of landmarks from the first neutral frame to the last peak frame, in an image sequence representing one emotional expression (see Figure 20). This method has been utilized in several related studies, such as Michel (2003) (see Figure 21), giving good results, but with high recognition accuracy differences between the basic facial expressions. Our first approach tracks the displacements more closely, increasing the accuracy.

Figure 21. The neutral and the peak frames producing vectors of displacements.

Our assumption was that part of the recognition error in tracking landmark displacements comes from head movements that occur between the neutral and the peak frame. Out of the 68 extracted facial landmarks (Lucey *et al*., 2010), we chose one landmark to present the base-point in every frame. This base-point should be fixed during emotion expressions, so it can be employed to detect head movement, without additional facial movements. Also, it should be located in the middle of the face, so that the calculations are equally sensitive to movements by all the other landmarks. In the six basic facial expressions, the nose region seems to move the least

Figure 22. Facial landmarks with the base-point (the septum).

during facial expressions. Therefore, by being in the middle of the facial region, the point between the two nostrils (called the septum), was chosen as our base-point in each frame (see Figure 22). It corresponds to the 34th landmark in the standard group of 68 extracted landmarks. The displacements of other landmarks were calculated with reference to this base-point, using the Euclidian distance:

$$d_i = \sqrt{(x_{i,P} - x_{34,P})^2 + (y_{i,P} - y_{34,P})^2} - \sqrt{(x_{i,N} - x_{34,N})^2 + (y_{i,N} - y_{34,N})^2} \quad (21)$$

where d_i is the subtraction of two Euclidian distances for a landmark i. The first distance is from the base-point (the 34th landmark) to the landmark in the peak (P) frame, while the second distance subtracts the base-point from the landmark in the neutral (N) frame.

By introducing the base-point in each frame, our calculations became more resistant to head movements that may appear between the first and the peak frame. Figure 23 shows an example of feature vectors from the previous method and from our proposed method.

The two feature vectors at the top of Figure 23, which represent anger and fear, were obtained by the previous method. They were classified incorrectly, because their feature patterns have a similar shape, which is confusing for a classifier. The same feature vectors, obtained using our proposed method, are displayed at the bottom of Figure 23. After removing the feature calculation "noise", caused by head movements, the differences in these two vectors are much more obvious. Using our method, these two vectors were classified correctly.

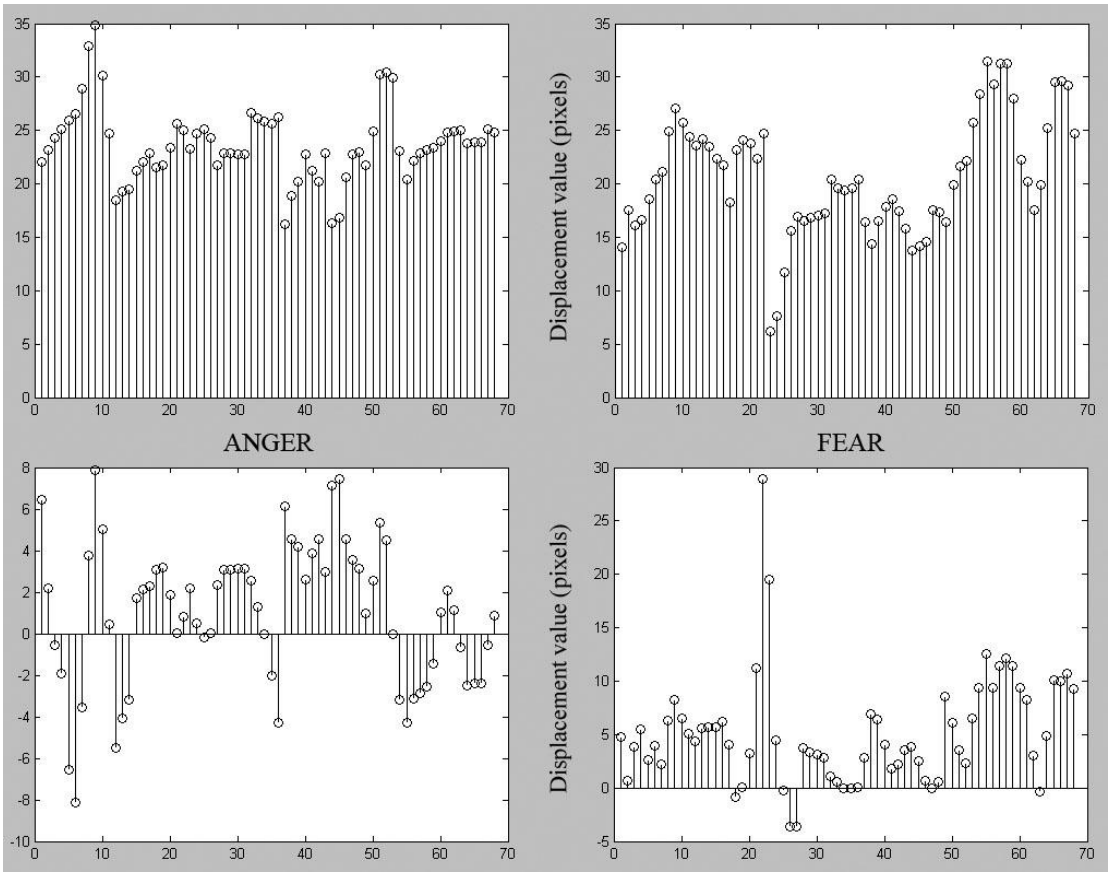

Figure 23. Two facial expression feature vectors for anger and fear, using the previous method (top), have a similar shape, leading to an incorrect classification. The same feature vectors, calculated using our proposed method (bottom), are classified correctly due to their much different shape.

3.3.1.2 Neurophysiologic approach

The important aspect of the speed with which facial emotions are processed and expressed has only recently been investigated in neurophysiology (Batty and Taylor, 2003). It seems that different emotions use different brain regions, so the time to process and express emotions differs. This difference is much more obvious when examining positive emotions (i.e. happiness, surprise) to negative ones (i.e. sadness, anger, fear), but it also differs from emotion to emotion. Since emotions are expressed over different durations, the time (the number of frames) needed for expressing an emotion in a time sequence was added to our system.

Furthermore, the movement of facial muscles is different for each emotion. For example, happiness is universally and easily recognized, and is interpreted as enjoyment, pleasure, and friendliness (Ekman *et al*., 2002). Happy expressions are frequently produced by people on demand in the absence of any real emotion, or to hide other emotions, or to deceive or manipulate. On the other hand, many cultures contain a strong censure against public displays of negative emotions, such as sadness and anger. Also, some emotions, such as fear, are not often seen in societies where personal security is typical.

As a consequence of neurophysiologic brain structure, social influences, and differences in usage frequency, different emotions trigger facial muscle movements in

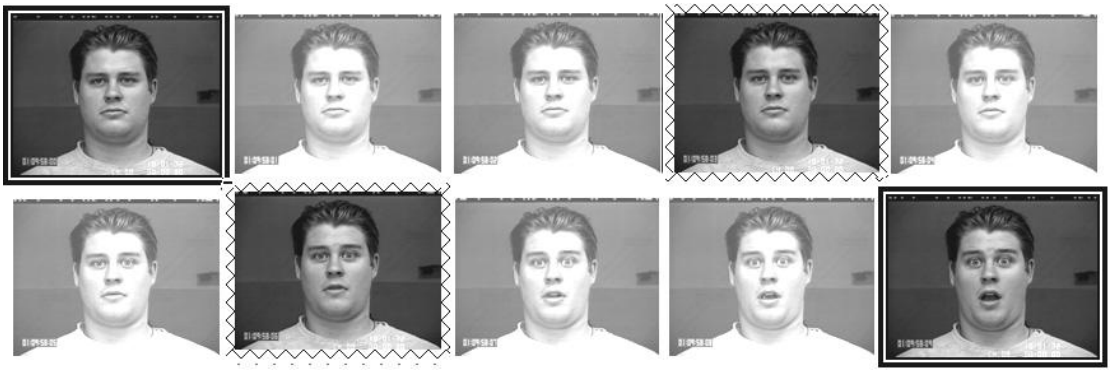

Figure 24. An example image sequence for expressing surprise taken from the CK+ database (straight line framed – the first and the peak frames; curved line framed – additional middle frames).

different ways over time. This observation is why our system examines time sequence frames at the one third and two thirds points of every sequence (see Figure 24).

Lucey *et al*. (2010) only examines the first (neutral) and the last (peak) expression frames, and does not take into account the facial changes that lead to the peak expression. However, middle frames were used during a visual inspection of the clip, to determine whether the expression is a good representation of an emotion. In our work, two frames, one third and two thirds along the image time sequence, proved an excellent way to distinguish precisely between emotions, and reduce the overall error.

Similarly to the 68-feature vectors described early, the distances from the first middle frame (M1) and the neutral frame, as well as the second middle frame (M2) and M1 were calculated using the Euclidian distance formulas:

$$d1_i = \sqrt{(x_{i,M1} - x_{34,M1})^2 + (y_{i,M1} - y_{34,M1})^2} - \sqrt{(x_{i,N} - x_{34,N})^2 + (y_{i,N} - y_{34,N})^2} \quad (22)$$

$$d2_i = \sqrt{(x_{i,M2} - x_{34,M2})^2 + (y_{i,M2} - y_{34,M2})^2} - \sqrt{(x_{i,M1} - x_{34,M1})^2 + (y_{i,M1} - y_{34,M1})^2} \quad (23)$$

In equation 22, $d1_i$ subtracts the Euclidian distance of each *i* landmark from the base-point (the 34th landmark) in M1 from the distance of the landmark from N. Equation 23 for $d2_i$ is similar but subtracts the distance from M2 and M1.

3.3.1.3 Movements along the x- and y-axes

After making our system more resistant to head movement errors and adding middle frames, it showed great improvements for recognizing all emotions, but errors were still present, especially when detecting sadness. Our data shows that happiness and sadness yields similar distances (from the neutral to the peak frames) for the points at the edge of the mouth. Unfortunately, those points are the most important for recognizing happiness, and so sadness can be confused with happiness. While smiling, edge points drastically change along the x-axis, with almost no changes in the y-axis. However, sadness shows small changes on both axes, which made the Euclidian distances for happiness and sadness at the edge points almost the same. To

address this problem, we factored the landmark changes on the x- and y-axes into our calculations, using:

$$dx_i = x_{P,i} - x_{N,i} - (x_{P,34} - x_{N,34}) \quad (24)$$

$$dy_i = y_{P,i} - y_{N,i} - (y_{P,34} - y_{N,34}) \quad (25)$$

dx_i and dy_i subtract movements in the x- and y-axes of landmark *i* from the base-point (the 34th landmark) in the peak (P) frame and the neutral (N) frame.

These feature vectors improve the results, giving perfect results for detecting happiness and sadness on CK+, and improving the recognition of the other emotions.

3.3.1.4 Experimental setup on CK+

The CK+ database (Lucey *et al*., 2010) includes 593 image sequences using 123 subjects. The sequences vary in duration (from 6 to 71 frames) and each one presents a subject's face from the first (neutral) frame to the peak formation of the given facial expression. An image sequence for a surprised expression is displayed in Figure 24.

Table 11. Frequency of emotions in the CK+ database.

Emotion	***Number of sequences***
Happiness	69
Sadness	28
Surprise	83
Anger	45
Fear	25
Disgust	59
Contempt	18
Total	*327*

Lucey *et al*. (2010) decided to evaluate the sequences by studying the middle frames in each image sequence. They concluded that only 327 of those 593 sequences represent a natural emotion expression. The other sequences, which failed their criterion, were discarded from their experiments. The final inventory of their selection process is given in Table 11.

Our work focuses only on the six basic facial expressions, so contempt sequences are not included, resulting in 327 – 18 = 309 samples for our experiments.

