

Emotion recognition from Burmese speech based on fused features and deep learning method

Lwin Lwin Mar¹, Win Pa Pa²

¹Faculty of Computer Science, Computer University, Pakokku, Myanmar

²Natural Language Processing Lab, University of Computer Studies, Yangon, Myanmar

Article Info

Article history:

Received Nov 7, 2023

Revised Mar 8, 2024

Accepted Apr 6, 2024

Keywords:

DenseNet-emotion

Discrete wavelet transform

Local binary pattern

Mel-frequency cepstral coefficient

Text-tone feature

ABSTRACT

Burmese language is challenging for speech emotion classification. Moreover, it is lack of resource and few research was made in this topic. To solve the challenging problem, novel feature extraction for Burmese language is proposed. For lack of resource, Burmese speech emotion corpus called BMISEC is built. To support the challenging problem, the advantages of feature extractions are fused to create a robust feature. Four features are fused. Novel text-tone feature, local binary pattern, mel-frequency cepstral coefficient and discrete wavelet transform are fused. To progress the performance, deep learning method called DenseNet-Emotion is used for classification. Support vector machine is used in DenseNet's classifier layer. To show the robustness of the proposed system, three types of experiments are made on Tensorflow framework. They are ablation study, experiments with three publicly available datasets and experiments with the previous research methods and they are compared with the proposed method. It is found that feature fusion is superior to only one feature in emotion recognition. BMISEC gets better performance than other datasets. Moreover, the proposed method gets the superior result than previous research methods. The proposed method gets the accuracy of 88.388% for 50 epochs.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Lwin Lwin Mar

Faculty of Computer Science, Computer University

Minn Dat Pakokku Road, Kan Hla, Pakokku Township, Pakokku, Myanmar

Email: lwinlwinmar@ucsy.edu.mm

1. INTRODUCTION

Speech is a key in communication of humans in daily life. Emotion recognizing is a difficult task for machines. The purpose of emotional speech recognition is to use a person's voice to automatically assess their emotional or physical state [1]. Human can recognize the emotion of others by listening speech, moreover facial expression and body gesture can also support in recognizing emotions. Emotion recognition from speech using machine is a challenging task. It is a method of detecting the emotional state of a speaker by using the speech signal [2]. Speech emotion recognition (SER) is a very active area of research that involves the application of current machine learning and neural network tools [3]. Speech contains important information such as intentions and emotions. Speech emotion recognition can be applied in great applications such as marketing, call center, health care, autonomous driver emotion detection and entertainment, moreover SER is important to make robots have humanoid intelligence.

Alghifari *et al.* [4] used deep neural networks (DNNs) to recognize emotion in speech. The authors used mel frequency cepstral coefficient (MFCC) to extract speech features from raw audio data. They experimented the optimal configuration with 13 MFCC and 25 MFCC. They compared 3 emotions and 4 emotions. The recognition rate of 4 emotions with 25 MFCC got the better result than 13 MFCC for 3

emotions. They limited emotions recognized for three and four types. Moreover, their results are poor because of language difference. Zielonka *et al.* [5] applied convolutional neural network and spectrogram and mel-spectrogram as input. And then they investigated which feature extraction method is better represents emotions. As the studies, mel-spectrograms are better suited datatype. The drawback is the recognition accuracy is low. The accuracy of only four emotion recognition was 55.89% and for six emotions, it got 57.42% accuracy. Speech emotion recognition was implemented using a self-attention based deep learning model that was created by combining a two-dimensional convolutional neural network (CNN) and a long short term memory (LSTM) network in [6]. They used only mel-frequency cepstral coefficient for feature extraction. They used only open-source database for materials. They used only 480 male recordings of RAVDESS from the total 1,440 recordings [6]. Cross-corpus speech emotion recognition used different corpus for training and testing data in [7]. They applied subspace learning and information entropy-based domain adaption to obtain common subspace. It got the average unweighted accuracy of 54.93%. It limited transferred knowledge from source corpus if there were complex acoustic conditions [7]. Wen *et al.* [8] also used different corpus for source and target. They used margin disparity discrepancy for corpus adaption. However, the separation between classes in the target corpus was not clear enough, therefore accuracy was reduced [8].

