DETECTING EMOTIONAL SPEECH IN THAI DRAMA

Sawit Kasuriya[1,2], Tanakom Banchaditt[1], Nantanit Somboon[1], Thanaruk Teeramunkong[1], and Chai Wutiwiwatchai[2]

[1]School of Information, Computer, and Communication Technology
Sirindhorn International Institute of Technology
Pathum Thani, THAILAND
[2]Speech and Audio Technology Laboratory,
National Electronics and Computer Technology Center,
Pathum Thani, THAILAND
E-mail: sawitk@nectec.or.th, nntk@gmail.com, thanaruk@siit.tu.ac.th, chai.wutiwiwatchai@nectec.or.th

Abstract

This paper presents the experiments of implementing emotional speech classification in four classes of emotions (happiness, anger, sadness, and fear) by using Thai drama corpus. In this research, there are three main approaches for this classification: baseline system, emotional segmentation, and binary classification. The result of binary classification has shown the accuracy improvement with 53 percent. In addition, the last experiment was conducted to evaluate the performance of speaker independent speech classification by using another speech data as a test set. The results confirmed that the amounts data in training set was not enough for speaker independent modeling.

Keywords: emotional speech, Thai corpus, emotion classification

1. Introduction

Speech recognition is one of the amazing technologies. Nowadays, almost all of the machine can communicate with human via keyboard or mouse. Speech recognition has created a new way for machine to interact with human. While the speech recognition research field flourishes, another field of study has emerged; Emotion in speech recognition, which the growth rate of interest has spiked in these recent years.

We can sub-categorize the time-line of this field into two phases. In the first phase the researches are focus on how to make the machine act emotionally and the second phase is focus on how to make the machine recognize emotion from human [1]. One of the most popular ways in emotion recognition is facial expression detection. Besides human facial expressions, speech has been acknowledged as another feature that is effective for the automatic human emotion recognition [2]. This research focused on the emotional speech classification, which is one of the most challenge fields within speech processing.

The model itself can be applied and used by many organization, company, or individuals. For example, the program can be used to detect the moods of the customers who call to contact each department of the company and they can use such information to evaluate the department. If there are many angry customers' calls, the department should be checked and investigated if there is any problem. TV drama categorizing also receives benefits from the emotional speech classification. Since TV shows often have more than one episode for a series. To rate all episodes, it would take a lot of time and effort. Moreover, there would be many emotions in TV drama. Providing the precisely rating, each episode of TV drama should have an individual-rating label instead of a single rating label for a whole TV series or show.

About the classification engine, the most popular modeling technique in speech recognition is Hidden Markov models (HMMs). It was especially known for applications in temporal pattern recognition [1]. Emotional cues contained in an utterance cannot be assumed as specific sequential events in the signal. It can be at the start, at the middle, or at the end of the utterance. Because of these reasons, HMM is suitable for emotion recognition since every state can be reached in a single step from every other state. HMM is named as hidden because of the state of the model is hidden. However, we can estimate the possibility of the state by sequence of observation. More than one events associate

with each state and we must know every state and their relation. HMM cannot recognize things outside model and estimation of state along with several assumptions are required. If we can find best state sequence, we can compare multiple HMMs for recognition [3].

To recognize the speech of human, speech features are essential. Without them, we cannot do the process of recognition. However, there are many kinds of features. Selecting a suitable feature for a job is essential and need to be done wisely. Methods for estimating acoustic features that are frequently used in emotion classification are mel-frequency cepstral coefficients (MFCCs), linear prediction filter coefficients (LPCs), pitch, energy operator, and Vocal tract features [4].

This paper is organized into five sections: introduction, data resource, Thai emotional speech classification system, experiments and results, and conclusion. Next section is data resource, which explains about speech resource used in this research. Later section is giving some details about the classification system. There are three methodologies that were applied for emotional speech classification. The experiments and results will be described in the section IV. The last section is conclusion.

2. Data resource

In order to build the emotional classification, we have developed the Thai emotional speech corpus by collecting conversations of actors and actresses on TV drama. As developing Thai emotional speech corpus is an ongoing project since 2012, some emotional speech data in the corpus was used as training and test data in this research. Moreover, we also added more emotional speech data by recording volunteers' voice. The details of these emotional speech data are described in below.

2.1 Thai Emotional Speech Data

All emotional speech, collected from conversations by actors and actresses in a Thai TV drama, have been transcribed into text within time alignments as so called subtitle. Then we use the time alignments to divide conversations into utterances. These utterances were carefully defined to have only one emotion per utterance. The utterances with transcripts were labeled with four emotions: happiness, sadness, anger, and fear.