Our feature vector for each image sequence comprises 342 features comprised from the following:

- 68 features for the displacements (for each of the 68 landmarks) between the natural (N) and the peak (P) frames.
- 68 features for the displacements between the frame at the one third point of the sequence (M1), and N.
- 68 features for the displacements between the frame at the two thirds point of the sequence (M2), and M1.
- 2 x 68 features for movements along the x- and y-axes from N to P.
- 1 feature for the number of frames in the sequence.
- 1 feature for the presence/absence of nose wrinkles (Lucey *et al*., 2010).

Figure 25 presents examples of the feature vectors for each basic emotion. Our feature dataset is represented as a 309 x 342 matrix, relating each of the sequences to their features.

In several other studies, such as Lucey *et al*. (2010) and Michel (2003), linear SVM has produced good results, and proved to be simple and effective for classifying facial expressions. Motivated by those studies, we tested our 309 x 342 dataset with a linear one-versus-all (i.e. anger vs. not-anger, happiness vs. not-happiness, etc.) multi-class SVM classifier, utilizing the leave-one-subject-out cross-validations method. The MATLAB's *libsvm* toolbox (Chang and Lin, 2011) was used in our experiments.

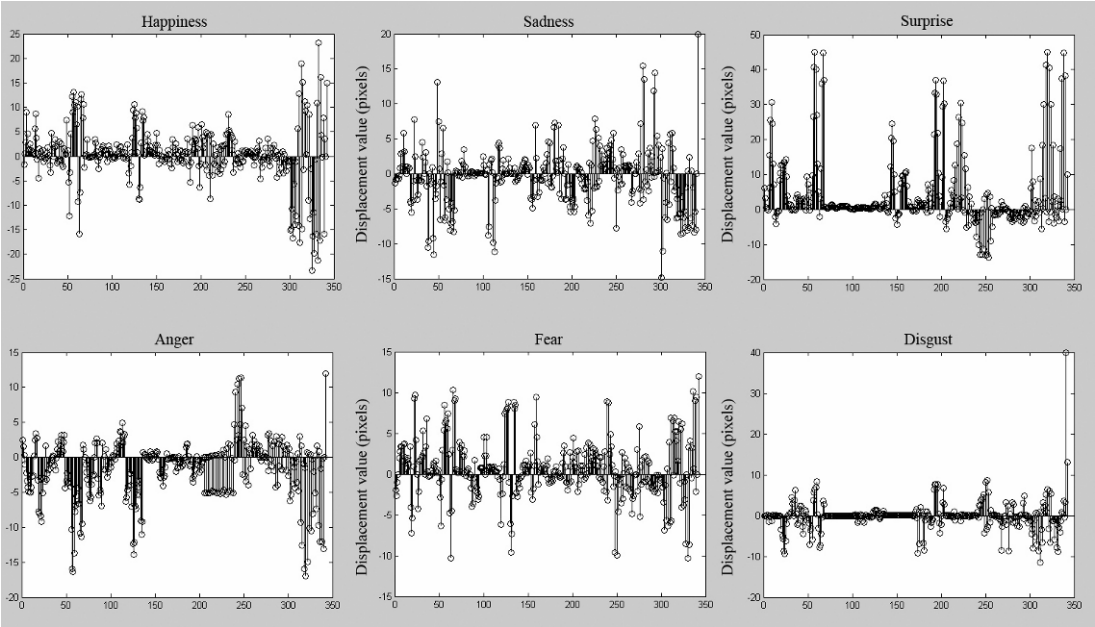

Figure 25. Example of facial expression feature vectors for each basic expression obtained by our proposed approaches.

3.3.1.5 Results on CK+

Our main focus was improving recognition results for the six basic facial expressions: happiness, sadness, surprise, anger, fear, and disgust. Lucey *et al*. (2010), however, added contempt as a new emotion, and used 118 different training and test sets for emotion detection. Removing contempt from our experiments raised the recognition rate slightly, but our usage of bigger training and test sets, made the rates drop, producing results similar to those of Lucey *et al*. (2010).

The use of frame base-points made our system more resistant to head movement errors. Also, adding middle frames (at the one third and two thirds point of every sequence) increased the distinction between emotions. Furthermore, our observation of movements in the x- and y-axes reduced the confusion between some emotions. Finally, additional features that represent a sequence duration (using the number of frames) for an emotion, and a nose wrinkle detector, as calculated by Lucey *et al*. (2010), improved our results by making a more significant difference between positive and negative emotions. A summary of our accuracy results are

displayed in Table 12, with the rows and columns: happiness, sadness, surprise, anger, fear, and disgust, respectably in both directions. The main diagonal represents correctly classified samples (happiness classified as happiness, sadness classified as sadness, etc.), while the other fields represent the system error.

Table 12. Our summarized facial expression recognition results.

***Accuracy* [%]**	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**100**	0	0	0	0	0
Sad.	0	**100**	0	0	0	0
Sur.	0	1.2	**97.6**	0	1.2	0
Ang.	0	0	0	**93.3**	0	6.7
Fear	8.0	4.0	0	0	**88.0**	0
Dis.	0	1.7	0	1.7	0	**96.6**

3.3.1.6 Result comparison

Table 13 compares the accuracy results obtained with our method and with those from several related papers.

Visutsak (2005) uses the displacements of only eight points from the lower part of the face for classifying basic expressions. His results are good for detecting happiness and surprise, but his 8-feature vectors are not informative enough for the other four emotions, yielding lower results, and a total accuracy of 74.5%.

Michel (2003) employed the *Eyematic FaceTracker* application to extract 22 facial landmarks, but outperforms our results only for the recognition of surprise, due to his smaller training and test sets (20 samples per emotion). His final recognition rate is 86.3%.

Sebe *et al*. (2007) presents an emotions database composed from spontaneous reactions caught using hidden cameras. They utilized Bayesian networks, k-nearest neighbor (kNN), and SVM classifiers, producing an accuracy average of 93.6%. Their results outperform ours when recognizing fear (see Table 13), probably due to the usage of spontaneous expressions in their experiments. Fear is particularly difficult to

express on demand, so their approach captures facial movements that are absent in acted databases such as CK+. However, our total accuracy, using a simple method, outperforms their results.

Table 13. Facial expression recognition result comparison.

		Accuracy [%]						
Paper	*Database*	***Hap.***	***Sad.***	***Sur.***	***Ang.***	***Fear***	***Dis.***	*Total*
Proposed method	***CK+***	**100**	**100**	97.6	**93.3**	88.0	**96.6**	***96.8***
Lucey et al. (2010)	***CK+***	**100**	68.0	96.0	75.0	65.2	94.7	*88.6*
Sebe et al. (2007)	***Their own***	95.7	92.0	88.7	91.2	**94.7**	85.6	*93.6*
Michel (2003)	***CMU***	95.3	85,6	**98.8**	78.4	76.2	83.9	*86.3*
Visutsak (2005)	***JAFFE***	91.5	61.0	97.5	67.7	66.7	62.3	*74.5*

Lucey *et al*. (2010) extracted 68 landmarks with AAMs, and produced excellent results for detecting happiness and surprise. Their nose wrinkle detector, which was also utilized in our system, meant that the recognition of disgust had high accuracy. However, sadness, anger, and fear have much lower recognition rates than in our work. They reached final recognition rate of 88.6%.

Our results show that our system produces excellent recognition rates for all the six basic emotions. In particular, the recognition of sadness, anger, and fear are drastically improved with our approach. Our overall accuracy outperforms results of other works, with total recognition rate of 96.8%.

3.3.2 Automatic facial points extraction

MATLAB's VideoIO toolbox (Dalley, 2006) gives easy, flexible, and efficient read/write access to video files in MATLAB, and was used to extract video frames from our recordings. Then, facial features were tracked and analyzed from sequences of images.

3.3.2.1 Selecting important landmarks and regions of interest

In order to make our system completely automatic, our experiments do not employ AAM approach as in study by Lucey *et al*. (2010), which is a semi-automatic approach that requires manual notation of facial landmarks. Instead, we propose fully automatic facial landmarks detection methods that show excellent results on CK+ and our Thai emotion audiovisual database.

In several related studies on facial expression recognition, it is concluded that eyebrows carries biggest piece of information on expressed emotion. For example, unless it is a part of the blinking stage, movements of eyes are almost always followed by even more prominent eyebrow movements, thus it seems that eyes are, comparing to eyebrows, completely unnecessary in correct recognition of facial expressions. This assumption was put on test (see Table 14), and showed that, from the group of 68 landmark points, obtained by Lucey *et al*. (2010) on CK+, using only points on eyebrows yields great results for happiness and surprise, which are two emotions that we were focused on to improve from speech emotion recognition results. Furthermore, Table 14 also shows that from all points on eyebrows, middle eyebrow points show best results.

Table 14. Facial expression recognition results using different eyebrow landmarks.

Selected points	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
5 points on each eyebrow	59.4	0	97.6	0	0	5.1
3 points on each eyebrow	58	0	97.6	0	0	5.1
Mid-point on each eyebrow	75.4	0	96.4	28.9	0	39

Table 14 proves that using only the middle eyebrow points can yield very good results in detecting happiness and surprise. These preliminary tests, as a part of our point selection phase, were conducted as described in subchapter 3.3.1, but using only several eyebrow points out of the set of 68 points, and without using the middle frame in calculations, as proposed earlier on. The final results on facial expression recognition, after extracting facial points using our proposed approaches, are presented in subchapter 3.3.2.5 (on CK+) and subchapter 3.3.2.6 (on our database).

Beside eyebrows, points on mouth can reveal a lot about expression. However, due to the fact that subjects in our database move lips while speaking, standard approach of tracking mouth points displacement in each frame would not be useful, because of the speech interference. For example, pronouncing vowel “o” in happy expression could be misclassified as an expression of surprise, and so on. Of all the mouth points, lip-corners are especially important for emotion recognition, as they are used to detect smiles (lip-corners pulled), in happy expressions, but can also detect sadness (lip-corners depressed), and surprise (upper lip raised, making lip-corners closer). Thus, in our calculations the mean value of displacement of lip-corners was used. Since all recorded subjects read the same text of 972 most commonly used Thai words, the mean movement of the lip-corners throughout the whole recording file should differ from emotion to emotion. For example, the mean value of all the recordings of happy expressions should show wider lip-corner positions than in the recordings of other emotions.

To this list of points, the septum was added, so that the head movement corrections (see subchapter 3.3.1), explained earlier on in this thesis, can be calculated and involved in our proposed method.

The following text in this subchapter gives a detailed explanation of techniques used for obtaining five facial landmarks (2 x middle eyebrow point, 2 x lip-corner points, 1 x septum) used in our system, followed by the results on CK+ and our Thai emotion audiovisual database.

As a starting point, in each frame ROI for face, eyes, and mouth were selected using Face Detector. Figure 26 shows face detection results on CK+ and our database.

Figure 26. Face detection results for CK+ and our database.

Due to the much less controlled illumination in our database, in some frames we experienced a problem with the Face Detector, resulting with unrecognized facial ROI. To address this problem, if the recognition was unsuccessful, the Gabor mask, which lit a frame and made a stronger contrast, was introduced. The Gabor mask was re-used, making a face region lighter, until the recognition of ROI of a face was successful.

3.3.2.2 Detecting middle eyebrows points

Two already selected ROIs for eyes were expanded so that they cover both eyes and areas under the eyes close to the mouth region. These two regions were then removed from image in order not to influence the recognition process of eyebrows and the septum, and two regions above were selected as ROI of the left and the right eyebrow. Figure 27 presents an example of masking eyes inside the already selected ROI of a face, and show selected ROI for both eyebrows.

