ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

The Comparison of Thai Speech Emotional Features for LSTM Classifier

Choopan Rattanapoka[1], Monkon Duangdoaw[2] and Noppanut Phetponpun[3]

ABSTRACT

The human voice, in its various tones, can effectively express human emotions. It is undeniable that human emotions significantly influence our daily lives. Many studies have attempted to improve machines' ability to comprehend human emotion to develop better Human-Computer Interaction (HCI) applications. As a result, this study presents the design and development of models for emotion recognition from Thai male speech. We examined the utilization of the chromagram (Chroma), Mel spectrogram, and Mel frequency cepstral coefficient (MFCC) with seven Long-Short Term Memory (LSTM) networks to distinguish four emotions: anger, happy, sad, and neutral. Additionally, we created a dataset consisting of 1,000 audio files recorded from 11 Thai males, divided into 250 audio files per emotion. Subsequently, we trained our seven models using this dataset. Our findings revealed that the model utilizing only the MFCC feature yielded the best results, with precision, recall, and F1 scores of 0.730, 0.739, and 0.732, respectively.

Article information:

Keywords: Machine Learning, Deep Learning, Audio Features, MFCC

Article history:
Received: January 28, 2022
Revised: June 10, 2023
Accepted: October 26, 2023
Published: November 25, 2023 (Online)

DOI: 10.37936/ecti-cit.2023174.247471

1. INTRODUCTION

The tone of the human voice can indicate human feelings. The ability to perceive the feelings of human beings' sound can be beneficial in many ways, whether to make a machine that can comfort sad people, to check customer feelings when talking to customer service employees, or to develop better HCI applications. Thus, the speech emotion recognition field becomes an active and interesting research topic in the past few years. However, speech emotion recognition is a highly challenging problem as the vocal features capable of distinguishing emotions remain unclear.

Nowadays, deep learning models are becoming widely popular. There is a deep learning model called Convolutional Neural Network (CNN) [1] that is good at recognizing images. Examples of applications that use this type of deep learning model are a real-time gender classification from facial images [2], a tomato disease and pest detection [3], or even a recognition of twelve human-like sign-language actions [4]. Moreover, there is another type of deep learning model that is appropriate for tasks involving continuity of time (voice data or text information). This model is called Recurrent Neural Network (RNN) [5]. For example, W. Khan, et al. [6] present a deep recurrent neural network model with word embeddings for recognizing Urdu named entities. From their experiment, this model achieved the best F1 score at 81.1%. However, RNN gives poor results when the network is too deep because of the loss of information in deeper layers. Later on, it was further developed into Gated Recurrent Unit (GRU) [7] and Long-Short Term Memory (LSTM) [8]. Examples of applications using LSTM include the use of the bidirectional long short-term memory (BLSTM) network to automatically assess the proficiency of Korean speech [9]. Also, F. Ertam [10] uses a deep LSTM network to recognize a person's gender from voice data. The result shows that it is better than other traditional models. Moreover, E. Swedia, et al. [11] designed a model for recognizing numbers from Indonesian numeral speech with Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficient (MFCC) feature on the LSTM network. Their model can achieve a prediction accuracy of 96.58%.

The Thai language is one of the languages that use intonation to distinguish the meaning of each word; thus, Thai has an obligatory lexical tone. Eash sylla-

[1,2,3] The authors are with College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Thailand, E-mail: choopan.r@cit.kmutnb.ac.th, s6003051613133@gmail.com and noppanut2553@gmail.com

ble pronounces with one of five tones: low, mid, high, falling, or rising. The tone must be spoken correctly for the intended meaning of a word to be understood. Thus, recognizing emotion in Thai speech may be difficult due to these intonations. Therefore, we need to learn and find a solution to recognize emotion in Thai speech. Unfortunately, a dataset for Thai speech is not yet available.

In this paper, our objectives are to present the design and development of a deep learning model for recognizing emotion from Thai speech, specially focusing on Thai male speech using LSTM networks. Because we want to concentrate on the male voice first to find the best model, rather than dealing with the difference between male and female tones simultaneously. Also, we have prepared our dataset of 1,000 Thai male audio files recorded by 11 individuals and divided them into four emotions: angry, sad, happy, and neutral (250 files for each emotion). From the total of 1,000 audio files, we divided the 800 audio files for the training set, 100 audio files for the validation set, and 100 audio files for the test set. Furthermore, we experiment with seven LSTM models using a combination of sound features, including MFCC, Mel spectrogram, and chromagram values, as their input to determine which features yield the best results.

2. BACKGROUNDS

To design and implement a speech emotion recognition system based on LSTM networks, which are deep learning models, we need to study a concept of deep learning and LSTM. Then, we need to understand the audio features that can be used as input to the model. Finally, we have to study the programming libraries that are used to create a deep learning model and to extract audio features from audio.

2.1 Deep Learning and LSTM

Deep learning is a field of artificial intelligence and machine learning that mimics the human brain's activity in data processing. It is relatively new area of machine learning research that focuses on learning multiple levels of representation and abstraction to make sense of data, including images, sound, and text. There are several types of deep learning networks. The popular network for analysing images is the convolutional neural network (CNN), which divides the image into sub-areas of each pixel for analysis. It is particularly suitable for computer vision tasks. Another type of deep learning network is the recurrent neural network (RNN), which is suitable for processing sequential data such as sound and text.

