Bimodal Emotion Recognition Using Deep Belief Network

Apichart Jaratrotkamjorn[1] and Anant Choksuriwong[2]

ABSTRACT: Emotions are essential in daily human life. Our goal is to make a machine which can recognize the human emotional state, and which can intelligently respond to the needs of humans. Accomplishing this is very important to support human-computer interaction (HCI). The majority of existing work concentrates on the classification of six basic emotions only. This research work proposes a bimodal emotion recognition system to be used for human emotion detection. In this method, we used facial landmarks combined with a Gabor filter bank for the extraction of facial features. We also used the pyAudioAnalysis library for extraction of speech features. Both facial and speech features are fused at feature-level and forwarded to the Deep belief network for the classification of eight basic emotions. Finally, we used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) in the experiment, for both unimodal and bimodal emotion recognition. The our results show that the performance of bimodal emotion recognition using the deep belief network (DBN) that was tested on the RAVDESS database in the classification of eight basic emotions with facial and speech information achieved an overall accuracy rate of 97.92%, which is better than unimodal emotion recognition (facial or speech expression).

Keywords: Human-Computer Interaction (HCI), Emotion recognition, Face Detection, Speech Detection, Feature level fusion, Deep Belief Network (DBN)

DOI: 10.37936/ecti-cit.2021151.226446
Received November 22, 2019; revised February 26, 2020; accepted May 18, 2020; available online January 13, 2021

1. INTRODUCTION

Emotions are essential in daily human life. Many psychologists see emotional importance. They have researched human emotions, and they have created many different emotional models (discrete emotion models, dimensional models, meaning oriented models, and componential models) [1]. There are two models found in most research: discrete and dimensional [2]. The dimensional models are an emotional model that has both 2D and 3D models. Each emotion is continuous, and it is not separated clearly. For example, in the 2D model, emotion is determined on 2D space (Valance and Arousal). Valance shows the polarity of emotions (positive vs. negative). Arousal shows the intensity of emotions (low vs. high). Thus, the dimensional models and discrete emotion models are different because discrete emotion models are the basic emotions that separate each emotion clearly. Each emotion has no continuity. The variety of theories causes the number of emotions in each group to be different, according to each method and principles used to explain it, including differences in the objectives of psychologists. Therefore, both models are used for human emotion detection. Emotion recognition is applied in many areas such as robots, automobile safety, animations, affective computing, human-computer interaction (HCI), etc [3-5].

Human emotion can be detected with many approaches. For example, emotion detection can be done through analysis of facial expressions, speech, body gestures, and physiological signals. Previous work found that combining two modalities gives better performance than using only one single modality [6]. Nevertheless, data integration has different features such as facial and speech expression, facial expression and body gesture, and speech expression and body gestures. Previous works found the integration of two modalities also had little research [7]. Most of the previous approaches focus on decision level fusion methods or model level fusion methods, and feature level fusion methods are seldom utilized [8]. Very few papers discussed the classification of eight emotions

[1,2]Department of Computer Engineering, Faculty of Engineering, Prince of Songkla University, Songkhla 90110, Thailand., E-mail: 5910120080@email.psu.ac.th and anant.c@psu.ac.th

[9, 10].

In this paper, we propose a bimodal emotion recognition system using multimodal data including facial and speech features fused at the feature-level, which makes features into a high-dimensional feature vector. The deep belief network method is suitable for large datasets with complex structures and is high-dimensional. It can associate relationships of data by inferring the probability of data. The deep belief network approach was used for the classification of eight basic emotions (neutral, calm, happy, sad, angry, fearful, disgust, and surprised) based on the RAVDESS database. It was used for the training and testing model by using the deep belief network approach. The experimental results display that the integration of two modalities at the feature level allows for higher performance and a better emotion recognition rate. The remainder of the paper is organized as follows: The bimodal emotion recognition system based on facial and speech expression, including database, number emotion, classification method, and experimental results are discussed briefly in Section 2. Section 3 explains the overview of the system and the database used for the experiment. Sections 4 and 5 describe the feature extraction method and the classification method. Section 6 shown results from the experiment. Finally, the conclusion is shown in Section 7.

2. LITERATURE REVIEW

The research work is about a bimodal emotion recognition system based on facial and speech expression by using the audio-visual signals in analyzing human emotions. The emotion classification method is briefly discussed.

Busso et al. [11] presented the strengths and weaknesses of an emotion recognition system based on facial expression or acoustic information, including the combination of two modalities at the decision level and feature level. They used a database recorded by an actress. The audio-visual information performed the classification of four emotions using a support vector machine (SVM). It was found that the bimodal emotion classification had better performance than either of the unimodal systems. The result from classification at the decision level was 89.00% and the feature level had an accuracy of 89.01%.