Feature is most important in SER process. Considerable amount of literature has been published on improved features for emotion in speech. Sun *et al.* [9] used empirical mode decomposition on speech signals to extract features. From empirical mode decomposition, 43-dimensional time-frequency features were extracted and classified with convolutional recurrent neural network. But noise residuals and mode mixing problem had bad effects on frequency distribution [9]. Peng *et al.* [10] used modulation-filtered cochleagram (MCG) as feature and multilevel attention network in classification. Despite the fact that MCG feature was effective, it had two limitations. They were high computational complexity and inefficient feature extraction [10]. Certain amount of literature has also been published on low resource language that are not available publicly. Messaoudi *et al.* [11] built Tunisian speech emotion dataset and different models: LSTM, VGGish and wav2vec2, were used for speech emotion recognition task. As features, mel-spectrogram, wav2vec2 and MFCC were used. They needed to augment the dataset size and apply the enhanced technique in order to improve the satisfactory results [11]. Nguyen *et al.* [12], both video facial expression and speech were used as input. Nguyen *et al.* [12] extracted face regions by using Viola Jones-based algorithm and three channels of log mel-spectrograms from audio stream. The extracted features were fed into PathNet for emotion recognition. This architecture was likely to be stacked up a large number of parameters [12].

Speech emotion recognition with global aware fusion module on multi-scale neural network was implemented in [13]. They used MFCC as feature extraction. They obtained the result for only four emotions [13]. Multi-domain model that trained multiple languages simultaneously was implemented in [14]. The author used five features and domain-specific gating network was utilized to assign weights for five features. The used five features are MFCC, Allosaurus, wav2vec, ge2e and byol. But multi-domain model had negative transfer problem [14]. Zou *et al.* [15] implemented speech emotion recognition using multi-level features with co-attention mechanism. They used MFCC, spectrogram and raw audio signal that were extracted by using different encoders and the obtained features were fused with co-attention method. But the obtained results were only with four emotions [15]. Tu *et al.* [16] used the method combining data augmentation, feature selection, feature fusion. They used spectral vector from Log-Mel spectrum and IS09-emotion, IS10-paraling, IS13-ComParE, emobase feature extracted from OpenSMILE. The features were fused and classified emotions with multi-head attention mechanism-based convolutional recurrent neural network. But they lacked in the optimal set and feature fusion of emotional acoustic features. Moreover, they needed to improve the speed of system recognition [16].

Some literatures were published on deep learning methods for speech emotion recognition. Speech emotion recognition using attention-based dense LSTM was implemented in [17]. They used frame-level speech features that were based on ComParE openSMILE features and attention-based dense long short-term memory (LSTM) was used for classification. They added weight coefficients to skip-connection of each layer. The author added weights to the time dimension and feature dimension of the LSTM output to distinguish emotional information difference. They were less meaningful for continuous emotion recognition [17]. Goncalves and Busso [18] improved speech emotion recognition by using self-supervised learning. The author used 65 frame-based features including fundamental frequency, energy, mel-frequency cepstral coefficients. They used the complementary relationship between acoustic and facial features. But they needed to include more modalities and more domain specific pre-text tasks in the architecture [18]. Saleem *et al.* [19] used fusion of spectral and temporal features of input speech. They classified emotions with attention-based GRU module. But the improvements were needed on the model and contextual information needed to enhance and parameter size was needed to reduce [19]. Liao and Shen [20] used log spectrogram, log mel spectrogram and 3-dimension spectrogram that combines static and dynamic features as features and swin

transformer network for classification. They found that mel spectrogram was the best for the network. But it can be applied on large-scale data, otherwise it made a large number of network parameters a burden [20].

Only one feature cannot achieve high accuracy and results. When they are fused, their good characteristics are fused and the method is the best. MFCC is effective in distinguishing different emotions. Discrete wavelet transform has good feature for recognizing speech emotion. In the proposed system of Burmese speech emotion recognition, their effective characteristics are fused, the best feature is obtained. Deep learning also improves recognition. In the past twenty years ago, there was a research on speech emotion recognition for Burmese language and Mandarin language. It used log frequency power coefficients (LFPC) as feature extraction and hidden Markov model was used for classification. Burmese language is a tonal language and it has many challenges for emotion recognition. Moreover, the researches [4]-[20] except the author [10] worked speech emotion recognition on existing popular resources. From the overview, it can be deduced that there is little published data on newly created speech resources. Popular speech emotion datasets are available, but the under-resourced language are not available because of the lack of resources. Burmese speech emotion dataset is non-existence resource. In section 2, a novel feature extraction for the challenges of Burmese speech emotion recognition is presented. In section 3, all methodologies and dataset material used in the system are presented with their detailed procedures. In section 4, results of the proposed system are presented and discussion about them is explained. Section 5 presents conclusion of the paper.