In this paper, we focus on a deep learning model called long short-term memory (LSTM), which is based on recurrent neural network architecture. However, LSTM has an advantage over RNN, which is the insensitivity of gap length. In RNN, the network cannot handle a long series of data because it will start to forget what it learned at the starting point. The structure of an LSTM cell, as shown in Fig. 1, helps to address the problems that RNNs face.

Each LSTM cell has additional parameters and a system of gating units that control the flow of information. The long-term memory is typically referred to as the cell state, enabling the storage of information from previous intervals within the LSTM cell.

In Equation (1), the current cell state ($c(t)$) is equal to the previous cell state ($c(t-1)$) multiplied by the output of the forget gate (f) and adds new information through the output of the input gates (i), adjusted by the input modulation gate (g).

$$c(t) = c(t-1) * f + g * i \tag{1}$$

The forget gate (f) determines which information from the previous interval will be forgotten. The input gate (i) and the input modulation gate (g) together determine which information should enter the cell state. The output gate (o) determines which information in the new state will be sent as output. Equation (2) – (5) reprensent the formulas for calculating the forget gate, input gate, input modulation gate, and output gate, respectively. The $h(t–1)$ refers to the output of the previous interval, whereas $x(t)$ is the input. The term W_f, W_i, W_g, W_o and b_f, b_i, b_g, bo represent the weights and biases of the forget gate, input gate, input modulation gate, and output gate, respectively. Finally, the output of the LSTM unit ($h(t)$) is calculated by Equation (6).

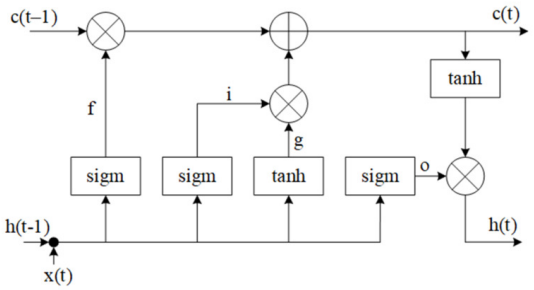

Fig.1: *Structure of an LSTM unit.*

$$f = \sigma(W_f[h(t-1), x(t)]) + b_f \tag{2}$$

$$i = \sigma(W_i[h(t-1), x(t)]) + b_i \tag{3}$$

$$g = \tanh(W_g[h(t-1), x(t)]) + b_g \tag{4}$$

$$o = \sigma(W_o[h(t-1), x(t)] + b_o \tag{5}$$

$$h(t) = o * \tanh(c(t)) \tag{6}$$

2.2 Audio Features

Audio features or sound characteristics are acoustic attributes that can be utilized in sound analysis tasks. In this paper, we are concentrating on distinguishing the emotions of Thai male speeches. Emotion should not be determined solely by the loudness of speech; rather, it should be based on the characteristics of sound frequency. Thus, there are three main audio features that we are interested in using to analyze and recognize emotions in Thai male speeches: chromagram (chroma), Mel spectrogram (Mel), and Mel Frequency Cepstral Coefficient (MFCC).

The chroma [12] is related to the twelve different pitch classes. The main property of chroma is to capture the harmonic and melodic characteristics of music. The twelve chromas are represented by the set of the twelve pitch spelling attributes used in music notation, which are C, C♯, D, D♯, E, F, F♯, G, G♯, A, A♯, and B.

The Mel spectrogram represents an acoustic time-frequency representation of a sound in the Mel scale, which was named by Stevens, Volkmann, and Newman in 1937 [13]. The name "Mel" comes from the word "melody", indicating that the scale is based on pitch comparisons. Also, The Mel scale is a frequency scale that closely matches to what the human ear can hear. Thus, in Mel scale, the difference between 500 and 1000 Hz is noticeable, whereas the difference between 7500 and 8000 Hz is barely noticeable.

In sound processing, the Mel-frequency cepstrum (MFC) [14] represents the short-term power spectrum of a sound. It is based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel frequency cepstral coefficients (MFCCs) are the coefficients that collectively constitue an MFC. MFCC considers human perception for sensitivity at appropriate frequencies by converting the conventional frequency to the Mel scale. Thus, it is well suitable for speech recognition tasks.

2.3 Programming Libraries

The implementation of models in this paper is written in Python. There are three main programming libraries that we used to implement our models: TensorFlow, Keras, and Librosa.

TensorFlow [15] is an open-source machine learning library created by the Google Brain team. It bundles deep learning models, algorithms, and supports Python. TensorFlow can run on multiple CPUs and GPUs, and it is available on 64-bit Linux, macOS, Windows, and mobile computing platforms.

Keras [16] is an open-source library that provides a Python interface for artificial neural networks and acts as an interface for the TensorFlow library. Until version 2.3, Keras supported multiple backends, including TensorFlow, R, Theano, etc. Keras contains numerous implementations of commonly used neural network components, such as layers, activation functions, and optimizers making it easy for developers to write code for deep neural networks.

Librosa [17] is a Python open-source library for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Additionally, it can be used to extract audio features that are utilized in the model implementation, such as MFCC, chroma, and Mel.