Wang et al. [12] proposed an emotion recognition system using audiovisual information. The facial features were extracted by a Gabor filter bank while the speech features were extracted by prosodic, Mel-frequency Cepstral Coefficient (MFCC), and formant frequency. The combination of facial and speech features was concatenated directly into a single long feature vector. The stepwise method was used in order to select the essential features for the classification of six emotions with Fisher's Linear Discriminant Analysis (FLDA) approach. The overall performance of the system achieved 82.14% accuracy.

Yan et al. [8] presented a novel bimodal emotion recognition approach from facial and speech expression. The openSMILE software was used to extract the speech emotional features. The Scale Invariant Feature Transform (SIFT) was used to extract emotional features from facial expression. The Sparse Kernel Reduced-Rank Regression (SKRRR) fusion approach is a combination of facial and speech features. Support Vector Machine (SVM) and Sparse Representation (SR) approaches were used for the classification of six emotions from the eNTERFACE'05 database. The results from classification emotions by the SVM approach achieved 87.02% accuracy, and the SR method achieved 87.46% accuracy. This was better than the bimodal emotion recognition rate among some state-of-the-art approaches.

Nguyen et al. [13] proposed a novel approach which combined 3 Dimensional Convolution Neural Networks (C3Ds) and Deep Belief Networks (DBNs) for multimodal emotion recognition from facial and speech expression. The audio-visual information was extracted by employing C3Ds and DBNs approaches. It was used for the classification of six basic emotions on the eNTERFACE database at score level fusion. Overall performance of emotion recognition achieved 89.39% accuracy that was better than the state-of-the-art approach by about 2%.

Miao et al. [9] proposed the multimodal emotion recognition system from facial expression and acoustic signals. The Chinese Natural Audio-Visual Emotion Database (CHEAVD 2.0) consists of eight emotions. It was used for the experiment. The toolkit OpenSMILE was used for the extraction of speech features and the convolution neural network (CNN) was used for the extraction of facial features from Audio-Video information. Classification of human emotion used traditional machine learning (SVM, REPTree, Random Forest) and deep learning (DBN, RNN) approaches. The overall performance of the system achieved 41.89% accuracy and 33.74% macro average precision (MAP) on the test dataset.

Xu et al. [10] presented emotion recognition based on the integration of facial expression and human voices. The experiments used the Chinese Natural Audio-Visual Emotion Database (CHEAVD 2.0) whose data consists of eight emotions. The toolkit OpenSMILE was used for the extraction of the audio signal. Geometric features and histogram of gradient features (HOG) were used instead of facial expression information. The facial and speech features were classified with the SVM approach. Then, the classification results were integrated at the decision level using Bayesian rules. The emotion at the highest score value was selected as the final output. The overall performance of multimodal emotion recognition achieved an accuracy of 38.41% for the dataset of CHEAVD 2.0.

Prasada Rao et al. [14] analyzed the performance of a bimodal emotion recognition system based on facial and speech features. The decision level fusion used the SVM approach in the classification of six basic emotions. The facial features did gender classification by using the AdaBoost algorithm. The Indian Face Database and Berlin Speech Database were used in the experiment. Results from the experiment have an overall accuracy of 78%. This method had 2-3% better performance than the unimodal recognition.

3. METHODOLOGY

We propose a new bimodal emotion recognition system. The data used in the experiment is comprised of video and audio datasets from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [**?**]. In the facial feature extraction process, we used a Gabor filter bank and facial landmarks for the extraction of the specific facial features that imply emotions from the facial expressions in the video dataset (eight basic emotions). Furthermore, in the speech feature extraction process, we used the pyAudioAnalysis library to extract the specific speech features that imply emotions from speech expression in the audio dataset (eight basic emotions). Then, we used the technique of feature-level fusion (Early fusion) to increase performance of emotion recognition by used two features (facial and speech) with deep belief network (DBN) method. Therefore, these features were forwarded to the DBN method to classify the eight basic emotions. Our software was implemented in three parts: Emotion detection through facial expression, speech expression, and bimodal data, as shown in Fig. **??**.

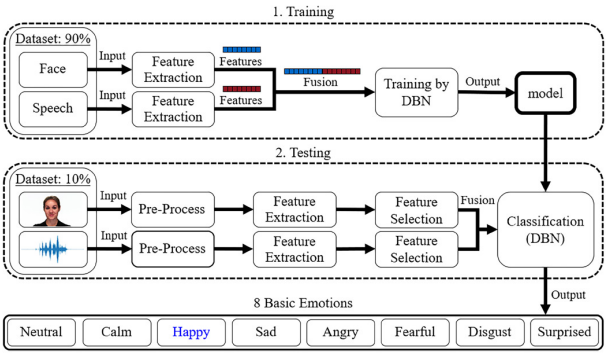


Fig.1:: Overview of Bimodal Emotion Recognition.