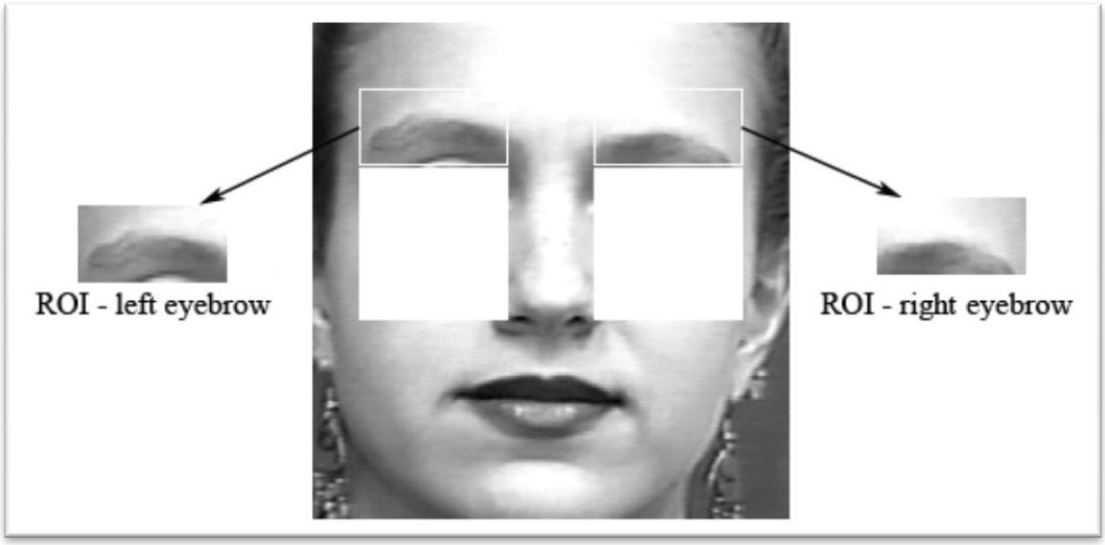

Figure 27. Selected ROIs for the left and right eyebrows, with masked eyes.

On the selected ROIs of both eyebrows, the Canny edge detector with a variable threshold was employed. In the Canny edge detector method, the threshold can vary from 0 to 1, where smaller values increase the detector's sensitivity (see Figure 28).

Figure 28. Canny edge detector outputs for a right eyebrow image, with values of 0.1, 0.3, 0.5, 0.7, and 0.9.

Increasing the threshold reduces the detector's sensitivity to noise, at the expense of losing some of the finer details in the image. Therefore, a number of edges that are recognized depend on an image itself. The same threshold can detect only one short edge in one image and detect many edges in the other, depending on the image quality, contrast, size etc. Our proposed solution is to use a variable threshold.

```
Percentage_of_white_pixels = 0
Threshold = 0.99

WHILE Percentage_of_white_pixels < 2 AND Threshold > 0

    Edges = Detect_edges(Eyebrow_image,'Canny',Threshold)

    Image_without_vertical_edges = Median_filter(Edges)

    Reconnected_edges_by_45 =
           Dilate_image(Image_without_vertical_edges,'line',45)
    Reconnected_edges_by_135 =
           Dilate_image(Reconnected_edges_by_45,'line',135))

    Image = Reconnected_edge_by_135

    Percentage_of_white_pixels = `
           100 * (No_of_white_pixels(Image) / No_of_pixels(Image))

    Threshold = Threshold - 0.005
    %lowering down the threshold until we reach 2% of white pixels
END WHILE
```

Figure 29. Pseudo-code of edge detection processing on eyebrow images.

Prior to our experiment with a different threshold in the Canny edge detector, edges along x-axis in output edge detector frames were removed using the MATLAB's *medfilt2* function. This step filtered out some unnecessary vertical edges in the frames, such as hair and wrinkles. Unfortunately, due to the curvy shape of the eyebrow, removing some vertical edges also influenced this main eyebrow edge by

disconnecting it in several places. Using the MATLAB's *imdilate* function, disconnected edges are successfully re-connected in 45° and 135°. Figure 29 presents the pseudo-code of this filtering process, while outputs of the edge detector in each proposed step is shown in Figure 30.

With vertical edges removed, and the main eyebrow edge re-connected, the best edge detector threshold was calculated. In our experiments, both on CK+ and our database, best detection results were shown when the percentage of edges (white pixels) per detector's output was around 2%. Accordingly, the threshold started with the highest value of 1, and was decreased until the area of the edges was just above the 2% of the whole frame (see Figure 29).

Figure 30. Example of left eyebrow edge detection with: a) Canny edge detector, b) after removing vertical edges, and c) after re-connecting the edges.

As displayed in Figure 30, vertical edges became only small disconnected areas, while the main eyebrow-edge presents the largest shape. By finding a center of

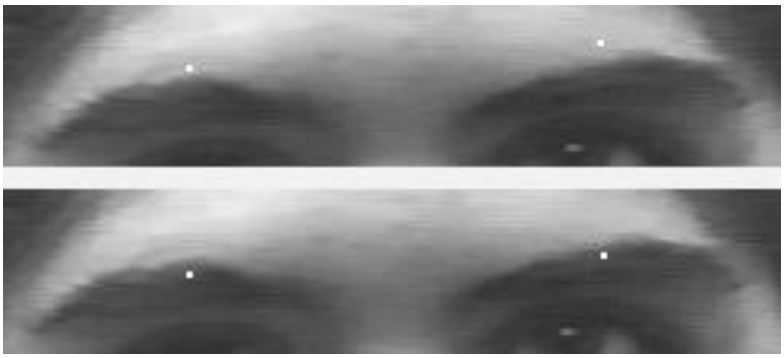

Figure 31. Example of middle eyebrow points detection using AAM (top), and our method (bottom).

the largest area in a final image, we extracted the middle eyebrow point that was then used in our experiments on facial expression recognition. Figure 31 shows examples of middle eyebrow landmark detection results on CK+, with extracted points (Lucey *et al*., 2010) using AAM (top), and using our proposed method (bottom).

3.3.2.3 Detecting lip-corner points

After selected ROI of mouth, the Canny edge detector was performed to extract lip-edges. Figure 32 presents an example of selected ROI of mouth on CK+ (left) and on our database (right). Due to the smaller picture resolution in our database (320 x 240 pixels), comparing to CK+ (640 x 480 pixels or 640 x 490 pixels), it was much more difficult to find edges in the selected mouth-region of our database.

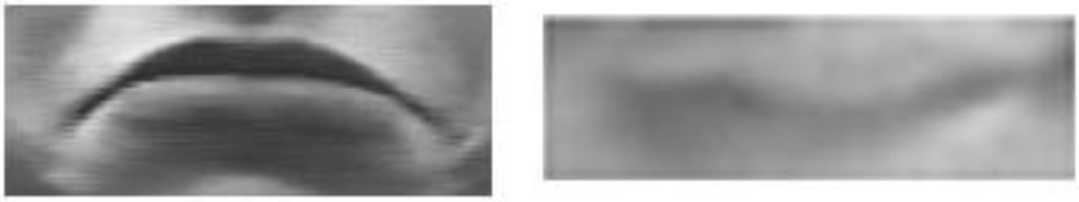

Figure 32. Example of ROI of mouth on CK+ and our database.

More importantly, unlike in eyebrow edge detection, a variable Canny detector threshold in detecting lip-edges showed poor results. The reason for this is that there are not many new edges that can occur with eyebrow movements, apart from some wrinkles that almost completely vanish removing the edges in x-axis (as already explained above), while lip movements can reveal new edges, such as teeth. For example, teeth seem to be more easily detected with the edge detector than lips. Accordingly, if the fixed threshold of 2% of white pixels is used in the image sequence that represents a happy expression, the first neutral frame will correctly detect lip edges, but the second "smiling" frame will not. In the second frame, detector lowers down the variable threshold detecting edges that corresponds to location of teeth. It stops detection when white pixels reach 2% of overall frame, before even starting to detect lips, resulting with only edges for several teeth. Example of this problem is shown in Figure 33.

Figure 33. Lips edge detection on neutral and peak frames for a happy expression from CK+. Edges in the first frame were found correctly, but the presence of teeth in the second frame produced edges that resulted in an incorrect detection of lips.

Hence, in detecting lip edges, the fixed threshold of 0.7 yielded best results and was used for experiment on CK+, and a threshold of 0.4 for our Thai audiovisual database. These thresholds on these two databases detect lips in all frames regardless of presence/absence of teeth and other additional edges.

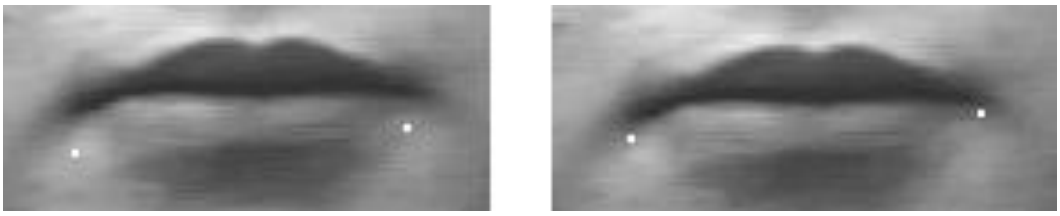

Figure 34. Lip-corner points extraction for CK+ utilizing AAM (left) and our method (right).

Similarly to eyebrow point extraction, the largest area in recognized lip-edges frame was selected, and the lowest and highest value in x-axis were chosen as two lip-corners. Results of lip-corners extraction on CK+ using AAM and our proposed method are shown in Figure 34.

3.3.2.4 Detecting septum

ROI for septum was chosen, in y-axis, as everything under the eyes masks and above ROI for mouth, and in x-axis as the region between two eye masks (see Figure 27). Then, a mean value over y-axis was calculated on that region, with the darkest part pointing at the line that goes through both nostrils and the septum (see Figure 35), which represents a piece of a skin between two nostrils. The region just under the

nose is almost always the darkest part, because of the nose shadow and the presence of two nostrils, which are two darkest parts of that region. Figure 35 clearly shows this fact.

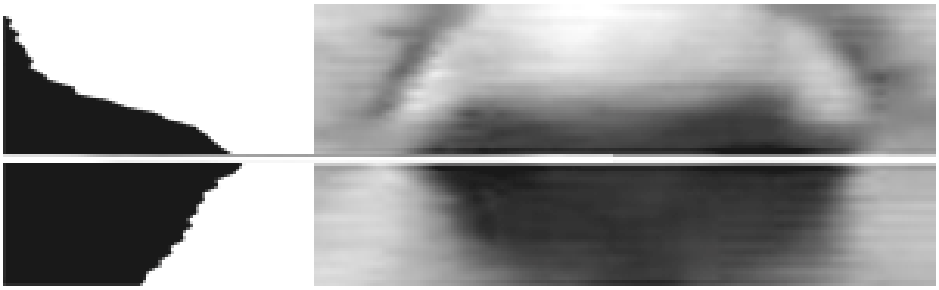

Figure 35. Finding the darkest y-axis line in the nose ROI, which goes through both nostrils.

Figure 36 present three examples from CK+, and three from our database of successfully finding the y-value of two nostrils and the septum.

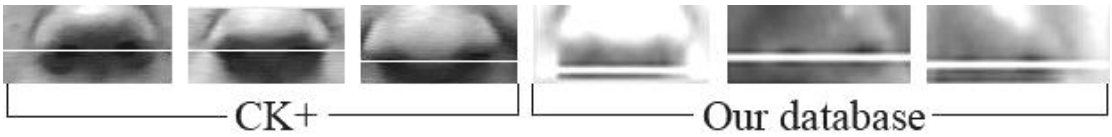

Figure 36. Y-axis line detection results (three from CK+ and three from our database).

With the y-value successfully extracted, the next step was to find the x-value of the septum. Again, two nostrils play an important role. Firstly, few lines above and under the already calculated y-line are added to y-line as a new ROI for nose, in order to reduce the possible error during extraction of y-line. Then, the histogram over x-axis new ROI of nose is calculated, similar to the method in finding the y-value.

Figure 37 presents a method of finding x-values of two nostrils. As seen from this example, observing now x-axis, two nostrils present two darkest regions. This histogram was then smoothened, deleting a few unimportant small local maximums and minimums, and resulting with at least two global maximums. Two central maximums where selected, as they represent the x-values of two nostrils.