2.4 Related Works

In the past few years, there has been research on speech emotion recognition. Y. Xie, et al. [18] proposed an automatic speech emotion recognition using attention-based LSTM that distinguishes six categories of emotion: anger, fear, happy, neutral, sad, and surprise on CASIA [19] and eNTERFACE [20] dataset. Also, they tested their model with GEMEP [21] dataset, which has 12 emotion categories. The result showed that their model can outperform other traditional models.

B. Atmaja and M. Akagi [22] also proposed speech emotion recognition using LSTM with an attention model that categorized emotions into four categories: anger, excited, neutral, and sadness, using IEMOCAP [14] dataset. Their model achieved 65% - 70% of accuracy. Meanwhile, Z. Zhu, et al. [23] proposed a speech emotion recognition model based on Bi-GRU and focal loss, capable of distinguishing six emotions on IEMOCAP dataset. According to their experiments, the model achieved around 70% accuracy.

Also, J. Zhao, X. Mao, and L. Chen [24] compared a traditional deep belief network and a CNN with their model, which combines 2D CNN and LSTM network for speech emotion recognition. The results from their experiment show that their models outperform traditional models, achieving 95% accuracy on the Berlin EmoDB dataset and 89% accuracy on the IEMOCAP dataset. Moreover, W. Lim, D. Jang, and T. Lee [25] proposed the combination between CNN and RNN in their model to recognize emotion from German speech into seven categories: anger, fear, disgust, sadness, boredom, happy, and neutral. Their model can achieve an average accuracy of 88%.

For predicting emotion from audio speech, U. Garg, et al. [26] used MFCC, MEL and Chroma as feature vectors on multiple machine learning models. They found that random forest gave the best result in their experiments. However, the neural network model used in the research only employed simple dense layers and 1-D convolutional layers. F. Abri et al. [27] proposed several regression models to predict arousal and valence for predicting emotions from sounds. They used Emo-soundscapes, which provided numerous audio features, including MFCC and Chroma. Finally, A.Jara Jaratrotkamjorn and A. Choksuriwong [28] used 11 audio features (MFCC, Chroma, etc.) and face landmarks in a deep belief

network to achieve better results.

3. METHODOLOGY

The overall architecture of our proposed system is shown in Fig.2, which users can use via two options.

(1) The first option is the command line, developed with Python. Users must provide a directory name containing audio files as an argument. Then, the command line will operate the emotion recognition process on each audio file in that directory. The result of the operation is a CSV file containing the percentage of each emotion that appears on each audio file, including the dominant emotion.

(2) The second option is the desktop application, developed with Python and TKinter library. The users can choose an audio file. Then, the application will analyze that audio file and display recognized emotions every 3 seconds time interval (with a 1-second overlap). The application also shows the overview of all the emotions contained in the audio file in a donut chart format.

The process of speech emotion recognition is shown in Fig. 3. The process begins with the input, which is the speech in the .wav audio format. Then, we extract features from the audio input and, if necessary, pad the audio input to ensure the size of audio features is consistent for all inputs. Next, we design, train, and evaluate deep learning models. Finally, we pass the audio features to deep learning models for classifying emotions. In this paper, we focus on four emotions: angry, sad, happy, and neutral.

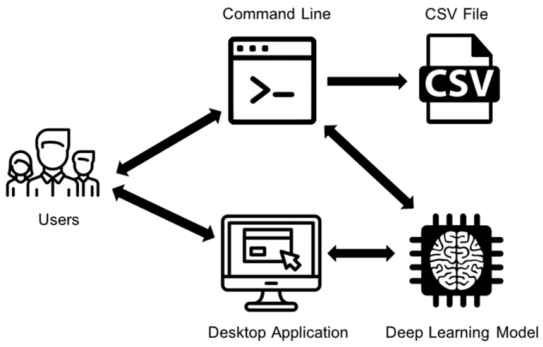

Fig.2: *System architecture.*

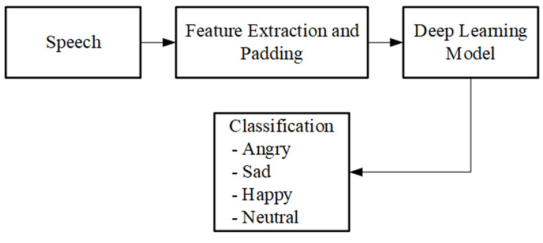

Fig.3: *Speech emotion recognition process.*

3.1 Dataset Preparation

To prepare the dataset, we enlisted 11 Thai men as participants to record audio files. Each of the ten participants recorded 12 utterances in 4 emotions through simulation, repeating each emotion twice. As a result, we have a total of 960 audio files with 240 audio files per emotion. The last participant spoke freely for 10 utterances in 4 emotions (1 time per emotion). Consequently, we obtained an additional 40 audio files (10 audio files per emotion). In total, our dataset comprises 1,000 audio files, with 250 audio files per emotion, and the maximum utterance length is 3 seconds.

Notably, all the audio files are recorded using mobile phone without any environment control. Some audio files may contain noise, which we believe is beneficial for deep learning models to extract audio features when used in real-world environments.