3.1 Database

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a relatively new database, published on May 16, 2018 [15]. We chose to use this database for the training and testing model. It consists of two main parts: facial and speech expression. The RAVDESS database has a total size of 24.8 GB. There are three forms of data (Audio-Only, Audio-Video, and Video-Only), consisting of 7,356 files. Each file is approximately three seconds long. All three forms are divided into two types. The first type is the song. There is a dynamic emotion expression of six basic emotions, such as neutral, calm, happy, sad, angry, and fearful. All six basic emotions have the intensity level of different emotions with two levels (normal and strong level) by using 23 professional actors (11 females and 12 males). The age of the actors was between 21-33 years. The second type is speech. There is a dynamic emotion expression of eight basic emotions such as neutral, calm, happy, sad, angry, fearful, disgust, and surprised. All eight basic emotions have the intensity level of different emotions with two levels (normal and strong level) by using 24 professional actors (12 females and 12 males). The age of the actors was between 21-33 years. See Table 1 for details.

Table 1:: RAVDESS Database

No.	Form	Detail	Type	Emotions	Number of Files
1	Audio-Only	16bit, 48kHz (.wav)	Song	6	1,012
			Speech	8	1,440
2	Audio-Video	720p H.264, ACC 48kHz (.mp4)	Song	6	1,012
			Speech	8	1,440
3	Video-Only		Song	6	1,012
			Speech	8	1,440
Total Number of Files					7,356

We choose the forms of Audio and Video data in types of speech, which have eight basic emotions. The emotion classification through facial expression used video data, which is 1,440 video files. It consists of 96 samples of neutral emotion, and 192 samples of calm, happy, sad, angry, fearful, disgust, and surprised. The emotion classification through speech expression used audio data, stored in 1,440 audio files. It consists of 96 samples of neutral emotion, and 192 samples of calm, happy, sad, angry, fearful, disgust, and surprised. Finally, both audio and video datasets were used in the section of emotion classification through bimodal data.

4. FEATURE EXTRACTION

4.1 Facial Features

We now describe the facial feature extraction method used for the video dataset. The dataset was loaded into the system. It consists of the video files obtaining from the RAVDESS database. The video is separated into frames inside the system. The frames brought into the system were used to detect the face position and 68 positions of the facial landmarks by

using the Dlib library [16]. Fig.2 and Table 2 present details of the facial landmarks. When the position of a face from an image was found we cut the face area from the image to convert the data from a color image to the standard format of a grayscale image. It was used in the facial feature extraction process by using a Gabor filter bank method that joins with 68 positions of the facial landmark (The facial landmarks are adjusted to the cropped image). The facial feature extraction from the image is done by bringing the grayscale image of the face through the Gabor filter bank process that has four scales (4, 7, 10, 13) and eight orientations (0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°) and can make a total of 32 patterns as shown in Fig. 3, 4, and 5 respectively. After that, 32 patterns (Magnitude part) are combined with 68 positions of the facial landmark to determine the one position of the facial landmark consisting of 32 features. Therefore, there are 2,176 (32 x 68) facial features per frame in a video. After that, the facial features are collected and the process is repeated until all features in one video have been extracted, according to Fig. 6. Next, we repeat the same process, according to Fig. 6, until we have finished (1,440 videos, eight basic emotions). Finally, bringing together all features we calculated the average of feature vectors in each frame of the video until we finished processing all 1,440 video files. Therefore, one-row data of the features contains 2,176 features for a video. Then, 1,440 videos yields 1,440 data rows.

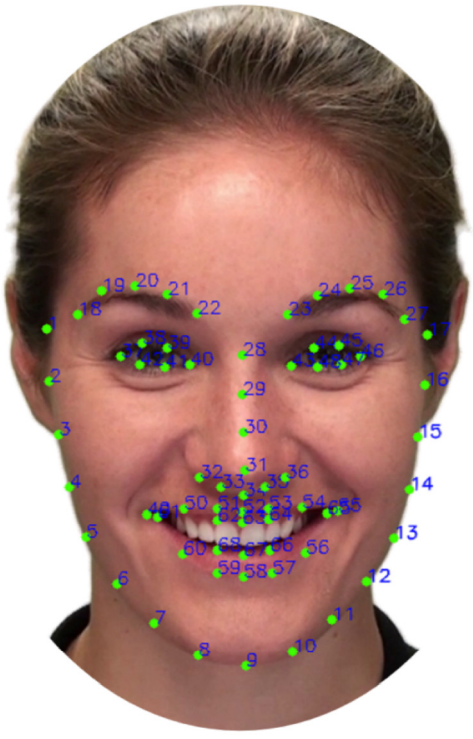


Fig.2:: 68 Facial Landmarks

Table 2:: Details of 68 Facial Landmarks

Geometric Features	Point Range	Description
Jaw	1-17	Face Contour Landmark
Right Eyebrow	18-22	Eyebrows Landmark
Left Eyebrow	23-27	
Nose Bridge	28-31	Nose Landmark
Lower Nose	32-36	
Right Eye	37-42	Eyes Landmark
Left Eye	43-48	
Outer Lip	49-60	Mouth Landmark
Inner Lip	61-68	