2. TEXT-TONE FEATURE EXTRACTION

Burmese is a tonal language and it is challenging for speech emotion recognition. Moreover, it is under-resourced language and resource dataset is not available publicly. For many years, speech emotion recognition for Burmese language did not worked. Feature extraction is very important part of SER system. It is the heart of speech emotion recognition system [21]. For challenging task of speech emotion recognition system in Burmese language, new feature is introduced and it is coming from emotion text and based on four important tones of Burmese. In feature fusion, to add more emotion information, new feature extraction for Burmese language is proposed.

Tonal language means phonemic contrasts can be made on the basis of the tone of the vowel [22]. There are four important tones in Burmese. They are low tone, high tone, creaky tone and stopped tone (checked tone). In tonal language, pitch is an important feature [23]. Not only pitch, but also phonation, loudness, duration and vowel quality are also important. Therefore, four tones can be distinguished by using these features. Figure 1 demonstrates the steps of Text-tone feature extraction. To get text-tone features, audio speech signals are transformed into emotion sentences and sentences are segmented into tone words. Tone word speech signals are obtained again from words. Pitch, loudness and duration are extracted from tone word speech.

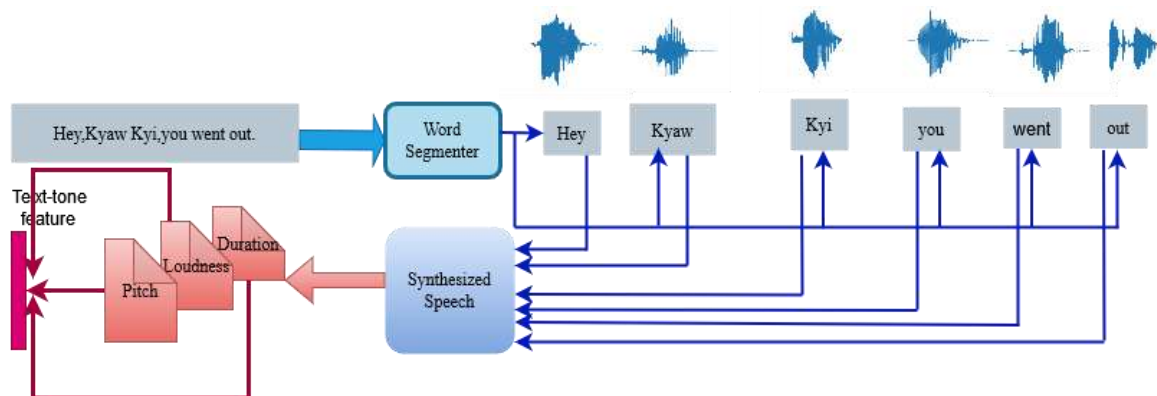


Figure 1. Text-tone feature extraction

Emotion sentences are distinguished into four types:

- You went out (Burmese) is a type of high tone.
- My sister has come (Burmese) is a type of creaky tone.
- It's bad (Burmese) is a type of low tone.
- They obeyed. (Burmese) is a type of low tone.

3. RESEARCH METHOD AND MATERIAL

Burmese speech emotion recognition is implemented based on the four feature extraction methods and deep learning architecture DenseNet-Emotion. Only one feature gets less effective result in recognition. Advantages of many features are more effective for the result. Deep learning is widespread used for speech emotion recognition in the recent researches. Figure 2 illustrates the system architecture for Burmese speech emotion recognition. Seven emotion classes are classified in the system. Features are extracted from Burmese text, spectrogram, and speech signal.