3.2 Audio Feature Extraction and Padding

In this paper, we use a library called Librosa library to extract the audio features. For the Librosa audio feature extraction parameters, we set the audio sample rate (sr) to 44,100 Hz and the hop length of 512 samples. According to Equation (7), where N_{sample} is the number of samples, s is the audio length in seconds, and sr is the audio sample rate. Thus, the number of samples for each audio is 259 samples. However, for audio files with a length less than 3 seconds, we apply the zero paddings at the end of the audio to ensure the number of samples reaches 259 samples.

$$N_{sample} = \frac{s \times sr}{hop\_length} \tag{7}$$

Then, we extract the MFCC, Mel, and chroma values from each sample. Each kind of audio feature has a different number of values (M). MFCC has 40 values, Mel has 128 values, and chroma has 12 values. For example, an audio file with 3 seconds is transformed into an MFCC feature matrix of the size of (M $\times$ Nsample) or (40 $\times$ 259).

We try to find the best audio features for our model, so we combine these three audio features and implement seven models called Model 1 to Model 7. The detail of each model is shown in Table 1.

3.3 Model Designing and Training

We design our models as shown in Fig. 4. The main components of the models consist of 8 layers as follows:

(1) An LSTM layer with 128 hidden units, approximately half the length of the number of samples, which takes the audio feature matrix as input.

(2) A Dropout layer with a 0.3 or 30% dropout rate.

Table 1: *Models and their audio features.*

Model	Audio Features	Number of Feature Value (M)	Size of Audio Features
Model 1	MFCC	40	(40 × 259)
Model 2	Mel	128	(128 × 259)
Model 3	chroma	12	(12 × 259)
Model 4	MFCC and Mel	168	(168 × 259)
Model 5	MFCC and chroma	52	(52 × 259)
Model 6	Mel and chroma	140	(140 × 259)
Model 7	MFCC, Mel, and chroma	180	(180 × 259)

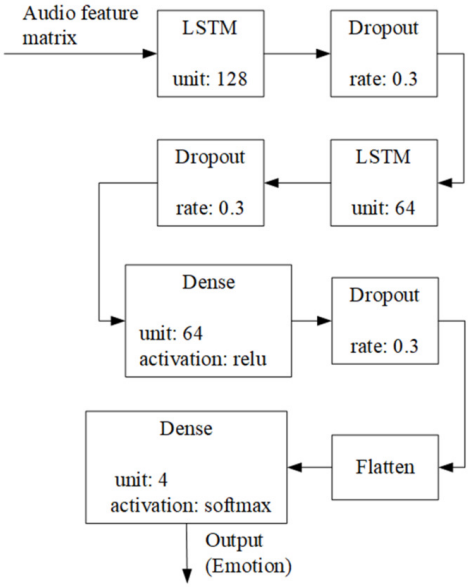

Fig.4: *Model architecture.*

(3) An LSTM layer with 64 hidden units to reduce the size of the 2nd LSTM layer by half from the 1st LSTM layer.
(4) A Dropout layer with a 0.3 dropout rate.
(5) A Dense layer with 64 hidden units using RELU as an activation function.
(6) A Dropout layer with a 0.3 dropout rate.
(7) A Flatten layer to transform 2-dimension data to 1-dimension data.
(8) A Dense layer with 4 hidden units using softmax as an activation function to output an emotion.

Dropout layers are included to prevent model overfitting. We chose a 0.3 or 30% dropout ratio as we believe it is a suitable ratio to avoid forgetting too many learning parameters. However a more thorough investigation should be conducted to determine an optimal dropout ratio. All seven models share the same architecture, with the only difference being the input size of the first LSTM layer, which depends on the size of the audio feature matrix (M × 259) for each model.

Then, we created our models by using Python with Tensorflow and Keras libraries. In the model learning process, we devided our 1,000 audio dataset into 800 audio for training, 100 audio for validation, and 100 audio for testing. We set the optimizer in the learning process to RMSprop with an accuracy metric, a learning rate of 0.00001, and categorical cross-entropy as the loss function. Each of our models was trained for 500 epochs with a batch size of 16. Additionally, we configured the model checkpoint to store the model weights from the epoch that produced the best validation loss value to a file for later use in the model evaluation step.

3.4 Model Evaluation

There are terms used to compare the result of the classifier (model) with the actual result when evaluating the model of classification tasks. The terms "true" and "false" are related to whether the model output (classifier's prediction) matches the actual result, while the terms "positive" and "negative" refer to whether the model output (classifier's prediction) matches the actual result.

Thus, a true positive (TP) implies that the model correctly predicts the positive output, while a *true negative (TN)* implies the model correctly predicts the negative output. A *false positive (FP)* implies the model incorrectly predicts the positive output, and a *false negative (FN)* implies that the model incorrectly predicts the negative output.

Then, we measure the model performance by using the model evaluation metrics, namely precision, recall, and F1 score.

Precision is the ratio between the true positives and all the positives. It is used to find out what proportion of positive output was correct, as shown in Equation (8).

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall is the ratio between the true positives and all correct predictions. It is used to find out what proportion of actual positives was identified correctly, as shown in Equation (9).

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

The F1 score is the harmonic mean of precision and recall. The F1 score gives the value that balances between the precision and the recall, as shown in Equation (10).

$$F1scroe = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

4. EXPERIMENT RESULTS

4.1 Model Evaluation Results

We have trained our models on the Google Colab [29]. For each model, the model training time of 500 epochs took approximately 3 hours. The accuracy of the validation set of all seven models is depicted in Fig. 5. Notably, Model 3, utilizing only chroma as an audio feature, exhibits significantly lower performance. Its accuracy on the validation set is around 40%. Model 1, employing only MFCC as an audio feature, and Model 5, utilizing both MFCC and chroma as audio features, demonstrated the highest accuracy on the validation set, achieving approximately 70% to 75%.