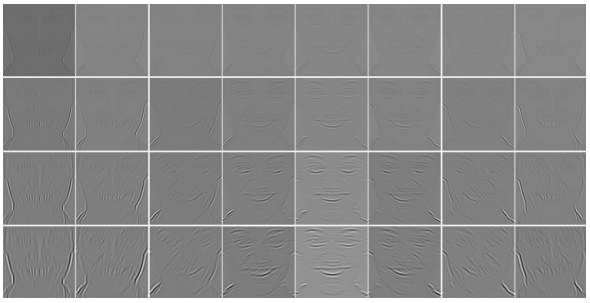


Fig.3:: Real Part using Gabor Wavelet

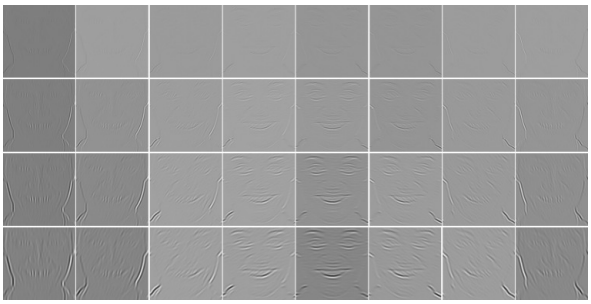


Fig.4:: Imaginary Part using Gabor Wavelet

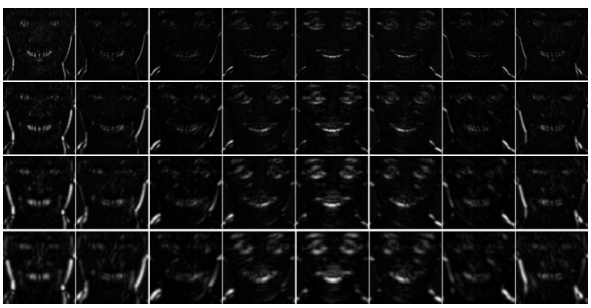


Fig.5:: Magnitude Part using Gabor Wavelet

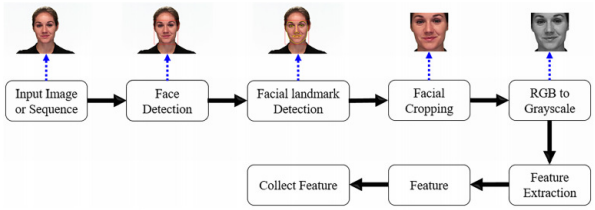


Fig.6:: Facial Feature Extraction Process

4.2 Audio Features

The speech feature extraction approach from the audio dataset will now be described. The information put into the system was an audio signal. The audio signal was divided into several frames (short-term windows) by using the pyAudioAnalysis library [17] with the short-term and non-overlapping time frame data. Each frame was calculated with 34 features, as shown in Table 3.

Table 3:: Audio Features [17]

Index	Name	Description
1	Zero Crossing Rate	The rate of sign-changes of the signal during the duration of a particular frame.
2	Energy	The sum of squares of the signal values, normalized by the respective frame length.
3	Entropy of Energy	The entropy of sub-frames' normalized energies. It can be interpreted as a measure of abrupt changes.
4	Spectral Centroid	The center of gravity of the spectrum
5	Spectral Spread	The second central moment of the spectrum.
6	Spectral Entropy	Entropy of the normalized spectral energies for a set of sub-frames.
7	Spectral Flux	The squared difference between the normalized magnitudes of the spectra of the two successive frames.
8	Spectral Rolloff	The frequency below which 90% of the magnitude distribution of the spectrum is concentrated.
9-21	MFCCs	Mel Frequency Cepstral Coefficients form a cepstral representation where the frequency bands are not linear but distributed according to the mel-scale.
22-33	Chroma Vector	A 12-element representation of the spectral energy where the bins represent the 12 equal-tempered pitch classes of western-type music (semitone spacing).
34	Chroma Deviation	The standard deviation of the 12 chroma coefficients.

In Table 3, audio features are divided into two groups. The first group has the features 1 to 3 which are related to the time domain features. The second group has the features 4 to 34 which are related to the frequency domain features. After that, 34 features were used to calculate the average and standard deviation resulting in 68 features of statistical values per audio signal. This data was then collected resulting in the features as shown in Fig. 7. Next, we repeated the same method until we finished processing all of the audio files (1,440 audios, eight basic emotions), according to Fig. 7. Therefore, one row of data of the features is equal to 68 features per audio sample. Then, 1,440 audio samples yielded 1,440 data rows.