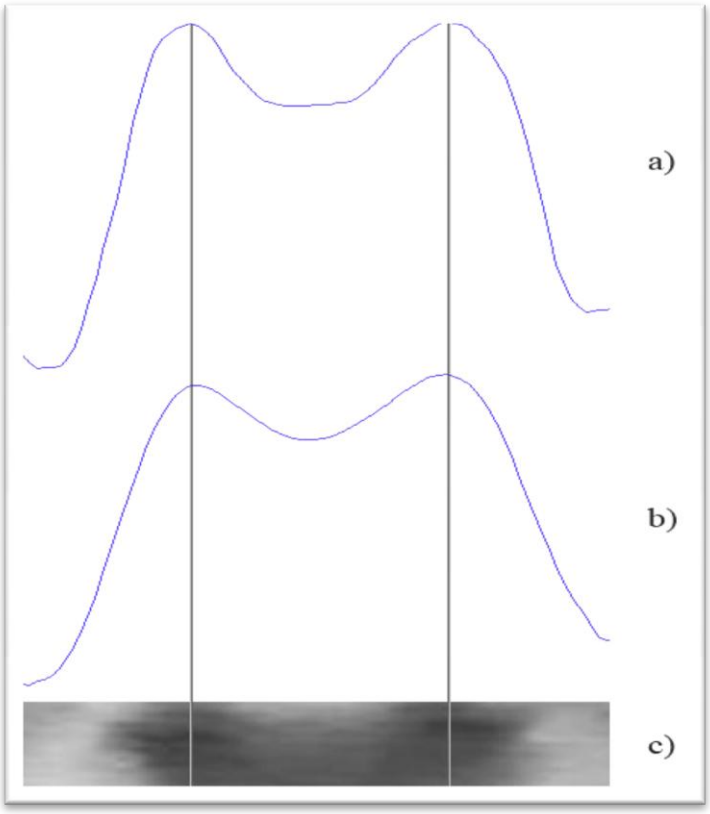

Figure 37. Finding x-axis values for the nostrils: a) the histogram of the ROI of the nose, b) the smoothed histogram, and c) the new ROI of the nose with x-axis values for the nostrils.

The x-value of the septum was calculated as a mid-position of x-values of nostrils. An example of the extracted septum using our proposed method, on an image from CK+ and our database, is displayed in Figure 38. In both left and right images, the septum is represented as a 3 x 3 white dot, which, in the frame from our database (right), looks much larger due to the lower image resolution compared to CK+ (left).

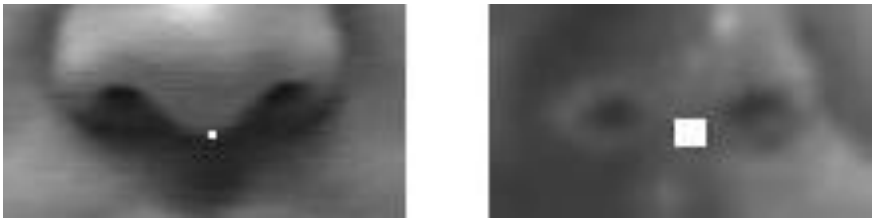

Figure 38. The septum detected in CK+ (left) and in our database (right), using our method

3.3.2.5 Experimental results on CK+

As already presented in subchapter 3.3.2.1, out of 68 extracted facial landmarks (Lucey *et al*., 2010), we have selected only five (2 x middle eyebrow point, 2 x lip-corner point, 1 x the septum, to perform as a reference point in each frame) that seem to be the most important for correctly recognizing facial expression, especially happiness and surprise – two emotions that show low speech emotion recognition rates on our database, and whose recognition rates needed to be improved using facial expression recognition.

Our previous tests were performed utilizing SVMs, so that they can be compared with results from other related papers that use this same classification method. However, in order to be able to combine our facial expression results with those from speech emotion, the following experiments are trained and tested, similarly to experiments on emotion speech segments classification, using NNs.

Table 15. NN facial expression recognition results on CK+ using middle eyebrows points, extracted by AAM.

***Accuracy* [%]**	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**76.9**	0	0	7.7	7.7	7.7
Sad.	0	**83.3**	16.7	0	0	0
Sur.	0	11.1	**77.8**	5.6	5.6	0
Ang.	37.5	12.5	12.5	**37.5**	0	0
Fear	0	100	0	0	**0**	0
Dis.	27.3	0	0	0	9.1	**63.6**

In Table 14, the results on SVMs using only middle eyebrow points are presented, showing good results for happiness and surprise, while Table 15 shows results of the same experiment, but using NN. Both experiments use only two eyebrow points and the septum as a reference point, extracted by Lucey *et al*. (2010), without employing middle frames.

Results in both Table 14 and Table 15 show good recognition for happiness and surprise, which was our main goal.

In the following test, displayed in Table 16, lip-corner points were involved in calculations, which increased the overall results. Again, all points are extracted with AAM (Lucey *et al*., 2010), and the tested dataset was composed of 309 rows (number of image sequences in CK+) and 12 features (displacement from the neutral to the peak frame, and movements in x- and y-axes for 4 selected points – 2 eyebrow points and 2 lip-corners). The septum was used as a reference point in all frames, in order to increase the recognition accuracy, as presented previously in this thesis.

Table 16. NN facial expression recognition results for CK+ using the middle eyebrows and lip-corner points, extracted by AAM.

***Accuracy* [%]**	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**85.7**	0	0	0	14.3	0
Sad.	0	**71.4**	0	0	28.6	0
Sur.	5.6	5.6	**83.3**	0	5.5	0
Ang.	0	0	0	**42.9**	0	57.1
Fear	0	14.3	14.2	14.3	**42.9**	14.3
Dis.	0	0	12.5	12.5	0	**75**

With overall recognition rate of 72.1%, the use of lip-corners increased the results, especially for happiness, surprise, and fear.

Previous results were calculated using extracted facial landmarks (Lucey *et al*., 2010), so our next step was to test our proposed facial point extraction methods on CK+. Table 17 displays results obtained with the same experimental setup as the previous test, but with points extracted with our proposed methods, instead of AAM. As in the previous test, the dataset is composed of 309 samples and 12 features, that is 309 x 12 matrix.

Overall accuracy of 77.6% produced with our proposed approach improved the overall results, comparing to the results obtained with extracted points by Lucey et al. (2010). Also, we can conclude that perfect facial landmarks have been chosen, as results in Table 17 show good performance for emotions that previously did not reach high rates from our speech emotion recognizer.

Table 17. NN facial expression recognition results for CK+ using the middle eyebrows and lip-corner points, extracted with our methods.

***Accuracy* [%]**	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**81.3**	0	0	6.2	12.5	0
Sad.	0	**50**	0	0	50	0
Sur.	0	4.8	**90.5**	4.7	0	0
Ang.	50	0	0	**0**	0	50
Fear	50	0	50	0	**0**	0
Dis.	6.7	0	0	13.3	0	**80**

The following step in our research, presented in the following subchapter, was to test our proposed methods on our Thai emotion audiovisual database.

3.3.2.6 Experimental results on our database

The experiments on our Thai emotion audiovisual database, using a reference point and five selected points, classified with NN show overall recognition of 75%, which is similar to the same experiment conducted on CK+ (see Table 17).

This subchapter presents facial expression recognition results on our database using all proposed methods in this thesis, so use of middle frames was introduced. However, unlike in tests on CK+ (see subchapter 3.3.1.4), video frames from our database could not be involved linearly over time. The CK+ represents a database of image sequences from neutral to peak frame, while ours is an audiovisual database in which subjects express a certain emotion and can reach many peaks over time. Hence, different approach of selecting middle frames had to be implemented.

Firstly, as a part of our database, a few initial/neutral video clips were recorded. The first frames from each of those recordings were selected as neutral frames in our experiments. Secondly, five selected points were extracted employing our proposed techniques in all frames of all recordings from our database. Then, in each group (see subchapter 3.1), a displacement of the extracted points in each frame was calculated and scaled. Unlike in CK+, where the subjects' head movements are gentle, the subjects in our database moved their heads in all directions, as a part of

their emotion statement. With use of a reference point, our calculations have become more resistant to left-right and up-down head movements, but we needed to make our calculations resistant to forward-backward movements as well. Thus, all calculated point displacements were scaled compared to ROI of face in the neutral and observed frame. Finally, a peak frame has been chosen as the one in which the displacement of facial points is the largest.

The displacements of four selected points (2 x middle eyebrow point, 2 x lip-corner point), with respect to an extracted reference point (the septum), were calculated between the neutral and the peak frame, and the mean values of the displacement of lip-corner points in all other frames in that group (that represents a middle frames) are added as facial features. Since the expression of emotion in our database is not linear over time, the middle frames cannot be chosen as frames at one third and two thirds over time. Also, our database involves speech that can influence the recognition process. For example, a happy expression, while pronouncing the vowel “o”, can be misclassified as a possible surprise expression, due to the round shape of the mouth, therefore, calculating the peak frame only with lip-corners is not useful on any bimodal system. However, the mean value of lip-corner displacements can reveal the group’s emotion. Since all subjects read the same text, average displacement of lip-corners vary from emotion to emotion – a mean value in recordings of happiness should show pulled lip-corners, depressed in recordings of sadness, etc. Our results show that this assumption is correct, as the final results on our database yields great recognition rate.

Our final feature set involves 18 features:

- 4 features (2 x middle eyebrow point, 2 x lip-corner point) for the displacements between the natural (N) and the peak (P) frame.
- 8 features (2 x 4 selected points) for movements in the x- and y-axes from N to P.
- 2 features (2 x lip-corner point) for the displacements between the natural (N) and the mean value of all middle frames (M).
- 4 features (two sets of two lip-corner points) for movements in the x- and y-axes from N to M.

The final number of groups, same as in speech emotion recognition, is 305, so our final data was 305 x 18 matrixes.

In this final experiment on facial expression recognition, all of our proposed approaches have been combined together. Firstly, all points were extracted using our proposed methods. Then, employing the reference points, our calculations have become more resistant on head movements. The movements in x- and y-axes also improved the results, and involving the middle frames made the whole process more natural and realistic. Our final set of results on facial expression recognition reaches the overall recognition rate of 95.1%, and is presented in Table 18.

Table 18. Facial expression recognition results on our database, combining all our proposed methods.

***Accuracy* [%]**	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**100**	0	0	0	0	0
Sad.	0	**100**	0	0	0	0
Sur.	0	0	**80**	10	0	10
Ang.	0	0	0	**100**	0	0
Fear	0	0	0	0	**100**	0
Dis.	0	0	9.1	0	0	**90.9**

Results in Table 18 show excellent results for all facial expressions, with the lowest recognition rate reached for surprise. This is probably due to the speech that influences our subject while recording. Even though happiness and surprise are considered as two most easily recognized facial expressions, this may not be true in a real-life situation. Many related papers show excellent recognition of surprise, but most of those works were tested on acted databases, such as CMU and CK+, where subjects can freely express any facial expression, and some of them even exaggerate. Many of those expressions are not likely to be seen in public. Our experiments prove that with presence of speech, where subjects were not able to, for example, open their mouth widely in a surprise, recognition of this facial expression fails. Nevertheless, our final set of facial expression recognition results on our database reveals that all of

our proposed approaches made an excellent improvement in the recognition rate, yielding similar performances for all basic emotions.

After reaching excellent facial expression results, the next chapter presents combining emotion recognition from speech and facial expressions, as well as the comparison to human performance.

CHAPTER 4

FINAL RESULTS

This chapter presents comparison of our previous results to human performance, more detailed explanation of the method for fusing results obtained from speech and facial expression, and final results obtained using audiovisual information from our presented bimodal system.

4.1 Human Performance

There is some work focused on human emotion recognition performance from audio, video, and combined audio-video recordings.

De Silva *et al*. (1997) concluded in their study that happiness, surprise, and anger are video dominant, sadness and fear are audio dominant, while disgust (in their experiments categorized as dislike) shows mixed results.

In his research on Spanish and Sinhala, Chen (2000) presented that Spanish speakers' emotions are more easily recognized from video recordings, while audio emotion recognition yielded better results for Sinhala, with constant misclassification of anger as disgust, and disgust as anger for both languages. Table 19 presents his final human performance results.

Table 19. Human emotion recognition performance for Spanish and Sinhala (Chen, 2000).

	Spanish	*Sinhala*
Video-only	53.8 %	26.8 %
Audio-only	41.7 %	32.3 %
Audio-video	53.4 %	39.9 %

Analyzing Table 19, the interesting thing is that video emotion (facial expression) recognition in Spanish achieved slightly better results than audiovisual performance. However, both in Spanish and in Sinhala, audiovisual performance showed much better results than average performances using only audio and only video recordings.