The validation loss of all seven models is shown in Fig.6. We found that the loss on the validation set of Model 3 using only chroma as an audio feature did not significantly improve during 500 training epochs. Once more, Model 1 relying solely on MFCC as an audio feature, and Model 5, incorporating both MFCC and chroma as audio features, exhibit the most favorable loss on the validation set. However, from the 100th to the 200th epoch, all models except Model 3 demonstrate signs of overfitting.

Then, we input 100 audio files (25 files per emotion) from the test set to all seven models. After that, we used the confusion matrix [30] to gain insight into the prediction results. The top two models that exhibited the best performance are Model 1 and Model 5. Their respective confusion matrices are depicted in Fig. 7 and Fig. 8. Examing the confusion matrices, it is evident that the prediction accuracy of Model 1, leveraging only MFCC as an audio feature, for four emotions is approximately 68%. This model demonstrates a high accuracy in predicting the emotion sadness. Conversely, the prediction results of the Model 5, incorporating both MFCC and chroma as audio features, exhibit high accuracy for the emotions of happiness and sadness, reaching around 88% accuracy. However, the accuracy of predicting the emotion of happiness is relatively low, standing at around 50%.

Table 2. presents a summary of precision, recall, and F1 scores for each model, measured with 100 audio files on the test set. The results reveal that the most proficient model in predicting the emotion "angry" is Model 5, achieving precision, recall, and F1 scores of 0.840 each for emotion "angry". In predicting the emotion of "happy," Model 1 emerges as the top performer, attaining precision, recall, and F1 scores of 0.640, 0.552, and 0.593, respectively. Additionally, Model 1 excels in predicting the emotion of "neutral" with precision, recall, and F1 scores of 0.720, 0.750, and 0.735, respectively. However, for the emotion "sad," the best prediction model is Model 5, achieving precision, recall, and F1 scores of 0.920,