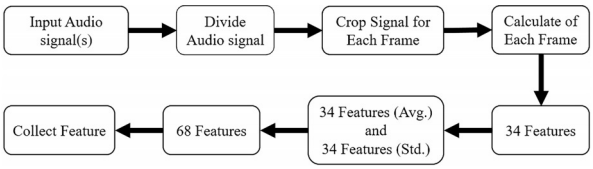


Fig.7:: Speech Feature Extraction Process

4.3 Bimodal Feature

In this section, we used the facial and speech features from the previous facial and speech feature extraction processes mentioned in section four. Therefore, the bimodal features (facial and speech features) were used together, which can be summarized as in Table 4.

Table 4:: Summary, All Features used for Training and Testing Model

No.	Feature	Number	Features used for Train and Test model
1	Face	68	2,176
2	Speech	34	68
Total Bimodal Features			2,244

In Table 4, The facial and speech features have been combined at the feature-level. Both of the two modalities (facial and speech features) were concatenated directly in the form of a single long feature vector (a single high-dimensional feature vector) [18] by using the feature-level fusion technique (Early fusion), as shown in Fig. 8, before forwarding it to the deep belief network (DBN) method in order to be applied to emotion recognition.

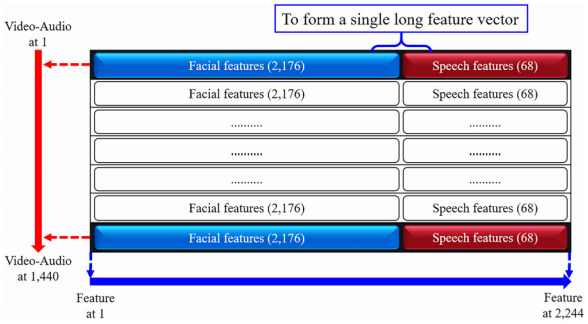


Fig. 8:: Feature-Level Fusion

5. CLASSIFICATION

The classification of eight basic emotions used a deep belief network method. We separated the classification into three parts. There is emotion classification through facial expression, speech expression, and bimodal data. First we brought all features through the feature scaling process to adjust the data range. Then all features went through a normalization process and all data was divided into two parts. The first part used 90% for the training model. In the second part, 10% was used for the testing model by classification of eight emotions with the deep belief network approach, as shown in Fig. 9.

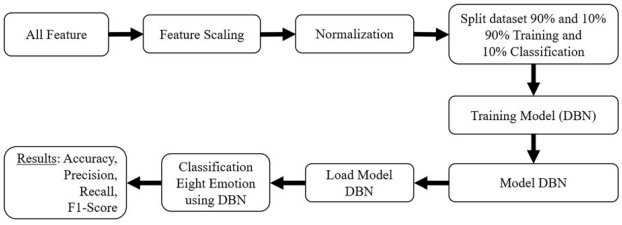


Fig. 9:: Procedure to Classify Eight Basic Emotions

The Deep Belief Network (DBN) [19-21] is a probabilistic generative model. It contains many layers of binary random variable nodes, and each layer consists of undirected graphical models, which are called Restricted Boltzmann Machine (RBM). The RBM has two layers (visible and hidden layers). The visible layer consists of visible nodes and the hidden layer is composed of hidden nodes. Therefore, each visible node connected to a hidden node. The weight value has a two-way direction movement to help reduce the mistakes. Then DBN constructed from the multiple RBM. Thus, the hidden layer of the first RBM became a visible layer of the next layer of RBM. This model can increase the effectiveness of machines in the classification of emotions and could also handle noise and missing values of features by probabilistic inference, as shown in Fig. 10.

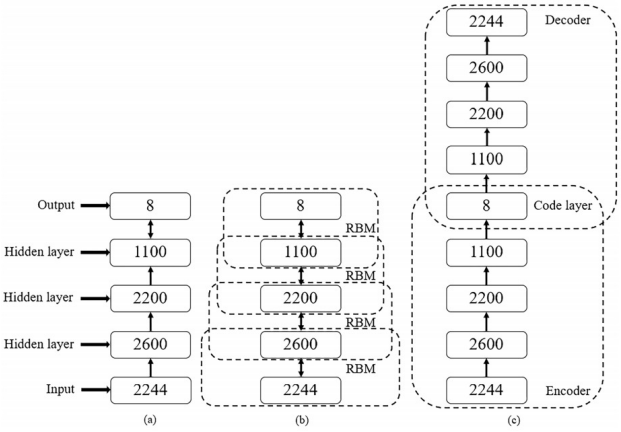


Fig. 10:: Adapted from [22]. (a): The structure of DBN. (b): Separating DBN into RBM for pre-training. (c): Unrolling DBN into DA for fine-tuning

Fig. 10 shows the structure of DBN used for emotion classification through facial expression, speech expression, and bimodal data. There are five layers. Layer 1 is the input layer, layers 2, 3, and 4 are the hidden layers, and layer 5 is the output layer. Layers 2, 3, and 4 have different nodes that depended on the input layer based on the identification of learning parameters. For example, the learning rate and the iteration of the pre-training phase, batch size, activation function, dropout, and hidden node amount affected the model accuracy. The final layer was the output layer that performed the eight emotion classification.