The following table, Table 20, displays human emotion recognition performance in Thai, using our database. Results are listed for only audio, only video, and audio-video performance of each six emotions.

Table 20. Human emotion recognition performance on Thai, from audio-only, video-only, and audio-video recordings.

Accuracy [%]	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*	*Overall*
Audio	95.7	69.1	0	28.2	0	48.1	*44*
Video	63.9	93.3	0	32	34.5	49.9	*46.8*
Audio-video	70.8	96.4	7.7	63.6	27.8	40	*59.8*

Comparing to the results in Spanish and Sinhala, human performance in Thai yields higher accuracies. This is probably due to the fact that, unlike in the experiments in Spanish and Sinhala, where employed subjects were not familiar with these two languages, our tests gathered two Thai native speakers. Having that in mind, it is understandable that human performance on our database shows better results in audio, video, and joint audio-video recordings. Being natives, our subjects are familiar with ways of expressing emotions in speech, as well as with common facial expressions for Thai people.

Furthermore, it seems that expressing and recognizing emotions in Thai differ from some universal emotion recognition experiments. For example, happiness yields good audio-video results, which is expected, but it performs much better results using only audio cues. Happiness is generally considered as a strong visual emotion, which was not the case in our experiments. Also, surprise reaches particularly low results, with the highest results obtained for sadness. Surprise is, like happiness, a visual emotion, but this low performance shows that this is not the case when speech is

involved. Speech interferes with the subjects' facial expressions, so they were not able to express it as strongly as in any acted facial expression database. Also, surprise alone acts more like a trigger emotion, to switch our attention to something, so it was difficult to express it constantly over recording time, keeping, for example, the mouth wide open in a surprise. Also, fear seems to be difficult for both expressing and recognizing.

These results can tell us much about Thai emotion expression and public behavior, where happy and sad emotions seem to dominate, therefore, they are much easier to recognize. However, our system overall accuracies for audio, video, and audio-video performances, overtop the human performance.

The following two tables, Table 21 and Table 22, present results comparison of human performance and our system performance for audio and video recordings.

Table 21. Comparison of human and our system speech emotion performance.

Accuracy [%]	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*	*Overall*
Audio (human)	95.7	69.1	0	28.2	0	48.1	*44*
Audio (system)	12.9	75	42.3	97.5	100	77.2	*72.4*

Table 22. Comparison of human and our system facial expression performance.

Accuracy [%]	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*	*Overall*
Video (human)	63.9	93.3	0	32	34.5	49.9	*46.8*
Video (system)	100	100	80	100	100	90.9	*95.1*

Comparison of our final audiovisual results with human performance from audio-video recordings is presented in subchapter 4.3.

4.2 Speech-Facial Result Fusion and Classification

In our speech emotion recognition and facial expression recognition experiments, we utilized 2-layer (different number of input nodes) feed-forward back-

propagation NN with 1 hidden layer and different numbers of hidden nodes, implemented using MATLAB's NN toolbox. As shown previously, use of 40 hidden numbers in a hidden layer yields best performance.

The classification of speech features and face features was done separately, with the aim of achieving high accuracy from the speech emotion recognition, and then improving it by analyzing face features. Accordingly, face features were not examined as equally important features, but rather as additional, auxiliary features to be used only if the classification of a segment of an examined audio file is below a certain threshold.

After the NN classifier had been trained, and string of emotion speech segment grouped together, the classification results for that group might look like [0.6 0.1 0.2 0 0.1 0], meaning that this group is closest to be classified as emotion 1.

On the other side, facial features were extracted from relatively low quality video and could have misled the classifier. Therefore, tracking was carried out on the face features taken from low quality video and utilized as an additional set of features. Even though these features were extracted from low quality video clips, they still improve the overall system accuracy.

Examining the output data of speech processing, it is clear that the speech emotion recognition errors come from resulting data with a low top value. For example, a speech group could be classified with a low top value [0.52 0.49 0 0.02 0 -0.05]. That group would be placed in emotion class 1, but with a high error probability, because there is a high chance (49%) that it should be in class 2.

Our experiments show that part of the system error comes from the classification of segments for which the highest emotion value is lower than 0.54. If the classification result for a segment is lower than this set threshold (e.g. [0.52 0.35 0 0 0 0.13] has highest value of 0.52), then face features are introduced to improve the classification decision with both speech and facial features.

The second threshold that was put to test was the usage of speech-face ratio if the face features are involved in classification. On our Thai audiovisual database, ratio of 32%-68% (speech-to-face) yields best results.

4.3 Final Results using Our Proposed Bimodal System

The main goal of this thesis was to research what speech features can improve the speech emotion recognition, build a speech emotion recognition system, and improve it with face feature analysis.

Results on speech emotion recognition are presented in subchapter 3.2.2, and results on facial expression recognition in subchapter 3.3.2.6, with a next step to combine those results together.

Our experiment on bimodal information involved two thresholds.

```
MAX_RATE    =  0
TOP_1       = -1
TOP_2       = -1

Threshold_1 = 0

WHILE Threshold_1 < 1

      Threshold_2 = 0

      WHILE Threshold_2 < 1

            FOR i = 1 TO No_of_elements_in_speech_group

                  IF max(Speech_group(i)) < Threshold_1

                     Group(i) = Threshold_2 * Speech_group(i) +
                              + (1 - Threshold_2) * Face_group(i)
                  ELSE
                     Group(i) = Speech_group(i)

                  END IF
            END FOR

            IF Recognition_rate(Group) > MAX_RATE

                  MAX_RATE = Recognition_rate(Group)
                  TOP_1 = Threshold_1
                  TOP_2 = Threshold_2
            END IF

            Threshold_2 = Threshold_2 + 0.01
      END WHILE

      Threshold_1 = Threshold_1 + 0.01
END WHILE
```

Figure 39. Pseudo-code of finding the best values for two thresholds for fusing emotion speech and facial expression recognition rates.

The first threshold represents the speech emotion top probability under which facial expressions are introduced and mixed with the speech results. For example, each group is classified as a string of values [0.8 0.1 0.1 0 0 0], which present the probability that this examined group should be classified in class 1, class 2, class3, class 4, class 5, and class 6 respectively. If our first threshold is, say 0.9, then if the top probability is under that threshold, facial expression results are involved in the experiment; if not, the final decision is based only on speech. In the previous example, the top probability of 0.8 is lower than the set threshold, therefore, the facial feature for that group would be included in the calculations.

The second threshold is connected with the ratio of speech and face features if the facial features are included.

Accordingly, (see Figure 39) the first threshold in this final experiment shows when to involve facial results, and the second threshold, if the facial features are involved, answers how much. Our final experiments using both speech and facial features show that the best results are obtained when the first threshold is 0.54, with the second threshold of 0.32. This means that facial results will be included only when the top speech emotion probability is under 0.54, and, if it is below that value, with the radio of 32% of speech and 68% of facial information. This proves that speech emotion recognition can produce relatively good results, but if the recognition is uncertain, then facial expression results play more important role with 68% of overall accuracy.

Table 23. Emotion recognition results after fusing speech and facial results.

***Accuracy* [%]**	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*
Hap.	**71**	3.2	9.7	0	12.9	3.2
Sad.	2.1	**89.6**	0	2.1	6.2	0
Sur.	1.9	0	**86.6**	9.6	0	1.9
Ang.	0	0	0	**100**	0	0
Fear	0	0	0	0	**100**	0
Dis.	0	3.5	7	0	0	**89.5**

Our overall accuracy using 0.54 and 0.32 thresholds shows 91.1% recognition rate, with more detailed results presented in Table 23. Interestingly enough, highest recognition accuracies for each of the six basic emotions were reached with these two threshold values, which proves that we found the perfect speech-face combination for recognizing emotions in our study.

The following figure, Figure 40, displays overall recognition rate depending on these two thresholds.

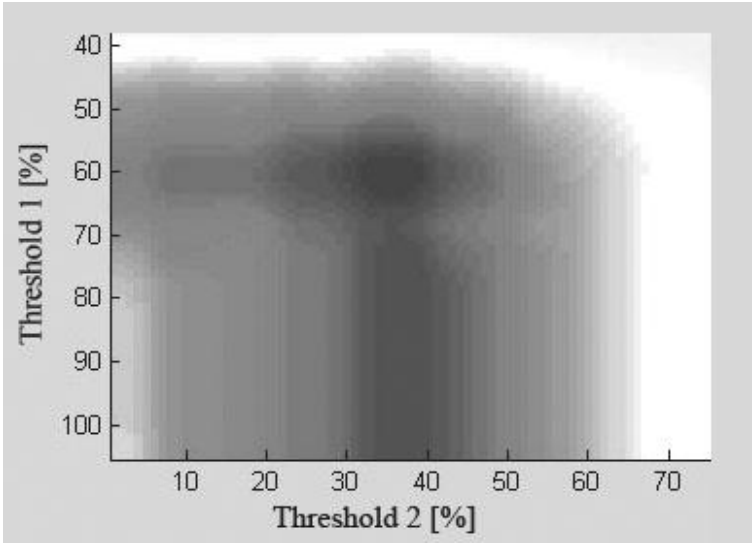

Figure 40. Part of our system's emotion recognition rate's sensitivity, depending on the two thresholds (light gray = a low recognition rate; dark gray = a high recognition rate).

Figure 40 partly displays our system's sensitivity depending on two thresholds. Both thresholds are represented in percentages, where threshold 1 is percentage below which facial expressions are included in the calculations and threshold 2 the percentage of used speech results if both audio and visual information are used. Low recognition rates are plotted with light gray and high rates with dark gray color. As seen in Figure 40, the highest recognition rates are achieved with threshold 1 around 60%, and threshold 2 around 35%.

Table 24 displays our final system performance compared with human performance from audio-video recordings.

Table 24. Comparison of human and our system's final emotion recognition performances (audio + video).

Accuracy [%]	*Hap.*	*Sad.*	*Sur.*	*Ang.*	*Fear*	*Dis.*	*Overall*
A+V (human)	70.8	96.4	7.7	63.6	27.8	40	*59.8*
A+V (system)	71	89.6	86.6	100	100	89.5	*91.1*

Our future work will be to focus on improving audio-video performance for sadness, which shows better results in human recognition, and improve the speech emotion recognition for happiness. Nevertheless, our final system performance far surpasses the human performance with 91.1% to 59.8% recognition rates.

The fact that both our system and the human emotion recognition performance for audio-video recordings show better results than using only audio or only video, proves that both hearing and vision play an important part in recognizing emotions in Thai language, which was the main goal of this thesis.

CHAPTER 5

DISCUSSION AND CONCLUSION

This chapter includes the discussion and conclusion of our study, and presents research contribution of this thesis. At the end of the chapter, and the thesis, a list of suggested future improvements is provided.

5.1 Discussion

Emotion recognition systems that utilize only speech or facial expressions do not represent a realistic way of communication and expressing emotions. In everyday life, humans use both vision and hearing to recognize emotions, thus this bimodal approach presents a more realistic and intuitive way of detecting emotional states.

In our research, the source of all difficulties came from our database. First of all, it seems that the recording of 972 words in six emotions by six students was too large for the first research of this kind for Thai. For example, in the speech emotion processing phase, after cutting all recording files into 30ms sound-segments (with 20ms overlapping), we ended up with a matrix of around 450,000 rows, that was too large to train in a single NN. Instead, our dataset was cut into blocks of maximum size that can be trained in a single NN (around 20,000 rows per block), and the final results calculated as an average value of all outputs from those blocks. This technique, however, does not produce results as accurate as if trained in a single NN.

While recognizing facial expression, our biggest problem was our database video resolution of 320 x 240 pixels. Even though image resolution and size is not an issue in image processing these days, we certainly came across several problems because of it. For example, ROI for eyebrows in our database presents a small image of say 20 x 20 pixels, which is then hard to find any edges in. Comparing our database with the standard database of facial expressions, such as CK+ that was also used in this thesis experiments, the resolution of some ROIs in the frames from our database

was too low, with less controlled lightning, and smaller contrast. Furthermore, in order to make them feel more relaxed, our subjects were told to move freely during recording, which resulted in probably more realistically expressed emotion, but with a more difficult job in the image processing stage, especially in capturing facial landmarks. However, we have proposed methods that show great results on both standard database CK+ and on our database.