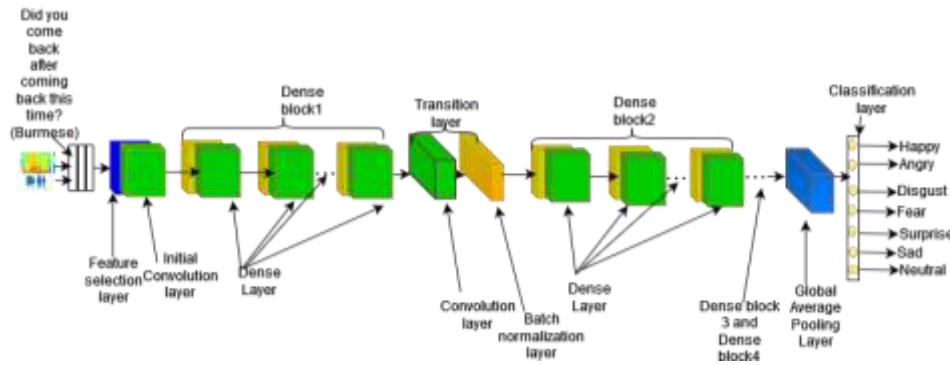


Figure 2. System architecture for Burmese speech emotion recognition

Only one feature cannot give effective emotion information. Speech emotion recognitions are limited because of small scale of datasets and lack of feature representation. Accurate recognition of emotion from speech is challenge. By combining the effectiveness of each feature extraction method, emotional information is more effective. Four feature extractions are fused to complement each other. To make progress the emotion recognition performance, deep learning method was chosen. To use features for the system, speech signals were extracted by using text-tone, local binary pattern (LBP), MFCC and discrete wavelet transform (DWT).

3.1. Local binary pattern feature extraction on burmese speech emotion spectrograms

Local binary pattern is used for facial and image classification. Ojala introduced it in 1996 [24]. LBP operator eliminates the need for contour tracing, denosing, and other prior processing [25]. LBP is an image operator which transforms an image into an array or image of integer labels describing small-scale appearance of the image. Two dimensional narrow band speech spectrogram is a graphical display of the square magnitude of the time varying characteristics of speech [26]. Spectrograms can carry emotion information.

Burmese speech spectrograms have many intensities. LBP can support emotional features for SER. Speech spectrograms are extracted from speech signals by using python script that creates spectrogram matrices. There are seven types of emotion spectrograms. From skimage package, local-binary-pattern method is used to extract emotional features. Flatten data features are extracted for the system. The features are inputted into panda dataframe.

3.2. MFCC features extracted from speech corpus

MFCC is a feature extraction that have high performance and recognition. It is the feature widely used in automatic speech and speaker recognition [27]. However it cannot give accurate result if there are background noises in the speech signal. Different dimension of features was used in MFCC. The 39-dimension is an enough feature representation for emotion classification. Librosa version 0.7.2 package of python gives the functions to extract features from speech signals. From Librosa and its feature, 39-dimensional MFCC features were extracted. From 39-dimension, 13th parameter is energy coefficient. Energy calculation is demonstrated in (1). Delta values were also calculated to measure the changes from the previous frame of input speech signal to the next frame. First order derivative of features is in (2).

$$\text{Energy} = \sum_{t=t_1}^{t_2} x^2[t] \tag{1}$$

$$d(t) = \frac{c(t+1)-c(t-1)}{2} \tag{2}$$

3.3. Discrete wavelet transform feature extraction on speech signals

The DWT feature extraction is a powerful feature extraction tool for emotion information. Its benefits are data compression, approximation's coefficient's retainment of features and characteristics of original data and data's locally changes in detail coefficient [28]. The system used pywt version 1.1.1 and from it, wavedec was used to extract DWT emotion features. Used level in DWT was 3. The obtained DWT emotion features were stored in panda dataframe. The emotion information obtained from four feature extraction methods were stored in dataframe and the dataframe was inputted into deep learning method.

3.4. Feature fusion with panda dataframe

Panda gives powerful and expressive data structure and it is easy for data manipulation and analysis [29]. Features obtained from four extraction methods are in diverse forms. Some features can contain missing values. Pandas can support for this problem well. It provides streamlined representation of data. Pandas dataframe provides in tabular form. Combining dataframes is easy and provide robust features for classification. Pandas dataframes can be used with little coding and it can work much more. The final feature is demonstrated in (3).

$$y = x_t \oplus x_l \oplus x_m \oplus x_d \quad (3)$$

In (3), y is final emotion feature, x_t is Text-tone features, x_l is local binary pattern feature, x_m is MFCC feature and x_d is discrete wavelet transform features. Because deep learning method trains large amount of data, the training data of the system is in huge amount. Pandas can support for huge amount of data very well with limited memory.

3.5. Deep learning architecture for classification

Deep learning progresses the speech emotion recognition. Various deep learning methods were used in previous researches. Including neural network, CNN, LSTM were used in these works. Traditional deep learning methods have degradation problem. DenseNet is a network that connects each layer to every other layer in a feedforward fashion [30]. DenseNet alleviates vanishing-gradient problem. Features inputted from feature fusion of pandas dataframe were fed into DenseNet-Emotion. Feature selection layer was added into DenseNet-Emotion. As the used data are large, the dimension of feature vector is increasing. The best features were selected for the system. Simple method SelectKBest feature selection was used in the system.