6. EXPERIMENT RESULTS

6.1 Face Emotion Recognition

Fig. 10 exhibited two main parts of DBN. The first part (b) was the pre-training. The second part (c) was fine-tuning. Thus, the structure of DBN has a size of 2176-2200-1800-900-8. For the pre-training phase, the learning rate was 0.08, and the number of iterations was 6. For the fine-tuning phase, the learning rate was 0.5, and the number of iterations was 296. The batch size was 48. The activation function was the Rectified Linear Unit (ReLU). The dropout was 0.2. These parameters were used for the training model. It was used for the classification of eight basic emotions through facial expression. The results are displayed in Fig. 11. The overall performance of the system has an accuracy value at 96.53%, precision of 96.61%, recall of 96.53%, and f-measure of 96.55%. Of 144 samples, 139 samples were classified correctly, but five samples had the wrong classification. The diagonal components of the confusion matrix indicated that all emotions were recognized with more than 95.00% accuracy. The happy, angry, fearful, and disgusted emotions were recognized with very high accuracy. On the other hand, the sad emotion is less accurate than other emotions. The sad emotion

often was misclassified as calm. On the other hand, the calm emotion often was misclassified as sad.

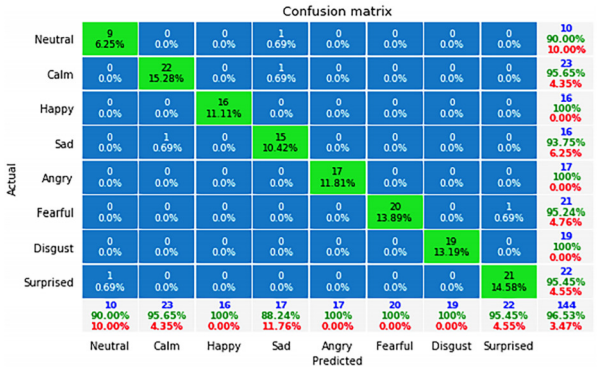


Fig. 11:: Confusion Matrix of Face Emotion Recognition

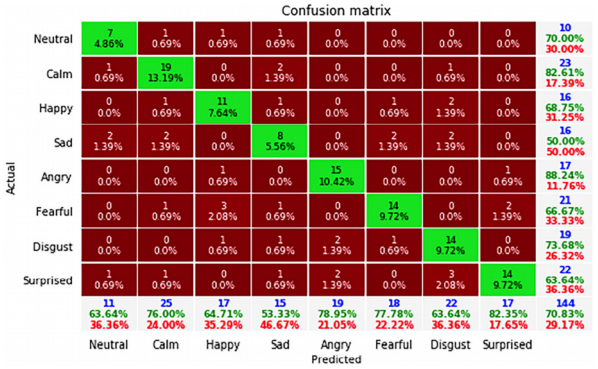


Fig. 12:: Confusion Matrix of Speech Emotion Recognition

6.2 Speech Emotion Recognition

Emotion classification through speech expression by the deep belief network approach will be discussed next. The structure of DBN has a size of 68-389-262-120-8. For the pre-training procedure, the learning rate was 0.08, and the number of iterations was 8. For the fine-tuning procedure, the learning rate was 0.5, and the number of iterations was 296. The batch size was 48. The activation function was the Rectified Linear Unit (ReLU). The dropout was 0.1. These parameters were used for the training model. It used for the classification of the eight basic emotions through audio signals. The results are displayed in Fig. 12. The confusion matrix showed the overall performance of the system. There was accuracy of 70.83%, precision of 71.31%, recall of 70.83%, and f-measure of 70.70%. The eight basic emotions were classified correctly for 102 samples. The number of wrong classifications was 42 from 144 samples. The angry emotion has the highest accuracy in each class. On the other hand, the sad emotion has lower accuracy than other emotions. The sad emotion had the wrong classification as neutral emotion, calm, fearful, and disgusted. Therefore, the bimodal emotion recognition allowed improving the performance to be better than the unimodal emotion recognition.