Finally, the largest problem with our database is the fact that all subjects during the recording get, at some point, tired of expressing a certain emotion, especially if they are to read almost 1,000 words. This problem is a universal issue in all acted databases. There are still many critics of acted database, saying that they do not represent a natural expression of emotions. However, recent studies show that the problem is not in the use of actors, but in an approach of recording acted databases. Firstly, as stressed in several studies on human behavior, subjects' actions become completely unnatural the moment they realize or suspect that they are recorded. This issue was tried to be resolved using hidden cameras and asking the subjects for their agreement on using those recordings, once the recordings are done, but even this approach, by some experts, is not completely correct. What one needs to do in order to make a perfect emotion audiovisual database is to record people's reactions in a real-life situation, without them knowing that they have been filmed, which is all together, of course, completely impossible. There lies the biggest problem in the field of emotion recognition – the inability to capture completely natural emotions. Hence, lacking some standard audiovisual emotion database with fully natural expressions, researches turn to acted emotions, as they can produce the largest testing dataset, and try to make all recorded subjects, as we did, as relaxed as possible, in order to capture emotions as naturally as possible.

Of course, there are several issues on emotions as well. For example, happiness is very easy to express, because it is the most frequently expressed emotion, but there is a big difference between acted and natural happy expression. On the other hand, fear is particularly difficult to express on demand. People are usually not sure, if asked on demand, how this facial expression even looks like. Fear is part of the

human defensive mechanism, thus being a highly instinctive emotion, difficult and rarely expressed.

Nevertheless, this thesis proposes several methods and approaches that reached excellent results in different databases, and surpassed our expectations with 91.1% recognition rate.

Our idea is that both audio and vision play a crucial and inseparable role in recognizing emotions. Accordingly, in order to successfully recognize emotions our system is based on building a speech emotion recognition system and improving it with face feature analysis.

There are some emotions that are strong “visual” emotions and some that rely more on hearing. Opposite to results of several related work on facial expression recognition, our results on emotion recognition from speech show great performance for all emotions, except happiness and surprise – emotions that are considered as most easily recognized facial expressions. Our feature selection methods chose ZCR and energy from short-time sound signals to be most informative in distinguishing our database speech in emotional groups, and were tested with the well-known MFCC technique. The reason why ZCR and energy show good results on our database probably lies in Thai language and culture. It seems that Thai people, unable to change tones of words in order not to change the meaning, tend to increase speech rate and/or use harder/softer voice to express their feelings.

On the other side, we proposed several approaches on improving facial expression recognition, with automatic facial landmark extraction. Experimental results on CK+ proves the quality of our methods, while using our database reached results that were then added to speech emotion process. With use of facial features, we have managed to increase the speech emotion results of 72.4%, utilizing two variable thresholds in pursuit of the best combination of speech and facial information, and yielded the final rate of 91.1%, over-performing human recognition results and proving that both speech and facial gestures are important in recognizing emotions.

5.2 Research Contribution

This project is the first of its kind for Thai, therefore, we had to record an emotion speech audiovisual database prior to our experiments. This database can be used in any related project for speech emotion recognition in Thai, in any facial expression recognition project, or used as a starting database in recording a newer version of audiovisual database for further research. The whole database will be left to professors and students of the Department of Computer Engineering at PSU, for further research, use, and development.

With usage of feature selection methods, we have found out and concluded which features are useful in detecting emotions from Thai speech. Two selected features, short-time ZCR and short-time energy, have reduced the error and, with the use of the standard MFCC method, produced good results on our database. However, results for happiness and surprise from speech did not yield good results, and our following step was to increase recognition using face feature analysis.

Our facial expression recognition system introduces several new approaches. First, only five most important facial landmarks were selected, and fully automatic extraction of those points proposed. Secondly, the significance of employing the reference point in all frames, to make calculations more resistant to head movements, with additional observation of displacements in x- and y-axes, was proven. Also, in our more neurophysiologic approach, the usage of middle frames in all facial expression recognition systems was presented. Facial emotions are expressed differently over time, so middle frames (frames that lead to the peak expression frame) can also reveal a lot about the emotional state of a subject. Facial expression tests were performed on standard database (CK+) and our database, and reached excellent performance.

With a proposed way of combining speech and face features, facial expression analysis increased our emotion speech recognition rate from 72.4% to 91.1%. These final results from bimodal information completely surpass human performance, and prove that our approach of utilizing emotion recognition from both speech and face was justified.

All proposed methods from this thesis show great results, are easily implemented, intuitive, and can be employed in any related project.

5.3 Future Work

Our several preliminary tests of speech emotion recognition on a portion of our database showed excellent results using ZCR and energy. However, this recognition rate dropped when used on the whole database, partially because we were not able to train it in a single NN. Our sound-segments of 30 ms were first completely randomized, and then the whole randomized dataset cut into blocks of maximum size that was possible to be placed and trained in a single NN. However, this approach reduces that overall accuracy, so part of our future work is to implement NN that is able to train our whole database in a single network. A possible solution is to change the MATLAB's NN toolbox so that it can read our data from a text file, instead of working in a current batch mode.

With the previous obstacle removed, speech emotion recognition rate would certainly be increased. Then, we will focus on how to more correctly capture happiness from Thai speech. Recognition of happiness show poor results from speech, which was expected, since happiness is a highly "visual" emotion, but it shows lower results than other emotions in our final bimodal experiments too. Accordingly, we will focus on finding a speech feature that distinguishes happiness alone from other emotions, and we shall hopefully increase the speech emotion recognition results.

Even though our final facial expression recognition, after involving all our proposed techniques, yields excellent accuracy of 95.1%, there is still room for improvement. In particular, disgust and surprise show mutual misclassification. This error could be reduced using, for example, nose wrinkle detection that will be a part of our future work. The expression of surprise produces almost no wrinkles, because of the wide open mouth without raising the upper lip, and with raised eyebrows. On the contrary, nose wrinkles are most present during the expression of disgust. Thus, our approach will be to decrease the misclassification of disgust and surprise by utilizing a simple nose wrinkle detector.

Finally, this thesis presents results on standard database (CK+) and on our database. Our next step would be to test our proposed methods on other publicly available databases.

REFERENCES

Aldrian, P., Meier, U., and Pura, A. 2009. “Extract feature points from faces to track eye’s movement,” University of Leoben, Austria

Ball, L. 2011. “Enhancing border security with automatic emotion recognition,” *International Crime and Intelligence Analysis Conference 2011 (ICIAC11),* November 2011, Manchester, UK.

Barrow, R. 2010. “Thai & English talking software dictionary,” *Paknam Books,* Samut Prakan, Thailand.

Batty, M. and Taylor, M. J. 2003. “Early processing of the six basic facial emotional expressions,” *Cognitive Brain Research,* Vol. 17, pp. 613–620.

Bettadapura, V. 2009. “Face expression recognition and analysis: the state of the art,” Columbia University, Computer and Information Science, USA

Bradski, G. and Kaehler, A. 2008. “Learning OpenCV,” *O’Reilly,* USA.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. 2005. “A database of German emotional speech,” *Interspeech’2005 - Eurospeech — 9th European Conference on Speech Communication and Technology,* Lisboa, Portugal, September 4-8, 2005, pp. 3-6.

Canny, J. 1986. “A computational approach to edge detection,” *IEEE Transaction on Pattern Analysis and Machine Intelligence,* Vol. 8, No. 6, pp. 679–698.

Chang, C.-C. and Lin, C.-J. 2011. “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology,* Vol. 2, No. 3, Article No. 27.

Chen, C. H. 1988. “Signal processing handbook (electrical and computer engineering),” *CRC Press*, New York, USA.

Chen, L. S.-H. 2000. “Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction,” *PhD Thesis*, University of Illinois at Urbana-Champaign, USA.

Cohen, I., Sebe, N., Garg, A., Chen, L. S., and Huang, T. S. 2003. “Facial expression recognition from video sequences: temporal and static modeling,” *Computer Vision and Image Understanding*, Vol. 91, pp. 160-187.

Comrie, B. 1990. “The world's major languages,” *Oxford University Press*, New York, USA.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. 1998. “Active appearance models,” *Proceedings of the European Conference on Computer Vision*, Vol. 2, pp. 484-498.

Cortes, C. and Vapnik, V. N. 1995. “Support-vector networks,” *Machine Learning*, Vol. 20, pp. 273-297.

Dalley, G. 2006. “VideoIO: granting easy, flexible, and efficient read/write access to video files in MATLAB on Windows and GNU/Linux platforms,” *Free Software Foundation*, Inc. 675 Mass Ave, Cambridge, MA 02139, USA.

Daugman, J. G. 1985. “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters,” *Journal of the Optical Society of America*, July 1985, Vol. 2, No. 7, pp. 1160–1169.

De Silva, L. C., Miyasato, T., and Natatsu, R. 1997. "Facial emotion recognition using multimodal information," *Proceedings of IEEE International Conference on Information, Communications and Signal Processing (ICICS'97)*, Singapore, September 1997, pp. 397–401.

Ding, C. and Peng, H. 2005. "Minimum redundancy feature selection from micro-array gene expression data," *Journal of Bioinformatics and Computational Biology*, Vol. 3, No. 2, pp.185-205.

Ding, C. and Peng, H. 2003. "Minimum redundancy feature selection from micro-array gene expression data," *Proceedings of the Second IEEE Computational Systems Bioinformatics Conference (CSB 2003)*, Stanford, CA, USA, August 2003, pp. 523-528.

Dornaika, F. and Davoine, F. 2008. "Simultaneous facial action tracking and expression recognition in the presence of head motion," *International Journal of Computer Vision*, Vol. 76, No. 3, pp. 257–281.

Ekman, P. and Friesen, W. V. 1971. "Constants across cultures in the face and emotions," *Journal of Personality and Social Psychology*, Vol. 17, No. 2, pp. 124-129.

Ekman, P. and Friesen, W. V. 1977. "Manual for the facial action coding system," *Consulting Psychologists Press*, Palo Alto, USA

Ekman, P., Friesen, W. V., and Hager, J. C. 2002. "Facial action coding system: a human face," Salt Lake City, Utah, USA

Fant, G. 1960. "Acoustic theory of speech production," *Mouton & Co*, The Hague, Netherlands.

Ghosh, S. 2001. “Formant tracker for MATLAB,” SpeechLab, Boston University, USA, http://cns.bu.edu/~speech/ftrack.php

Gustavsen, E. 2007. “Classifying motion picture audio,” *Master’s Thesis*, Department of Computer Science and Media Technology, Gjøvik University College, Norway.

Haas, M. R. 1964. “Thai-English student's dictionary,” *Stanford University Press*, 1st edition, Palo Alto, California, USA, June 1, 1964.

Hone, K. and Bhadal, A. 2004. “Affective agents to reduce user frustration: the role of agent gender,” *Human-computer interaction (HCI) 2004*, Vol. 2, pp. 173-174.

Hopfield, J. J. 1982. “Neural networks and physical systems with emergent collective computational abilities,” *Proceedings of the National Academy of Sciences*, Vol. 79, USA, April 1982, pp. 2554-2558.

Hu, H. and Zahorian, S. 2008. “YAAPT (yet another algorithm for pitch tracking) – fundamental frequency (pitch) tracking,” *The Journal of the Acoustical Society of America*, Vol. 123, No. 6, June 2008, pp. 4559-71.

Jovičić, S. T., Kašić, Z., Đorđević, M., and Rajković, M. 2004. “Serbian emotional speech database: design, processing, and evaluation,” *SPECOM’2004: 9th Conference Speech and Computer*, St. Petersburg, Russia, September 22-24, 2004, pp. 77-81.

Kanade, T., Cohn, J. F., and Tian, Y. 2000. “Comprehensive database for facial expression analysis,” *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*, Grenoble, France, pp. 46-53.