DenseNet-Emotion has four dense blocks and the dense blocks have 6, 12, 24, 16 dense layers each. As classifier layer, SVM was used. Other types such as softmax and softplus were also experimented. As the experiments, SVM layer got the best performance. The SVM equation for seven classes is:

$$a_7(x) = w^T x \quad (4)$$

3.6. Emotion dataset (BMISEC) [31]

Burmese emotion dataset is a lack of resource and it is under-resource language. And Burmese emotion recognition is a challenging task, therefore it is an important work to build the robust emotion dataset. Moreover, speech emotion corpus is a foundation of the speech emotion recognition. Burmese emotion dataset called Burmese movies interviews speech emotion corpus (BMISEC) was collected from Myanmar movies and interviews. Myanmar movies are mainly collected from mahar channel. It can be watched in website, www.youtube.com/MaharSeries. Interviews are collected from YouTube also. As it was collected from movies, there are many background musics and noises. They were preprocessed to remove. It is a reason because feature extractions can work accurately if there is no background noise.

BMISEC contains seven emotion types. They are happy, angry, disgust, surprise, sad, fear and neutral. There were five actors and three actresses in BMISEC. The utterances were extracted by using Praat tool with 16,000 Hz and mono stereo type. Emotion sentences are 3 seconds in length. Total length of BMISEC is 37,602 seconds or 10 hours, 26 minutes and 43 seconds. Train and test data separation of experimental dataset is shown as total utterances and total length in Table 1.

Table 1. Experimental data separation

Train	Test
9145 utterances	6097 utterances
06 hr:20 min:14 sec	04 hr:06 min:27 sec

In collecting BMISEC, it was found that some emotion types can confuse. Happy emotion utterance can be two types: loud voice and soft voice. Moreover, some happy speech appears with sadness. Similarly, angry voice can be loud and soft. In fear type, beside afraid voice, anxious voice was also added. Class distribution of BMISEC is demonstrated in Figure 3. In class distribution, surprise emotion is more speech utterances. In this figure, there are three amounts of percentage. The first is 13.7% of total dataset and happy, neutral, sad, disgust and fear have this percentage of total dataset, the second is 13.4% of total and angry emotion has this percentage. The third is 18.2% of total dataset and surprise emotion has this percentage.

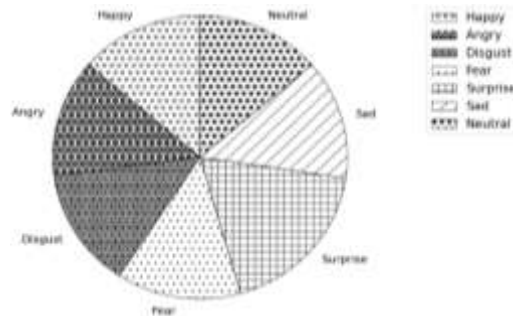


Figure 3. Class distribution of BMISEC

4. RESULTS AND DISCUSSION

Because of the challenging in speech emotion recognition for Burmese language, the advantages of four feature extractions are combined to obtain the expecting results. MFCC has advantage of high recognition accuracy. The features complement new feature extraction for Burmese language. Feature fusion gets improvement in recognition accuracy and recognition speed. Especially emotion recognition for Burmese language improves.

To proof the effectiveness of fusion and deep learning method, the experiments were done in different ways. To see the effectiveness of feature fusion, the experiments were made using only a new feature extraction, LBP, MFCC and DWT each and deep learning method. The four experimental results are compared with the proposed method. Moreover, the proposed system was experimented on other three different publicly available datasets and the proposed method was compared with previous methods.

4.1. Experimental setup

To show the effective results of the proposed system, experiments were made with Python language. These experiments were made in Anaconda 1.7.2 and Tensorflow framework 2.2.0. NVidia GeForce MX 250 was used. The same configuration was used for all experiments. For deep learning model, batch size was 32 and learning rate of 0.001 was used.

For new feature extraction method, word breaker is used to break the tone word from emotion sentences. The words are transformed into speech signals by using text-to-speech converter. For LBP, skimage package is used. For MFCC, Librosa toolbox is used. 39 dimensional MFCC feature vectors are extracted. For DWT, pywt package is used and from pywt, wavedec is used to extract discrete wavelet transform features. These features are combined with panda dataframe. Panda library supports high performance for large dataset even with limited memory resources. It can handle missed data very well. By fusing with panel data, the features can be faster processed. It is simplest, it takes many dataframes and appends them into long a particular axis.