The annotator was asked to label emotions into speech without showing any moving pictures or video. Therefore, two cues (text and audio) used as information in annotation. There are eighteen speakers (actors and actresses) in this Thai TV drama. In fact, this research used 2,507 utterances of Thai emotional speech corpus. Since these utterances were selected from conversations in TV drama, there normally contains many background music and noise within the speech. For this research, we firstly focused on detecting emotions from clean speech rather than considering about background music. It is the most important to have the same probability distribution in training and test set. Another reason, we did not have a huge amount of speech data to implement the models for environment effects at this stage. The utterances having noise and/or background music were not considered. As that result, only 866 utterances from the corpus were utilized in this work. The details of each emotion are shown in Table 1.

Table 1. Number of emotions in Thai drama corpus

Happiness	Sadness	Anger	Fear	Total
501	91	232	42	866

2.2 Non-professional Voice Actors

According to lacking of emotional speech data, we have to collect some data for evaluating our system. The additional data were collected by recording acting voice of six volunteers who are non-professional actors. All volunteers were strictly instructed to perform the given emotion by reading the transcription. Each utterance represents one emotion. The length of each utterance is varied from two to five seconds. This additional data are 220 utterances, which are balanced in 55 utterances for four emotions as shown in Table 2.

Table 2. Number of emotional speeches in actor's voice

Happiness	Sadness	Anger	Fear	Total
55	55	55	55	220

3. Thai emotional speech classification system

We aim to create a system with the main function to detect the emotions in conversation especially in TV drama. Firstly, all speech data were divided into utterances, which are predefined as phrase or sentence containing an emotion from a speaker. After getting annotated utterances with emotion labels, we developed the baseline of emotional speech recognition by using HMM toolkit (HTK) [5].

This section explains how we developed and implemented our system.

3.1 Emotions

There are a variety of theories that attempt to identify the primary or basic emotions that categorized the feelings which all other emotions fall into. Robert Plutchik's theory [6] of emotion is one of them and it would be best to consider his theory of emotion because it is one of the most influential classification approaches for general emotional responses. His model does not only list the basic emotions but also highlights the relations between them. In Plutchik's model, the emotions are classified into eight basic emotions: happiness, trust, fear, surprise, sadness, disgust, anger, and anticipation. However, some emotions are ambiguity in expression and perception such as, disgust and anger, and also rarely occur in normal situation of conversations such as, trust, surprise, and anticipation. It is not easily able to collect enough the number of data for all that eight basic emotions. Then, we consider selecting only four distinguished emotions from Plutchik's model, which are happiness, sadness, anger, and fear for this research.

3.2 Methodology

There are three methodologies that have been implemented in this work. Firstly, the baseline system generally built to classify four emotions by using HMM as speech engine in emotional speech recognition. Secondly, since we defined the utterances in this work as the turn of actor's speaking, which must have only one emotion, but sometime the actor/actress may express many emotions in the same time or they have some problem in expression about emotional ambiguity. Then some utterances may have more than one emotion within there. To solve this problem, the system was allowed to have many emotions within an utterance. In the last method, we tried to improve the accuracy by applying binary classification with HMM.

3.2.1 Baseline System

Figure 1. Baseline System

The baseline system (as shown in Figure 1) was developed by using HMMs to classify four emotions: happiness, sadness, anger, and fear. In this system, the input utterance was clearly assumed to contain only an emotion with it. For feature extraction, Mel-Frequency Cepstral Coefficients (MFCC) was chosen for this classification.

3.2.2 HMM training with segmented utterences

For this approach, the input utterances were assumedly contained more than one emotion as mentioned in previous section. The utterances were segmented into pieces of emotions according to the experiment in the next section. The system was also allowed to answer difference emotions in an utterance freely. For example, the input utterance was divided into five segments and all of segments were labeled as "Fear", but the output of classification might independently answers five different emotions like "Fear, Sad, Sad, Fear, and Fear" as displayed in Figure 2.

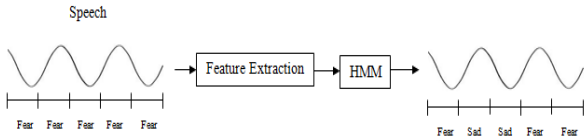

Figure 2. HMM training with segment approach

3.2.3 Binary classification

Applying the concept of binary classification, we created a HMM that can classify whether the utterance contains the particular emotion or not. There are four HMMs for this approach as depicted in Figure 3. For example, a HMM for happiness emotion is used to classify the input utterance having happiness or not. The output of each model will give two probabilities between having the particular emotion and not having that emotion. This classification system is resulted by comparing the probabilities of all four models. The model that gives the highest probability is the answer.