Klein, J., Moon, Y., and Picard, R.W. 2002 “This computer responds to user frustration: Theory, design and results,” *Interacting with Computers*, Vol. 14, pp. 119-140.

Krishna, S. 2008. “Open CV Viola-Jones Face Detection in MATLAB,” http://www.mathworks.com/MATLABcentral/fileexchange/19912

Lewis, M. P. 2009. “Ethnologue: languages of the world, sixteenth edition**,** ” *SIL International*, Dallas, Texas, USA.

Lienhart, R. and Maydt, J. 2002. “An extended set of Haar-like features for rapid object detection,” *IEEE ICIP 2002*, Vol. 1, September 2002, pp. 900-903.

Lisetti, C.L. and Schiano, D.J. 2000. “Automatic facial expression interpretation: where hci, artificial intelligence and cognitive science intersect,” *Pragmatics and Cognition*, Vol. 8, No. 1, pp. 185-235.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. 2010. “The extended Cohn-Kanade dataset (CK+): a complete expression dataset for action unit and emotion-specified expression,” *Proceedings of the Third International Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB 2010)*, San Francisco, USA, pp. 94-101.

Mermelstein, P. 1976. “Distance measures for speech recognition, psychological and instrumental,” *In Pattern Recognition and Artificial Intelligence*, New York, pp. 374–388.

Mermelstein, P. and Davis, S. B. 1980. “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-28, No. 4, August 1980, pp. 357.

Michel, P. 2003. “Support vector machines in automated emotion classification,” *CST Part II Project Dissertation*, Churchill College, UK.

Murray, I. and Arnott, J. 1993. “Toward the simulation of emotion in synthetic speech: a review of the literature of human vocal emotion,” *Journal of the Acoustic Society of America*, No. 93, Vol. 2, pp. 1097–1108.

Nakatsu, R., Nicholson, J., and Tosa, N. 1999. “Emotion recognition and its application to computer agents with spontaneous interactive capabilities,” *Proceedings of the Seventh ACM International Conference on Multimedia*, October 30-November 5, 1999, Orlando, Florida, USA, pp. 343-351.

Nicholson, J., Takahashi, K., and Nakatsu, R. 1999. “Emotion recognition in speech using neural networks,” Proceedings of the Sixth International Conference on Natural Information Processing (ICONIP’99), Perth, Australia, November 16-20, 1999, Vol. 2, pp. 495-501.

Omogbenigun, O. 2007. “SpeechCore,” London Metropolitan University, UK.

Pantic, M., Valstar, M. F., Rademaker, R., and Maat, L. 2005. “Web-based database for facial expression analysis,” *Proceedings of the Thirteenth Annual ACM International Conference on Multimedia (Multimedia’05)*, Singapore, November 6-11, 2005, pp. 317-321.

Pantic, M. and Rothkrantz, L. J. M. 2003. "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, Vol. 91, No. 9, September 2003.

Papageorgiou, C. P., Oren, M., and Poggio, T. 1998. "A general framework for object detection," *International Conference on Computer Vision*, January 4-7, 1998, pp. 555-562.

Peng, H., Ding, C., Long, F. 2005. "Minimum redundancy maximum relevance feature selection," *IEEE Intelligent Systems*, November-December, 2005, Vol. 20, No. 6, pp.70-71.

Peng, H., Long, F., and Ding, C. 2005. "Feature selection based on mutual information: criteria of maximal dependency, maximal relevance, and mininimal redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 8, pp. 1226-1238.

Petrushin, V. 2000. "Emotion recognition agents in real world," *Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium on Socially Intelligent Agents: Human in the Loop*, November 3-5, 2000, North Falmouth, Massachusetts, USA.

Picard, R. 1997. "Affective Computing." *The MIT Press*, Cambridge, Massachusetts, USA.

Sebe, N., Lew, M. S., Sun, Y., Cohen, I., Gevers, T., and Huang, T. S. 2007. "Authentic facial expression analysis," *Image and Vision Computing*, Vol. 25, pp. 1856-1863.

Sharma, N. S. 2009. “Short-time energy (MATLAB toolbox). A generalized linear filter approach for sonar receivers,” *Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop*, DSP/SPE 2009, IEEE 13th, Marco Island, Florida, USA, 4-7 Jan, 2009, pp. 507-512.

Sornlertlamvanich, V., Charoenporn, T., and Isahara, H. 1997. “ORCHID: Thai part-of-speech tagged corpus,” *Technical Report: Orchid Corpus, National Electronics and Computer Technology Center*, Thailand and Japan.

Tan, L. and Jiang, J. 2009. “Novel adaptive IIR notch filter for frequency estimation and tracking,” *IEEE Signal Processing Magazine*, USA, November 2009, pp.168-189.

Tian, Y., Kanade, T., and Cohn, J. 2001. “Recognizing action units for facial expression analysis,” *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 2, pp. 97–115.

Viola, P. and Jones, M. 2001. “Rapid object detection using a boosted cascade of simple features,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Vol. 1, Kauai, Hawaii, USA, December 08-14, 2001, pp. 511-518.

Visutsak, P. 2005. “Emotion recognition through lower facial expressions using support vector machines,” *Proceedings of the Fifth National Symposium on Graduate Research*, Kasetsart University, Bangkok, Thailand, October 10-11, 2005.

Vogt, T. and André, E. 2009. “Exploring the benefits of discretization of acoustic features for speech emotion recognition,” *Proceedings of the Tenth Conference of the International Speech Communication Association (INTERSPEECH)*, Brighton, U.K., September 1, 2009, pp. 328-331.

Williams, C. and Stevens, K. 1972. “Emotions and speech: some acoustical correlates,” *Journal of the Acoustic Society of America*, No. 52, Vol. 4, pp. 1238–1250.

Zhang, T. and Jay Kuo, C.-C. 2001. “Audio content analysis for online audiovisual data segmentation and classification,” *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 4, USA, May 2001, pp. 441-457.

Zhang, Y. and Zhou, J. 2003. “A study on content-based music classification,” *Proceedings of the Seventh International Symposium on Signal Processing and its Applications*, Vol. 2, USA, July 1-4, 2003, pp. 113-116.

Zhou, J. and Peng, H. 2007. “Automatic recognition and annotation of gene expression patterns of fly embryos,” *Bioinformatics*, Vol. 23, No. 5, pp. 589-596.

APPENDIX. THE LIST OF 972 MOST COMMON WORDS IN THAI *

* *only first two words are numbered, with 50 words (only in Thai) per page*

1. การ	ความ	นั้น	ผู้	ระบบ
2. และ	ไม่	ต้อง	ด้วย	กัน
ใน	ว่า	ซึ่ง	คือ	ขึ้น
ที่	พัฒนา	โดย	มา	ตาม
มี	ใช้	อยู่	แล้ว	ต่างๆ
ของ	ก็	ไป	ทำ	ข้อมูล
เป็น	เรา	จาก	แต่	กำหนด
จะ	นี้	คน	เรื่อง	งาน
ได้	หรือ	เพื่อ	มาก	ปัญหา
ให้	กับ	สามารถ	ถ้า	ต่อ

ประเทศ	สำคัญ	ผม	ทำงาน	อุตสาหกรรม
ทาง	ประเภท	เพิ่ม	ผลิต	นำ
ปี	เกิด	ทั้ง	รู้	บริหาร
ยัง	ข้าราชการ	ฟังก์ชัน	แสดง	นัก
ค่า	หนึ่ง	ถูก	ต้องการ	สนับสนุน
ด้าน	เขา	ไว้	ที่จะ	ทำให้
ดี	ศึกษา	เพราะ	ท่าน	สร้าง
สำหรับ	กว่า	แบบ	รัฐบาล	เศรษฐกิจ
จึง	ระดับ	ส่วน	กลุ่ม	ผล
เมื่อ	ถึง	เช่น	เห็น	ควร

อีก	เรียก	แต่ละ	เงิน	เพื่อให้
ดำเนินการ	ปรับปรุง	ส่งเสริม	ตัวอักษร	ประชาชน
โปรแกรม	พิจารณา	บาง	บริการ	สังคม
อย่าง	ประโยชน์	ระหว่าง	สิ่ง	คุณภาพ
สูง	นโยบาย	จำเป็น	วิจัย	หลัก
เรียน	เกี่ยวกับ	อาจ	ช่วย	สินค้า
ใหม่	เทคโนโลยี	เข้า	ปฏิบัติ	เปลี่ยนแปลง
ราชการ	จัด	ตัว	วิธี	ฝึกอบรม
เขียน	ได้รับ	จำนวน	โครงสร้าง	รายได้
คิด	แก่	ลักษณะ	เก็บ	อบรม

ป๋วย	ประกอบ	เหมาะสม	ต่อไป	บทบาท
รวมทั้ง	ขนาด	หน่วยงาน	ออก	อะไร
อ.	ลง	ควรจะ	อย่างไร	จัดการ
น้อย	บุคคล	ใด	รับ	ตำแหน่ง
ประกาศ	ไทย	เวลา	อัน	โครงการ
หลาย	สัมพันธ์	กำลัง	หัวหน้า	กระทำ
หน้าที่	อื่นๆ	ประการ	ข้อ	ประสิทธิภาพ
ทุก	มาตรฐาน	พยายาม	ควบคุม	ลด
ชื่อ	ลงทุน	มัน	เหล่านี้	คำ
ตนเอง	พื้นฐาน	บริษัท	ตัวชี้	เอกชน

กล่าว	วิธีการ	อ่าน	ปัจจุบัน	แผน
โอกาส	รูปแบบ	ตลาด	บ้าง	กิจกรรม
อาจจะ	ฝึก	ส่งกลับ	เอง	ลำดับ
ตน	หา	พื้นที่	เป้าหมาย	เสีย
ภายใน	เปลี่ยน	ก่อน	ขยาย	อัตรา
ราคา	การที่	ป้องกัน	ที่สุด	อื่น
ดังกล่าว	เข้าใจ	เน้น	ขอ	แนวทาง
ส่ง	เท่านั้น	แก้ไข	ประเทศไทย	อำนาจ
ตัวแปร	กระจาย	แรก	ชนบท	เนื่องจาก
มากขึ้น	ดังนั้น	แห่ง	รัฐ	ครั้ง

เกี่ยวข้อง	เริ่ม	คอมพิวเตอร์	เกษตร	รับผิดชอบ
ถือ	งบประมาณ	ปกครอง	เคย	เป็นต้น
เอา	ไปยัง	องค์การ	โลก	เฉพาะ
เศรษฐศาสตร์	หนังสือ	สอดคล้อง	กำลังคน	ธุรกิจ
ดังนี้	จริง	ฯ	ผ่าน	ค้า
วิชา	เพียง	พูด	ชุด	ฉบับ
เข้ามา	ชีวิต	ใหญ่	รวม	เดียวกัน
โดยเฉพาะ	ดำเนินงาน	ต่างประเทศ	ใช่	ตั้งแต่
นิพจน์	ตัวอย่าง	ดู	กฎหมาย	สถาบัน
อาร์กิวเมนต์	เมือง	สู่	มุ่ง	องค์กร

เพิ่มขึ้น	สอง	อาจารย์	สภาพ	ระยะ
เลย	ตั้ง	เดียว	ปฏิบัติงาน	ทราบ
ทั้งหมด	ทรัพยากร	มโนทัศน์	กระทรวง	โรงเรียน
มนุษย์	ส่วนใหญ่	ย่อม	มอง	ภาค
ยาก	เป็นไป	หลักสูตร	แข่งขัน	แตกต่าง
รักษา	ร่วม	ขาด	ข้อคำสั่ง	ระบุ
อธิบาย	จน	ง่าย	ข้างต้น	เสนอ
ที่ดิน	วางแผน	แทน	ร่วมมือ	พอ
มหาวิทยาลัย	ก็ได้	กรณี	วัตถุ	ฝ่าย
ชาว	สอน	สมาชิก	มาตรการ	เลือก