6.3 Bimodal Emotion Recognition

The emotion classification using bimodal data by a deep belief network method is discussed next. The structure of DBN has a size of 2244-2600-2200-1100-8. For the pre-training process, the learning rate was 0.09, and the number of iterations was 11. For the fine-tuning process, the learning rate was 0.7, and the number of iterations was 260. The batch size was 48. The activation function was the Rectified Linear Unit (ReLU). The dropout was 0.2. These parameters were used for the training model. It used for the classification of eight basic emotions using bimodal data. The results are displayed in Fig. 13. The overall performance of the system has an accuracy value at 97.92%, precision of 98.01%, recall of 97.92%, and f-measure of 97.90%. The samples were used for classification of emotions. There were 144 samples. They were classified correctly for 141 of the samples. The diagonal component exposed that all emotions can be recognized correctly more than 97.00% of the time in each class. The recognition of neutral emotions, calm, sad, and surprised can be improved to be better. The combination of facial and speech features are highly effective because of additional features from the feature-level fusion of features (face and speech) whose relationship can help improve the emotion recognition performance, including the DBN method and configuration. The better system performance is displayed in Table 5 and Fig. 14.

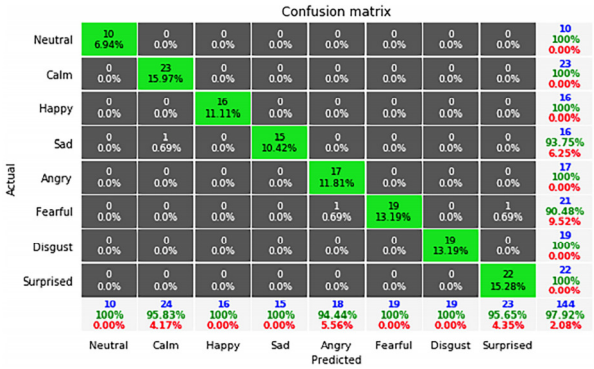
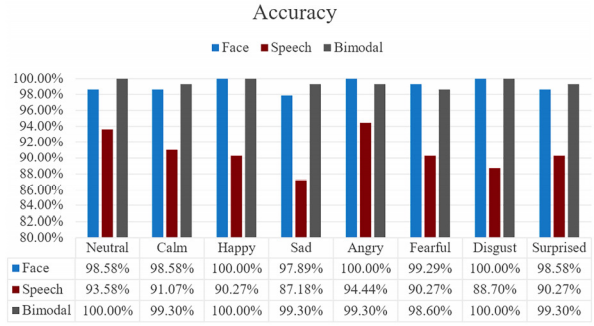


Fig.13:: Confusion Matrix of Bimodal Emotion Recognition

Table 5:: Summary of Result

Method	Accuracy rate (%)
Face Emotion Recognition	96.53
Speech Emotion Recognition	70.83
Bimodal Emotion Recognition	97.92



	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprised
Face	98.58%	98.58%	100.00%	97.89%	100.00%	99.29%	100.00%	98.58%
Speech	93.58%	91.07%	90.27%	87.18%	94.44%	90.27%	88.70%	90.27%
Bimodal	100.00%	99.30%	100.00%	99.30%	99.30%	98.60%	100.00%	99.30%

Fig.14:: Efficiency comparison between unimodal and bimodal emotion recognition

Fig. 14 shows that the sad emotion recognition has lower accuracy than other emotions (neutral, calm, happy, angry, fearful, disgust, and surprised) when using either facial or speech expressions. Therefore, bimodal emotion recognition which used the integration between facial and speech features in the feature-level was tried. It can improve the accuracy of recognition of sad emotions. The accuracy of detecting emotions (neutral, calm, sad, and surprised) improved compared with the unimodal emotion recognition (only using face or speech). The other emotions (happy and disgusted) have the same accuracy in comparison with the unimodal emotion recognition (face), but angry and fearful emotions have a little reduction in accuracy.

Furthermore, previous work proposed the bimodal emotion recognition system based on facial and speech expression. We found the classification of eight basic emotions was investigated in two papers in the year 2018. Each work used different methods and got different results on the same database, as shown in Fig. 15.

Refs.	Year	Propose	Bimodal	Method	Database	Emotion	Accuracy
[11]	2004	Emotion recognition	Face & Speech	SVM	Own database	4	89.01%
[12]	2008	Emotion recognition	Face & Speech	FLDA	Own database	6	82.14%
[8]	2016	Emotion recognition	Face & Speech	SR	eNTERFACE'05	6	87.46%
[13]	2017	Emotion recognition	Face & Speech	C3D & DBN	eNTERFACE'05	6	89.39%
[9]	2018	Emotion recognition	Face & Speech	Traditional & Deep learning	CHEAVD 2.0	8	41.89%
[10]	2018	Emotion recognition	Face & Speech	SVM	CHEAVD 2.0	8	38.41%
[14]	2019	Emotion recognition	Face & Speech	SVM	Indian Face database & Berlin Speech database	6	78%
Our	2020	Emotion recognition	Face & Speech	DBN	RAVDESS	8	97.92%

Fig.15:: Bimodal Emotion Recognition System

With our bimodal emotion recognition system based on facial and speech expression by deep belief network method for the classification of eight basic emotions on the RAVDESS database, the overall accuracy rate is 97.92%.