4.2. Experiment results

Three types of experiments were made. The first type is ablation study. Burmese speech emotion recognition was experimented by using only one of feature extraction methods and deep learning method. Their results are compared with proposed method (4 Fusion) to show its effectiveness. The results were obtained from 50 epochs. Table 2 shows the comparison of baseline methods and the proposed system. As shown in Table 2, the recognition performance is improved obviously on feature fusion. To evaluate the influence of more than one feature, ablation study is performed. As in table, feature fusion method improves the result to about 20% to 30% over the other methods. Deep learning method can also support emotion recognition to increase performance. The confusion matrix result of the proposed system (4 Fusion+DenseNet-Emotion) is presented in Figure 4.

Table 2. Ablation study of the proposed system

Method	Accuracy (%)
Text-tone+DenseNet-Emotion (Baseline)	54.04
LBP+DenseNet-Emotion (Baseline)	51.89
DWT+DenseNet-Emotion (Baseline)	69.41
MFCC+DenseNet-Emotion (Baseline)	74.32
4 Fusion+DenseNet-Emotion	88.388



Figure 4. Confusion matrix of the proposed system (Text-tone+LBP+MFCC+DWT+DenseNet-Emotion)

The second type of experiment is that the proposed system is experimented with three publicly available datasets. They are berlin, RAVDESS and Persian datasets. LBP, MFCC and DWT are feature extractions that are made on these datasets. Advantages of these feature extractions are combined to be more effective. New feature extraction is innovated for Burmese language, therefore it is not used in the second experiment. LBP is appropriate for spectrograms, so it can be used for various datasets. MFCC is high performance rate feature extraction if there is no background noises. Therefore 3 fusion can work for speech emotion recognition of these datasets. Table 3 shows the comparison of BMISEC and other datasets. BMISEC was extracted by only three feature extractions to be equal with other datasets. BMISEC gets the highest accuracy among the datasets.

Experimental results of the second type are demonstrated with deep learning curve in Figure 5. Figure 5(a) is comparison of accuracy and Figure 5(b) is of loss (error). In the experiments, berlin dataset did not plot the curve very well and it is considerable for the reason of only 535 data utterances.

Table 3. Comparison of experimental results on four datasets

Dataset	Method	Accuracy (%)
BMISEC(Burmese)	3 Fusion+DenseNet-Emotion	79.9
RAVDESS [32]	3 Fusion+DenseNet-Emotion	46.59
Berlin [33]	3 Fusion+DenseNet-Emotion	61.33
Persian [34]	3 Fusion+DenseNet-Emotion	54.15

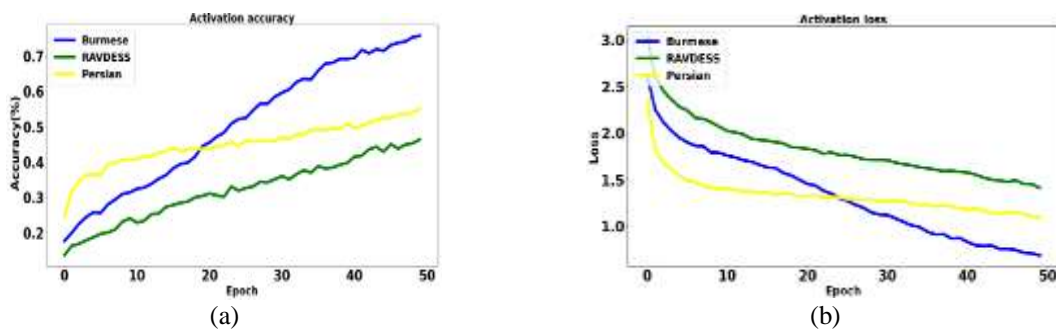


Figure 5. Comparison of BMISEC and other datasets in (a) accuracy and (b) loss (error)

The third type of experiment is to compare the proposed system with the previous research results. In past twenty years ago, speech emotion recognition in Burmese language was implemented with Log frequency power coefficients (LFPC) and hidden Markov model. During the years after this research, this topic has not been worked. To compare with the proposed system, BMISEC was extracted using lfpc and classified with HMM. Table 4 shows the results of the proposed system and the previous works and the effectiveness of the proposed method is presented. The third type uses three previous methods to compare with the proposed method. The first method is of Nwe *et al.* [35]. The authors used 720 utterances of Burmese and Mandarin languages. The experiment used BMISEC and obtained 78.18%. The second method is of Zielonka *et al.* [5] they used combination of four datasets. The method got the result of 55.3% for seven emotions. The last method is with MFCC and deep feedforward neural network. Alghifari *et al.* [4] implemented the recognition of speech with the method. It got 75.49% accuracy with seven emotions.