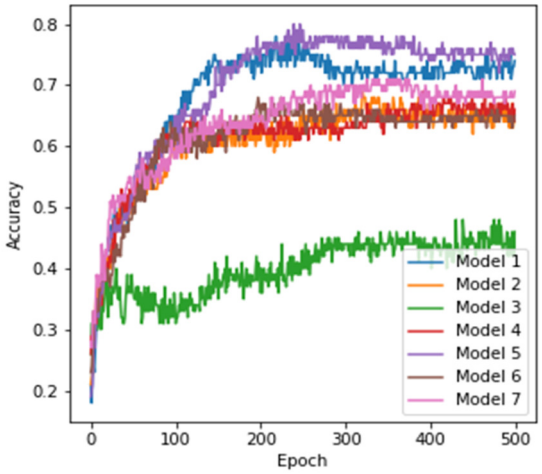

Fig.5: *Accuracy during the training on the validation set.*

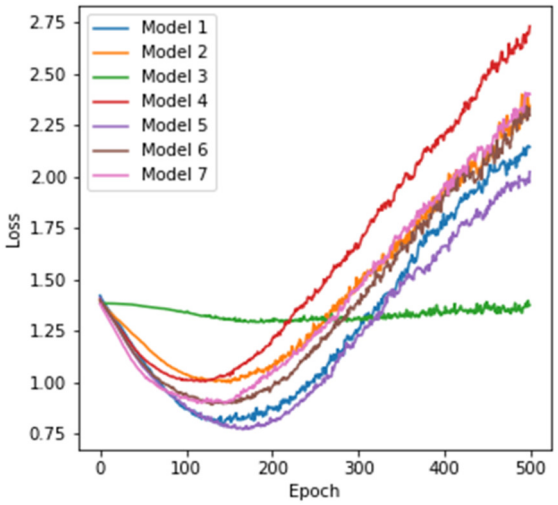

Fig.6: *Loss during the training on the validation set.*

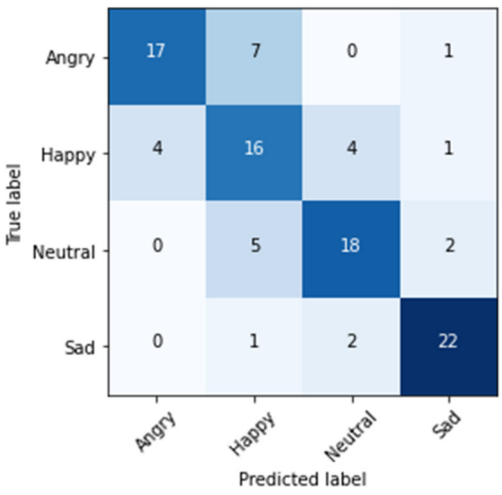

Fig.7: *Confusion matrix of the Model 1.*

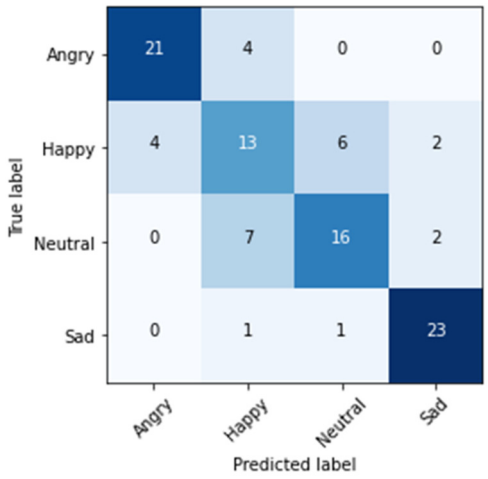

Fig.8: *Confusion matrix of the Model 5.*

Table 2: *The result of models' evaluations.*

Model name	Emotion	Precision	Recall	F1 score
Model 1	Angry	0.680	0.810	0.739
	Happy	**0.640**	**0.552**	**0.593**
	Neutral	**0.720**	**0.750**	**0.735**
	Sad	0.880	0.846	0.863
	Average	**0.730**	**0.739**	**0.732**
Model 2	Angry	0.760	0.704	0.731
	Happy	0.440	0.500	0.468
	Neutral	0.640	0.533	0.582
	Sad	0.640	0.762	0.696
	Average	0.620	0.625	0.619
Model 3	Angry	0.200	0.263	0.227
	Happy	0.240	0.240	0.240
	Neutral	0.520	0.464	0.491
	Sad	0.600	0.536	0.566
	Average	0.390	0.376	0.381
Model 4	Angry	0.680	0.607	0.642
	Happy	0.440	0.524	0.478
	Neutral	0.600	0.556	0.576
	Sad	0.760	0.792	0.776
	Average	0.620	0.620	0.618
Model 5	Angry	**0.840**	**0.840**	**0.840**
	Happy	0.520	0.520	0.520
	Neutral	0.640	0.696	0.667
	Sad	**0.920**	**0.852**	**0.882**
	Average	**0.730**	0.727	0.728
Model 6	Angry	0.760	0.679	0.717
	Happy	0.480	0.480	0.480
	Neutral	0.480	0.480	0.480
	Sad	0.680	0.773	0.723
	Average	0.600	0.603	0.600
Model 7	Angry	0.760	0.731	0.745
	Happy	0.530	0.542	0.531
	Neutral	0.480	0.545	0.511
	Sad	0.760	0.679	0.717
	Average	0.630	0.624	0.626

0.852, and 0.882, respectively.

When evaluating the overall performance of the models, both Model 1 and Model 5 exhibit the best precision, achieving a score of 0.730. However, Model 1 excels in overall recall with a score of 0.739. Consequently, Model 1 attains the best overall F1 score at 0.732. In summary, Model 1, leveraging only MFCC as an audio feature, emerges as the optimal model for Thai speech emotion recognition in our test data.

4.2 Desktop Application

The desktop application for Thai male emotion recognition was developed using Python and the TK-inter library. The user interface is depicted in Fig. 9. Users can click the "Browse" button on the top right to select an audio file (.WAV) for processing. Then, users click the green "Calculate" button to analyze the speech emotion of the audio file. The application then displays the result of speech emotion recognition, as illustrated in Fig.10.

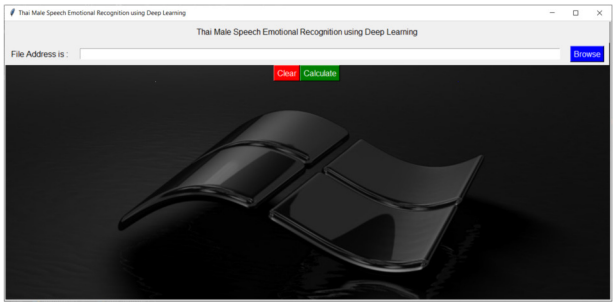

Fig.9: *Desktop application for Thai male speech emotion recognition.*

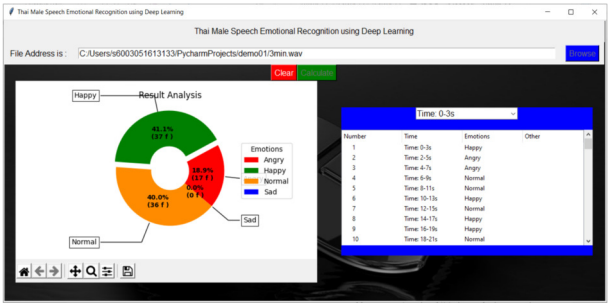

Fig.10: *Speech emotion recognition result of the desktop application.*

The audio file undergoes segmentation into several parts, each with a length of a 3-second time interval and a 1-second overlap. Subsequently, each audio segment undergoes processing by the LSTM deep learning model for emotion recognition.

The result displayed in the desktop application includes a donut chart on the left side, showcasing the overall emotions identified in the audio file, as illustrated in Fig. 11. The color scheme employs red

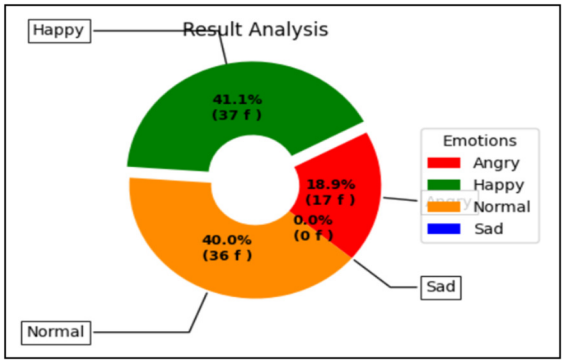

Fig.11: *Overall emotion found in the donut chart format.*

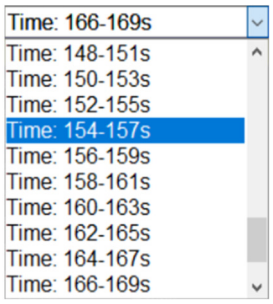

Fig.12: *Drop-down list for listening to the audio part in each interval.*

Number	Time	Emotions	Other
1	Time: 0-3s	Happy	
2	Time: 2-5s	Angry	
3	Time: 4-7s	Angry	
4	Time: 6-9s	Normal	
5	Time: 8-11s	Normal	
6	Time: 10-13s	Happy	
7	Time: 12-15s	Normal	
8	Time: 14-17s	Happy	
9	Time: 16-19s	Happy	
10	Time: 18-21s	Normal	