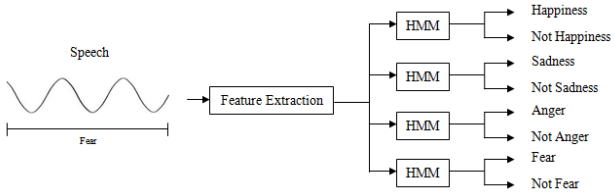

Figure 3. Binary classification approach

4 Experiments and results

Three experiments were set to evaluate the system. Beginning with the experiment varying the number of states and Gaussian mixtures of HMM is to find the optimal point for the emotional speech recognition system. Later on considering about emotional segmentation was the second experiment. Finally, we applied the binary classification technique with HMM to improve the accuracy of the system.

The speech data that was used as training and test set for the following experiments are described in details of each class of emotions in Table 3. The number of speech data in test set is around 30% of speech data. The happiness class is a large number of utterances while the fear is the least number of utterances in this emotional speech data.

Table 3. Number of emotional speech in training and test set

	Class of Emotions				
	Happiness	**Sadness**	**Anger**	**Fear**	**Total**
Training	372	79	190	24	665
Test	129	12	42	18	201
All	501	91	232	42	866

4.1 HMM with different numbers of states and Gaussian mixtures

Normally, in speech recognition, a 5-state HMM is usually used for phoneme modeling in speech recognition task. The 5-state HMM might be not suitable for classifying emotions in the whole sentence, which contains several phonemes. Therefore, we conducted this experiment to find the optimal number of HMM states and also that of Gaussian mixtures. This experiment was set to vary the number of HMM states starting at 8 to 32 states and the number of Gaussian mixtures from 1 to 16 mixtures.

In general, the number of Gaussian mixtures are represented the number of clustering. More clustering would be more accuracy, but it usually takes much time to process the system. In some case that training data is not much enough to provide all statistical probabilities of the distinguish features in clustering. For that reason, more number of clustering will decrease the accuracy. This problem is so called over-fitting [6].

The feature utilized in this research was MFCC since it showed the best accuracy in speech recognition. The results of this experiment are shown in Table 4. From the table, the accuracy reached the best at 50.75% when using 16 and 32 states of HMM with two mixtures. Then, we decided to use 16 states of HMM and two Gaussian mixtures for the rest experiments.

Table 4. Accuracy results of varying the number of HMM states and Gaussian mixtures (in percentage)

Feature	**State of HMM**	**Number of Gaussian mixtures**				
		1	***2***	***4***	***8***	***16***
MFCC	***8***	21.89	35.82	45.27	45.75	45.27
	16	41.79	**50.75**	41.29	40.30	45.77
	32	42.76	**50.75**	48.76	49.25	48.26

4.2 HMM with different numbers of segments

This experiment was conducted to find the appropriate number of emotional segmentation in utterances. In general, only one emotional state is represented in an utterance. As mentioned in previous section, we allowed an utterance having one to five emotional states as shown in Table 5. According to the previous experiment, the 16 states of HMM and two Gaussian mixtures were selected as the best accuracy parameter for this experiment.

Table 5. Accuracy results of HMM training with various number of segments (in percentage)

Number of segments				
1	***2***	***3***	***4***	***5***
50.75	46.27	42.79	41.67	40.9

From Table 5, the best accuracy of the emotional speech classification was one emotional segment. It could imply that almost all utterances were perfectly annotated with only one emotional state. Since having more emotional state in an utterance, it affected less accuracy following the number of emotional segments.

4.3 Binary classification HMM

In this experiment we used the utterances from Thai drama corpus without emotional segmentation to train sixteen-state and two-Gaussian mixtures HMMs for binary classification as mentioned in section III. Then we tested these four classes of HMMs with two data sets from Thai drama corpus and non-professional voice actor. Assumedly, the first test data set is represented for

speaker dependent evaluation and another set is provided for speaker independent evaluation.
The results of this binary classification are shown in Table 6. The accuracy of emotional classification can be improved by nearly 2.50% comparing with the baseline system. The accuracy of this classification dropped to 22% when testing in speaker independent.

Table 6. Accuracy results of binary classification HMM

Thai Drama Corpus	Nonprofessional Voice Actor
53.00%	22.00%

To analyze these results in details, we applied confusion matrix into the results of two data sets. From Table 7, it is confusion matrix for binary classification in Thai Drama Corpus. The highest accuracy of emotional class was happiness. On other hand, the sadness and fear classes had worse accuracies at 0.12 and 0.11, respectively. However, when focusing on the error of classification in anger, the most missed match was happiness. Moreover, the most error of predicted class for happiness was sadness. The emotional state between anger and happiness are quite different in state of pleasure. The happiness and sadness states are also totally stood in opposite point of pleasure. The question of these errors comes up with what features of these two opposite emotional states have shared together.