Table 4. Comparison of proposed method with previous research

Reference	Method	Accuracy (%)
[35]	LFPC+HMM	78.18
[5]	Mel-spectrogram+CNN	55.3
[4]	MFCC+DNN	75.49
Our work	4 Fusion+DenseNet+Emotion	88.388

The proposed system investigated the effects of features to increase the recognition performance. Nwe *et al.* [35] used lfpc as feature extraction and ergodic hidden Markov model was used as classifier. They found that the performance of systems employing traditional features such as lpcc degraded when more than two classes of emotion are classified. Zielonka *et al.* [5] intended to maximize recognition rate and Alghifari *et al.* intended to increase CNN performance. They explored the impacts of mel-spectrogram, MFCC and lfcc but they have not addressed the influence of fusion of the features on the speech emotion recognition system. We found that feature fusion of popular features gives highest accuracy and accelerates the processing speed. Our finding also demonstrates that innovated feature extraction is effective for Burmese speech emotion recognition. Our study suggests that DenseNet gets the higher result which is not high in CNN. The proposed system may benefit from not only feature fusion but also DenseNet and speech emotion recognition corpus (BMISEC). It is also found that BMISEC is a speech emotion corpus that does not contain any background noise and it can support for speech emotion recognition. Overall execution time and performance are better than the previous works.

This study explored the novel feature extraction for Burmese language SER and the effective fusion of feature extractions. When speech is inputted, speech to text converter is needed. Therefore, more effective converter is intended to use. Our study explored the innovative feature extraction and further studies may explore more effective feature extraction. More effective feature extraction will be researched. The study provides conclusive evidence that feature fusion can give more robust feature for speech emotion recognition. Fusion of four feature extractions gets increasing the Burmese speech emotion recognition performance.

5. CONCLUSION

To enhance the effectiveness of feature extraction, four feature extractions are fused. For the challenge of Burmese speech emotion recognition, innovated feature extraction for Burmese language is proposed. To complement this, local binary pattern and two popular feature extractions are used and their features are fused. Burmese speech emotion recognition is a neglected topic for researchers. Because Burmese language is lack of resource and speech emotion corpus is a foundation for the system, Burmese speech emotion corpus (BMISEC) is built. In fusing the emotion features, simple and powerful panel data is used. It can handle the missing data very well, moreover it can support high performance for the large dataset with limited memory. The robust features are inputted to deep learning model DenseNet-Emotion. For the robustness of the proposed system, three types of experiments are carried out. As the first experiment, effectiveness of feature fusion can be seen. In second experiment, BMISEC is compared with other three datasets. It is found that BMISEC got better result than other three publicly datasets. BMISEC is a best prepared speech emotion corpus. It is preprocessed to remove background noise and music, therefore MFCC and DWT can get the high result. As the last experiment, it can be found that the proposed method obtained the superior result than three previous research methods. The proposed method got the high result for seven emotions in only 50 epochs. It got the accuracy of 88.388% in 50 epochs. The proposed system can work overall execution and recognition much faster. In the future, more effective feature extraction for Burmese speech emotion recognition is intended to experiment.