Fig.13: *Speech emotion recognition results in each interval.*

for the emotion “angry,” green for “happy,” blue for “sad,” and orange for “neutral” or “normal.”

The drop-down list on the right side of the result screen, as shown in Fig.12, contains the list of each audio part in each time interval. Users can select the interval to listen to the audio during that specific duration.

Lastly, the table on the right side of the result screen, as shown in Fig.13, displays the emotion that the LSTM recognized in each time interval.

4.3 Command Line

The desktop application is suitable for users requiring in-depth speech emotion analysis of individual audio files. However, in scenarios where users need to analyze the overall speech emotion across multiple files simultaneously, the command line provides an alternative solution. We have implemented a Python

	A	B	C	D	E	F
1	File name	Angry	Happy	Normal	Sad	Emotion
2	Actor\08-01-01-01.wav	0.00%	0.00%	100.00%	0.00%	Normal
3	Actor\08-01-01-02.wav	0.00%	0.00%	100.00%	0.00%	Normal
4	Actor\08-01-02-01.wav	0.00%	100.00%	0.00%	0.00%	Happy
5	Actor\08-01-02-02.wav	0.00%	50.00%	50.00%	0.00%	Happy
6	Actor\08-01-03-01.wav	0.00%	0.00%	50.00%	50.00%	Normal
7	Actor\08-01-03-02.wav	0.00%	0.00%	50.00%	50.00%	Normal
8	Actor\08-01-04-01.wav	50.00%	0.00%	50.00%	0.00%	Angry
9	Actor\08-01-04-02.wav	100.00%	0.00%	0.00%	0.00%	Angry
10	Actor\08-02-01-01.wav	0.00%	0.00%	100.00%	0.00%	Normal
11	Actor\08-02-01-02.wav	0.00%	0.00%	100.00%	0.00%	Normal
12	Actor\15sec.wav	28.57%	14.29%	42.86%	14.29%	Normal
13						

Fig.14: *Output of command line in the CSV format.*

script named *cmd_emotions.py*, which accepts the directory name containing audio files as an argument. Upon executing the command line, the Python script processes each audio file in the specified directory using the LSTM deep learning model for speech emotion analysis. Subsequently, the output is generated in the form of a CSV file. The CSV file comprises six columns, as illustrated in Fig.14, including:

(1) filename
(2) percentage of emotion “angry” in that audio file
(3) percentage of emotion “happy” in that audio file
(4) percentage of emotion “neutral” or “normal” in that audio file
(5) percentage of emotion “sad” in that audio file
(6) the dominant emotion in that audio file.

5. CONCLUSIONS

We developed and assessed seven LSTM models aimed ar recognizing four emotions—sad, angry, happy, and neutral—in Thai male speech. The input features for these LSTM models are derived from audio characteristics, specifically MFCC, Mel, and chroma, all extracted using Librosa library. All seven LSTM models follows a uniform architecture, comprising two LSTM layers succeeded by two dense layers, with dropout layers interspersed between each LSTM and dense layers. Furthermore, we trained and evaluated the models using our audio dataset.

The experimental findings for Thai male speech emotion recognition employing the LSTM classifier reveal that the superiority of the model using solely MFCC as audio. This model outperforms counterparts utilizing Mel, chroma, or the combination of the three audio features, achieving notable metrics in precision (0.730), recall (0.739), and F1 score (0.732). Additionally, we observed that utilizing only chroma as an audio feature did not yield satisfactory performance in our model. We believed this limitation to the inadequacy of only 12 feature vectors for our complex model, where audio pitch struggles to effectively discern emotional nuances.

Furthermore, we implemented both a command line and desktop application utilizing our model for emotion recognition in audio clips. In future research, we plan to expand our dataset to include female voice samples and a broader spectrum of emotions. Further experimental will explore alternative audio features, such as pitch and the teager energy operator (TEO), and their impact on model performance. We also aim to explore hybrid approaches by combining our model with existing architectures to identify the most suitable model for Thai speech emotion recognition. Moreover, we aspire to apply these models to practical scenarios, particularly in real-world applications like customer services, where the system's ability to discern customer emotions during the calls can enhance service quality.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, Art. no. 7553, May 2015.

[2] S. S. Liew, M. Khalil-Hani, S. Ahmad Radzi, and R. Bakhteri, "Gender classification: a convolutional neural network approach," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 24, no. 3, Art. no. 40, pp. 1248-1264, 2016.