7. CONCLUSION

In this paper, we presented an emotion recognition system based on facial and speech expression, which used a deep belief network approach for the classification of eight basic emotions from video and audio information on the RAVDESS database. The combination of facial and speech features at feature level fusion was more effective than using single modality data. The experimental results showed that the emotion recognition system through facial expression had an accuracy rate of 96.53%. The speech emotion recognition system had an accuracy rate of 70.83%. Our bimodal emotion recognition system has an accuracy of 97.92%. From this research result we can conclude that the bimodal emotion recognition system has better performance than unimodal emotion recognition systems using only facial or speech expression.

References

[1] K. R. Scherer et al., "Psychological models of emotion," *The neuropsychology of emotion*, vol. 137, no. 3, pp. 137–162, 2000.

[2] Q. Li, Z. Yang, S. Liu, Z. Dai, and Y. Liu, "The study of emotion recognition from physiological signals," in *2015 Seventh International Conference on Advanced Computational Intelligence (ICACI)*. IEEE, 2015, pp. 378–382.

[3] R. W. Picard, *Affective computing*. MIT press, 2000.

[4] A. A. Varghese, J. P. Cherian, and J. J. Kizhakkethottam, "Overview on emotion recognition system," in *2015 International Conference on Soft-Computing and Networks Security (ICSNS)*. IEEE, 2015, pp. 1–5.

[5] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from au-

dio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[6] M. Mukeshimana, X. Ban, N. Karani, and R. Liu, "Multimodal emotion recognition for human-computer interaction: A survey," *System*, vol. 9, p. 10.

[7] S. Thushara and S. Veni, "A multimodal emotion recognition system from video," in *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*. IEEE, 2016, pp. 1–5.

[8] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, "Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech," *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1319–1329, 2016.

[9] H. Miao, Y. Zhang, W. Li, H. Zhang, D. Wang, and S. Feng, "Chinese Multimodal Emotion Recognition in Deep and Traditional Machine Leaming Approaches," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 2018, pp. 1–6.

[10] F. Xu and Z. Wang, "Emotion Recognition Research Based on Integration of Facial Expression and Voice," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISPBMEI)*. IEEE, 2018, pp. 1–6.

[11] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[12] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[13] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatio-temporal features for multimodal emotion recognition," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 1215–1223.

[14] K. P. Rao, M. V. P. C. S. Rao, and N. H. Chowdary, "An Integrated Approach to Emotion Recognition and Gender Classification," *Journal of Visual Communication and Image Representation*, 2019.

[15] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[16] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research* , vol. 10, no. Jul, pp. 1755–1758, 2009.

[17] T. Giannakopoulos, "pyaudioanalysis: An open-source python library for audio signal analysis," *PloS one*, vol. 10, no. 12, p. e0144610, 2015.

[18] S. Poria, E. Cambria, A. Hussain, and G.-B. Huang, "Towards an intelligent framework for multimodal affective data analysis," *Neural Networks*, vol. 63, pp. 104–116, 2015.

[19] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[20] A. Fischer and C. Igel, "Training restricted Boltzmann machines: An introduction," *Pattern Recognition*, vol. 47, no. 1, pp. 25–39, 2014.

[21] albertbup, "A python implementation of deep belief networks built upon numpy and tensorflow with scikit-learn compatibility," 2017. [Online]. Available: `https://github.com/albertbup/deep-belief-network`

[22] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

Apichart Jaratrotkamjorn was born in Songkhla, Thailand, in 1987. He received his B.Eng. degree in Electrical Engineering, Major Computer Engineering from Mahanakorn University of Technology (MUT), Thailand in 2010. He is currently a master student at the Department of Computer Engineering, Prince of Songkla University (PSU). His research interests are artificial intelligence, machine learning, deep learning, human-computer interaction, and emotion recognition.

Anant Choksuriwong received the Bachelor, Diploma, Master and Ph.D. degrees in 2000 (PSU), 2003 (UJF), 2004 (INPG) and 2008 from the School of Engineering in ENSI de Bourges. Currently he is a researcher at iSys Laboratory of Computer Engineering PSU, Songkhla, Thailand. He is also a lecturer at the Department of Computer Engineering, Prince of Songkla University (PSU), teaching courses in Advanced Image Processing, Machine Learning, and Principles of Robotics.