Table 7. Confusion matrix of binary classification with Thai drama corpus

Actual class	**Predicted class**			
	Happiness	*Anger*	*Sadness*	*Fear*
Happiness	0.63	0.14	0.21	0.01
Anger	0.42	0.58	0	0
Sadness	0.5	0.25	0.12	0.12
Fear	0.11	0.22	0.55	0.11

The confusion matrix in Table 8 is for binary classification with data from non-professional voice actor. This data set is balanced in the number of each emotional state as described in Table 2. The highest accuracy of this classification was happiness and the worst was fear. The interesting point of this confusion matrix is that almost anger state was predicted to be happiness state by 68%. More than an half of fear states were predicted to be happiness. However, this non-professional voice actor data set was not in the part of training set. This mismatch between emotional states may not strongly support evidences as same as the results of Table 7, but it would be the guideline of some explanation for future works.

Table 8. Confusion matrix of binary classification with non-professional voice actor

Actual class	**Predicted class**			
	Happiness	*Anger*	*Sadness*	*Fear*
Happiness	0.56	0.24	0.16	0.04
Anger	0.68	0.32	0	0
Sadness	0.28	0.44	0.24	0
Fear	0.56	0.32	0.12	0

5 Conclusion

This paper has reported the research work on emotional speech classification by using speech from Thai drama corpus. There are three methods (baseline system, emotional segmentation and binary classification) that were applied implementing in this work. The results of our experiments have shown that the binary classification with no emotional segmentation provided the best accuracy, which was 53 percent. In addition, another test set named non-professional voice actor was used to evaluate the binary classification in the last experiment. The result of these extra speakers, whose utterances do not contain in training set, was not greater than 25 percent. We can assume that our emotional classification was not having enough amounts of data for training speaker independent models. However, this research did not consider about meaning of words in transcripts. The word meaning would improve the accuracy of emotional classification.

References

[1] Bjorn Schuller, Anton Batliner, Stefan Steidl, Dino Seppi, Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, Speech Communication Vol.53, pp. 1062–1087, 2011.

[2] Alexander I. Iliev, Michael S. Scordilis, Joao P. Papa, Alexandre X. Falcao, Spoken emotion recognition through optimum-path forest classification using glottal features, Computer Speech and Language Vol.24, pp. 445–460, 2010.

[3] John-Paul Hosom, Speech Recognition with Hidden Markov Models, CS 552/652 Lecture 4, pp. 1-36, 2011.

[4] Chul min Lee, Automatic Recognition of Emotions from the Acoustic Speech Signal

[5] Jitendra Ajmera, Iain McCowan, Herve Bourlard, Speech/music segmentation using entropy and dynamism features in a HMM classification framework, Speech Communication Vol.40, pp. 351–363, 2003.

[6] Tin Lay Nwe, Say Wei Foo, Liyanage C. De Silva, Speech emotion recognition using hidden Markov models, Speech Communication Vol.41, pp. 603–623, 2003

[7] Tim Polzehl, Alexander Schmitt, Florian Metze, Michael Wagner, Anger recognition in speech using a coustic and linguistic cues, Speech Communication Vol.53, pp. 1198–1209, 2011.

[8] Dimitrios Ververidis, Constantine Kotropoulos, Emotional speech recognition: Resources, features, and methods, Speech Communication Vol.48, pp. 1162–1181, 2006.

[9] Louis ten Bosch, Emotions, speech and the ASR framework, Speech Communication Vol.40, pp. 213–225, 2003

[10] B.Yang, M.Lugger, Emotion recognition from speech signals using new harmony features, Signal Processing Vol.90, pp. 1415–1423, 2010.

[11] Jia Rong, Gang Li , Yi-Ping Phoebe Chen, Acoustic feature selection for automatic emotion recognition from speech, Information Processing and Management Vol.45, pp. 315–328, 2009

[12] Silke Paulmann, Marc D. Pell, Sonja A. Kotz, How aging affects the recognition of emotional speech, Brain and Language Vol.104, pp. 262–269, 2008

[13] Moataz ElAyadi, MohamedS.Kamel, Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognition Vol.44, pp. 572–587, 2011

[14] Alex S. Cohen, S. Lee Hong, Alvaro Guevara, Understanding emotional expression using prosodic analysis of natural speech: Refining the methodology, J. Behav. Ther. \& Exp. Psychiat. Vol.41, pp. 150–157, 2010

[15] John-Paul Hosom, Speech Recognition with Hidden Markov Models, CS 552/652 Lecture 4, pp. 1-36, 2011.

[16] Phil Woodland, Gunnar Evermann, Mark Gales, The HTK Book, Cambridge University Engineering Department 2006