[3] J. Liu and X. Wang, "Tomato Diseases and Pests Detection Based on Improved Yolo V3 Convolutional Neural Network," *Frontiers in Plant Science*, vol. 11, 2020, Accessed: Feb. 02, 2022. [Online]. Available: `https://www.frontiersin.org/article/10.3389/fpls.2020.00898`

[4] Y. Ji, S. Kim, Y.-J. Kim, and K.-B. Lee, "Human-like sign-language learning method using deep learning," *ETRI Journal*, vol. 40, no. 4, pp. 435–445, 2018.

[5] David E. Rumelhart; James L. McClelland, "Learning Internal Representations by Error Propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, MIT Press, pp.318-362, 1987.

[6] W. Khan, A. Daud, F. Alotaibi, N. Aljohani, and S. Arafat, "Deep recurrent neural networks with word embeddings for Urdu named entity recognition," *ETRI Journal*, vol. 42, no. 1, pp. 90–100, 2020.

[7] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv:1406.1078 [cs, stat]*, Sep. 2014, Accessed: Feb. 02, 2022. [Online]. Available: `http://arxiv.org/abs/1406.1078`

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[9] Y. R. Oh, K. Park, H.-B. Jeon, and J. G. Park, "Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition," *ETRI Journal*, vol. 42, no. 5, pp. 761–772, 2020.

[10] F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Applied Acoustics*, vol. 156, pp. 351–358, Dec. 2019.

[11] E. R. Swedia, A. B. Mutiara, M. Subali, and Ernastuti, "Deep Learning Long-Short Term Memory (LSTM) for Indonesian Speech Digit Recognition using LPC and MFCC Feature," in *2018 Third International Conference on Informatics and Computing (ICIC)*, pp. 1–5, Oct. 2018.

[12] R. N. Shepard, "Circularity in Judgments of Relative Pitch," *The Journal of the Acoustical Society of America*, vol. 36, no. 12, pp. 2346–2353, Dec. 1964.

[13] S. S. Stevens, J. Volkmann, and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, Jan. 1937.

[14] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern Recognition and Artificial Intelligence*, pp. 374–388, 1976.

[15] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," arXiv:1603.04467 [cs], Mar. 2016, Accessed: Feb. 02, 2022. [Online]. Available: `http://arxiv.org/abs/1603.04467`

[16] F. Chollet and others, "Keras," 2015. `https://github.com/fchollet/keras`

[17] B. McFee *et al.*, *librosa/librosa: 0.8.0. Zenodo*, 2020. doi: 10.5281/ZENODO.3955228.

[18] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou, and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1675–1685, Nov. 2019.

[19] S. Pan, J. Tao, and Y. Li, "The CASIA audio emotion recognition method for audio/visual emotion challenge 2011," in *Proceedings of the 4th international conference on Affective computing and intelligent interaction - Volume Part II*, Berlin, Heidelberg, pp. 388–395, Oct. 2011.

[20] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 Audio-Visual Emotion Database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*,Atlanta, GA, USA, pp. 8–8, Apr. 2006.

[21] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception," *Emotion*, vol. 12, no. 5, pp. 1161–1179, Oct. 2012.

[22] B. T. Atmaja and M. Akagi, "Speech Emotion

Recognition Based on Speech Segment Using LSTM with Attention Model," in *2019 IEEE International Conference on Signals and Systems (ICSigSys)*, pp. 40–44, Jul. 2019.

[23] Z. Zhu, W. Dai, Y. Hu, and J. Li, "Speech emotion recognition model based on Bi-GRU and Focal Loss," *Pattern Recognition Letters*, vol. 140, pp. 358–365, Dec. 2020.

[24] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, Jan. 2019.

[25] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–4, Dec. 2016.

[26] U. Garg, S. Agarwal, S. Gupta, R. Dutt and D. Singh, "Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma," *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, Bhimtal, India, pp. 87-91, 2020.

[27] F. Abri, L. F. Gutiérrez, A. Siami Namin, D. R. W. Sears and K. S. Jones, "Predicting Emotions Perceived from Sounds," *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, pp. 2057-2064, 2020.

[28] A. Jaratrotkamjorn and A. Choksuriwong, "Bimodal Emotion Recognition Using Deep Belief Network," *ECTI-CIT Transactions*, vol. 15, no. 1, pp. 73–81, Jan. 2021.

[29] E. Bisong, "Google Colaboratory," in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, E. Bisong, Ed. Berkeley, CA: Apress, pp. 59–64, 2019.

[30] K. M. Ting, "Confusion Matrix," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds. Boston, MA: Springer US, pp. 260–260, 2017.

Choopan Rattanapoka is a lecturer in Electronics Engineering Technology Department, College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Thailand. He completed B.Eng from Kasetsart University, Thailand in 2000. Then, he obtained M.Sc and Ph.D. from Strasbourg University, France seubsequently in 2004 and 2008. His research interests cover distributed computing, cloud computing, big data, and machine learning.

Monkon Duangdoaw is an undergraduate student on the Bachelor's degree of Engineering Technology in Eletronics Engineering Technology (Computer), Department of Electronics Engineering Technology, College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Thailand. His research interest is in the field of computer engineering.

Noppanut Phetponpun is an undergraduate student on the Bachelor's degree of Engineering Technology in Eletronics Engineering Technology (Computer), Department of Electronics Engineering Technology, College of Industrial Technology, King Mongkut's University of Technology North Bangkok, Thailand. His research interest is in the field of computer engineering.