ปรับ	ชัดเจน	แปลง	พบ	สำเร็จ
ก้าวหน้า	ถูกต้อง	ขยายตัว	พิมพ์	แฟ้มข้อมูล
ต่ำ	ทางไกล	ยิ่ง	เริ่มต้น	ดำเนิน
มักจะ	ท้องถิ่น	วิเคราะห์	ประมาณ	สิ่งแวดล้อม
มิ	เกิดขึ้น	ได้แก่	ก็คือ	เจริญ
ผิด	สนใจ	บน	ภาษา	เหตุการณ์
กรม	สาขา	ชี้	ละ	คงจะ
ซื้อ	หรือไม่	ช่วง	เขต	จัดทำ
ตลอดจน	เจ้าหน้าที่	ประสาน	กลับ	จำกัด
บรรทัด	ใคร	ช่วยเหลือ	ชนิด	ยิ่งขึ้น

อาชีพ	จำนวนเต็ม	ส่วนราชการ	เปิด	แม้ว่า
ตัวดำเนินการ	แยก	หมายถึง	ในด้าน	กลาง
เสมอ	ขาย	ขั้น	ไปสู่	ทั่วไป
เกษตรกร	ร้อยละ	ขั้นตอน	รัฐวิสาหกิจ	พวกเรา
คำสั่ง	ข้าราชการพลเรือน	จัดตั้ง	จำเป็นต้อง	แปล
แถวลำดับ	หลัง	วิทยาศาสตร์	เช่นนี้	ก็จะ
ทางเศรษฐกิจ	การเมือง	แฟ้ม	เดิม	บท
ปริมาณ	คง	อนาคต	นอกจาก	ภาษี
ภาคเอกชน	รู้สึก	สมัย	อาศัย	เหตุ
แก้	ข้าว	เปรียบเทียบ	เทศบาล	รุ่น

หน่วย	น้ำ	ขอบเขต	โดยตรง	หัวข้อ
จิตใจ	มัก	ฉะนั้น	บทความ	ผลลัพธ์
ชาติ	นี่	ภาษา	ผลผลิต	ให้กับ
หน่วยความจำ	พลังงาน	ครอบครัว	ยก	ทั้งนี้
เพราะฉะนั้น	ภายนอก	ข้าพเจ้า	เครื่องหมาย	ส่งออก
ไม่ว่า	ระเบียบ	วัน	ทั้งหลาย	ชอบ
จุด	อย่างไรก็ตาม	เครื่อง	ธรรมชาติ	บ้าน
บอก	อุปกรณ์	เด็ก	หลักการ	ญี่ปุ่น
ออกไป	เกิน	ต่อไปนี้	ใหม่ๆ	ดูแล
ค่าคงที่	แม้	สาย	ประชากร	รายละเอียด

นอก	เร่งรัด	จังหวัด	ที่สอง	ขนส่ง
เหมือน	การเงิน	พวก	สายอักษร	บล็อก
คุณ	ทุกคน	แถว	ให้แก่	สิทธิ
ทางด้าน	บุคลากร	แม่	ตัวถูกดำเนินการ	อาหาร
น่าจะ	ยาว	คำนวณ	ต่อเนื่อง	เชื่อมโยง
ภูมิภาค	รายการ	นอกจากนี้	ทฤษฎี	จ่าย
วาง	ร่วมกัน	รูป	ประโยค	พิเศษ
สื่อ	เหตุผล	การ “พ”	เทคนิค	ส่วนกลาง
เครื่องมือ	แรงงาน	จริงๆ	ไม่ได้	ขึ้นอยู่
เพียงพอ	ก็ตาม	ชุมชน	รู้จัก	คุณธรรม

หมาย	พารามิเตอร์	พ.ศ.	แน่นอน	หมด
การเกษตร	สะดวก	หมายความ	โดยทั่วไป	จบ
คลัง	เข้าไป	อยาก	ภาครัฐ	สรุป
ทิศทาง	แนว	แนวโน้ม	มากมาย	ส่วนรวม
บังคับ	ตัวเลข	ในช่วง	วิชาการ	อักษร
บังคับบัญชา	พร้อม	กล่าวคือ	โรงงาน	อุปสรรค
ปัจจัย	ยอมรับ	ขาดแคลน	ขึ้นมา	แบ่ง
มิได้	หาก	พนักงาน	ตัวระบุ	ค่าใช้จ่าย
ครู	แนวคิด	ยังคง	ปรากฏ	รองรับ
ประสบ	ดัง	หวัง	สมบูรณ์	ศูนย์กลาง

เล็ก	ไม่ใช่	ทดสอบ	ติดตาม	ประเด็น
กิจการ	ก.พ.	ข้อความ	สัญลักษณ์	ส่วนภูมิภาค
จัดสรร	ศูนย์	ตอบ	กลไก	กระบวนการ
ฐานะ	อย่างมาก	ทัศนคติ	ข้อผิดพลาด	ตรวจสอบ
ตัวเอง	เกษตรกรรม	บิต	ตัดสินใจ	ต้นทุน
ทำหน้าที่	เงื่อนไข	ระหว่างประเทศ	ทักษะ	ประสิทธิผล
ทุน	เล่ม	เชื่อ	อย่างรวดเร็ว	มาแล้ว
นับ	ก่อ	เผยแพร่	ในทาง	ยากจน
นั้นๆ	คณะ	เหลือ	งานบุคคล	สูงขึ้น
ใดๆ	ค่อนข้าง	จีน	นาน	เท่ากับ

ประจำ	สารสนเทศ	ตอน	ทบวง	เธอ
เก่ง	ประชาธิปไตย	ทั่วถึง	มั่นคง	ต่าง
เท่าที่	ประมวลผล	อันตราย	ลูก	ทัน
คุณลักษณ์	ฟ๊ง	เหมือนกัน	วัตถุประสงค์	ยุติธรรม
ถ้าหากว่า	วินโดว์	แหล่ง	ส่วนมาก	วงวน
ประเมิน	สัดส่วน	จนถึง	เรียนรู้	สังเกต
พิธีกรรม	องค์ประกอบ	บาท	โดยเฉพาะอย่างยิ่ง	เป็นไปได้
รัก	ที่ 1	อย่างต่อเนื่อง	ได้ผล	กรุงเทพมหานคร
เมนู	ประสบการณ์	เร่ง	ชั้น	ฐาน
เร็ว	ฉัน	เหล่านั้น	สุดท้าย	ณ

ทั้งสอง	ทรัพยากรธรรมชาติ	ด้วยกัน	พวกเขา	สาม
ผลงาน	ที่ 2	ตัด	ราษฎร	เสริม
พอสมควร	ประหยัด	ติดต่อ	อนุรักษ์	การคลัง
มากกว่า	ปล่อย	บริเวณ	อย่า	ตาย
รายจ่าย	สถานการณ์	รถยนต์	ออกมา	สำรวจ
ลดลง	อ้างอิง	อย่างมีประสิทธิภาพ	เลื่อน	อังกฤษ
ไฟฟ้า	รวดเร็ว	เครือข่าย	แผนพัฒนา	เลิก
ก็ยัง	รวบรวม	เช่นเดียวกับ	กลายเป็น	เอกสาร
ค่อย	เสริมสร้าง	กฎ	ข้อจำกัด	ก็ดี
ถนน	แท้จริง	ทำให้เกิด	ที่อยู่อาศัย	จัดหา

ผู้ผลิต	ผู้ใหญ่	พ่อค้า	เหมาะ	ลบ
สาธารณสุข	วิทยากร	ละคร	แลกเปลี่ยน	แพง
เกลี่ย	สิ้นสุด	จนกระทั่ง	ข้อมูลเข้า	แม้แต่
เครื่องจักร	เข้าถึง	ตำรา	ควบคู่	ใจ
เงินเดือน	ตลอด	ปรากฏ	ความหมาย	กระตุ้น
เป็นธรรม	นา	ยอม	ตาราง	ครอบคลุม
กลับมา	ปกติ	วิชาชีพ	ถาม	คุณค่า
คำนึง	ประเมินผล	ฯลฯ	พึ่ง	ที่ 7
จอมพล	ผลกระทบ	เท่า	มลพิษ	นาย
ทันสมัย	ผู้อื่น	เป็นอยู่	ยูเนียน	นำเข้า

ประสานงาน	ตามที่	ข้อเสนอแนะ	เรียกว่า	ระดม
ปลอดภัย	ทรัพย์สิน	ซ้ำ	เสถียรภาพ	หลีกเลี่ยง
ภัย	บริโภค	บันทึก	กว้างขวาง	เฉลี่ย
มีผล	สภาพแวดล้อม	บ้านเมือง	กิน	เพื่อน
ยุค	หลายๆ	ภาพ	ครับ	ในประเทศ
สาเหตุ	เพียงแต่	ราย	คุณสมบัติ	กำจัด
อย่างจริงจัง	เลี้ยง	สูงสุด	ต่อมา	คล่องตัว
เตรียม	กรอบ	ส่วนตัว	ทหาร	คำจำกัดความ
ใช้งาน	กำกับ	หมู่บ้าน	นอกจากนั้น	ดังต่อไปนี้
ค้นหา	ก่อสร้าง	เกินไป	ปฏิบัติการ	ตรวจ

ธรรมดา	จูงใจ	ตก	ในขณะที่	สาร
ผูกขาด	หลังจาก	ตัวแปลภาษา	ๆ	สินเชื่อ
ภายใต้	อายุ	บรรยาย	การพัฒนา	สื่อสาร
อยากจะ	เพราะว่า	ผลิตภัณฑ์	ตรง	เพียงใด
เจตนา	เพื่อที่จะ	มาโคร	ถัดไป	แผนงาน
เช่นกัน	แนะนำ	ล้าน	ทรรศนะ	แผ่นดิน
เดือน	กระทบ	ส่วนหนึ่ง	ทีเดียว	กว้าง
เพิ่มเติม	กำไร	อเมริกา	บรรจุ	ขีดความสามารถ
เสรีภาพ	ข่าวสาร	เหมือนกับ	บรรลุ	จริยธรรม
ก่อให้เกิด	ชายฝั่ง	แน่	ย่อย	ซึ่งกันและกัน

ทำไม	เป็นจริง	มีส่วน	แผนพัฒนาฯ	ผู้ใช้
นักเรียน	เรียง	มูลค่า	ดังที่	วัสดุ
นิยม	แง่	ร้อย	ถือเป็น	สำนักงบประมาณ
ผลประโยชน์	ในปัจจุบัน	สัมมนา	ถ้าหาก	สิ่งของ
ภาระ	กรรมการ	หุ้น	สงคราม	อย่างเป็นระบบ
ยกเว้น	กล่าวถึง	อย่างยิ่ง	เคลื่อนไหว	อิสระ
วัฒนธรรม	คำถาม	อีกด้วย	คาดหวัง	กฎระเบียบ
สุขภาพ	ถ่ายทอด	เก่า	นักศึกษา	ทั่ว
อ้าง	บวก	เติบโต	น่า	มากนัก
เกือบ	พลัง	เทียบ	ป้าย	ระยะเวลา

สมรรถภาพ	ด้อย
สอบ	ทะเล
สะท้อน	นับว่า
สุข	น้ำมัน
หยุด	ผู้บริโภค
หลักเกณฑ์	พร้อมทั้ง
แต่ว่า	ริเริ่ม
ขณะนี้	วัตถุดิบ
ขัดแย้ง	เรื่อยๆ
ชิ้น	โรค
คุ้มครอง	ตั้งใจ

VITAE

Name Mr. Igor Stankovic

Student ID 5110130027

Educational Attainment

Degree	Name of Institution	Year of Graduation
Master in Electrical and Computer Engineering	University of Novi Sad, Serbia, Europe	2008

List of Publications and Proceedings

Stankovic, I., Karnjanadecha, M., and Delic, V. 2011. "Improvement of Thai speech emotion recognition by using face feature analysis," *Proceedings of the Nineteenth IEEE International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'2011)*, Chiang Mai, Thailand, December 7-9, 2011, pp. 1-5.

Stankovic, I., Karnjanadecha, M., and Delic, V. 2012. "Improvement of Thai speech emotion recognition using face feature analysis," *International Review on Computers and Software (IRECOS)*, Vol. 7, No. 5 (Impact factor: 6.14)