REFERENCES

- [1] S. Padman and D. Magare, "Regional language speech emotion detection using deep neural network," *ITM Web of Conferences*, vol. 44, p. 03071, May 2022, doi: 10.1051/itmconf/20224403071.
- [2] A. Kumar and A. Kumar Goel, "Speech emotion recognition by using feature selection and extraction," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, May 2022, pp. 818–823. doi: 10.1109/ICAAIC53929.2022.9792796.
- [3] J. de Lope and M. Graña, "An ongoing review of speech emotion recognition," *Neurocomputing*, vol. 528, pp. 1–11, Apr. 2023, doi: 10.1016/j.neucom.2023.01.002.
- [4] M. F. Alghifari, T. S. Gunawan, and M. Kartiwi, "Speech emotion recognition using deep feedforward neural network," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 10, no. 2, p. 554, May 2018, doi: 10.11591/ijeecs.v10.i2.pp554-561.
- [5] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, "Recognition of emotions in speech using convolutional neural networks on different datasets," *Electronics*, vol. 11, no. 22, p. 3831, Nov. 2022, doi: 10.3390/electronics11223831.
- [6] J. Singh, L. B. Saheer, and O. Faust, "Speech emotion recognition using attention model," *International Journal of Environmental Research and Public Health*, vol. 20, no. 6, p. 5140, Mar. 2023, doi: 10.3390/ijerph20065140.
- [7] X. Cao, M. Jia, J. Ru, and T. Pai, "Cross-corpus speech emotion recognition using subspace learning and domain adaption," *Eurasip Journal on Audio, Speech, and Music Processing*, vol. 2022, no. 1, p. 32, Dec. 2022, doi: 10.1186/s13636-022-00264-5.
- [8] X.-C. Wen *et al.*, "CTL-MTNet: a novel capsnet and transfer learning-based mixed task net for single-corpus and cross-corpus speech emotion recognition," in *Proceedings of the Thirty-First International Conference on Artificial Intelligence*, California: International Joint Conferences on Artificial Intelligence Organization, Jul. 2022, pp. 2305–2311. doi: 10.24963/ijcai.2022/320.
- [9] C. Sun, H. Li, and L. Ma, "Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network," *Frontiers in Psychology*, vol. 13, Jan. 2023, doi: 10.3389/fpsyg.2022.1075624.
- [10] Z. Peng, W. He, Y. Li, Y. Du, and J. Dang, "Multi-level attention-based categorical emotion recognition using modulation-filtered cochleagram," *Applied Sciences*, vol. 13, no. 11, p. 6749, Jun. 2023, doi: 10.3390/app13116749.
- [11] A. Messaoudi, H. Haddad, M. B. Hmida, and M. Graiet, "TuniSER: toward a Tunisian speech emotion recognition system," in *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, 2022. [Online]. Available: <https://aclanthology.org/2022.icnlsp-1.27>
- [12] D. Nguyen *et al.*, "Meta-transfer learning for emotion recognition," *Neural Computing and Applications*, vol. 35, no. 14, pp. 10535–10549, May 2023, doi: 10.1007/s00521-023-08248-y.
- [13] W. Zhu and X. Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2022, pp. 6437–6441. doi: 10.1109/ICASSP43922.2022.9747517.
- [14] Z. Wang, Q. Meng, H. Lan, X. Zhang, K. Guo, and A. Gupta, "Multilingual speech emotion recognition with multi-gating mechanism and neural architecture search," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Jan. 2023, pp. 806–813. doi: 10.1109/SLT54892.2023.10022557.
- [15] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2022, pp. 7367–7371. doi: 10.1109/ICASSP43922.2022.9747095.
- [16] Z. Tu, B. Liu, W. Zhao, R. Yan, and Y. Zou, "A feature fusion model with data augmentation for speech emotion recognition," *Applied Sciences*, vol. 13, no. 7, p. 4124, Mar. 2023, doi: 10.3390/app13074124.
- [17] Y. Xie, R. Liang, Z. Liang, and L. Zhao, "Attention-based dense LSTM for speech emotion recognition," *IEICE Transactions on Information and Systems*, vol. E102.D, no. 7, pp. 1426–1429, Jul. 2019, doi: 10.1587/transinf.2019EDL8019.
- [18] L. Goncalves and C. Busso, "Improving speech emotion recognition using self-supervised learning with domain-specific audiovisual tasks," in *Interspeech 2022*, ISCA: ISCA, Sep. 2022, pp. 1168–1172. doi: 10.21437/Interspeech.2022-11012.
- [19] N. Saleem *et al.*, "DeepCNN: spectro-temporal feature representation for speech emotion recognition," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 2, pp. 401–417, Jun. 2023, doi: 10.1049/cit.12233.
- [20] Z. Liao and S. Shen, "Speech emotion recognition based on swin-transformer," *Journal of Physics: Conference Series*, vol. 2508, no. 1, p. 012056, May 2023, doi: 10.1088/1742-6596/2508/1/012056.
- [21] N. Shreya and G. Ms.Divya, "Speech feature extraction techniques: a review," *International Journal of Computer Science and Mobile Computing*, vol. 4, no. 3, pp. 107–114, 2015, [Online]. Available: <https://www.ijcsmc.com/docs/papers/March2015/V4I3201545.pdf>
- [22] M. Tian and A. Lee, "Burmese quotation intonation," in *Proceedings of the 19th International Congress of Phonetic Sciences*, Australia, 2019, pp. 2435–2439. [Online]. Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_2484.pdf.
- [23] J. Watkins, "Tone and intonation in burmese," in *International Phonetic Association*, 2003, pp. 1289–1292. [Online]. Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_1289.pdf
- [24] "Local binary patterns (LBP) & Histogram of oriented gradient (HoG)." 2016. [Online]. Available: <https://biomisa.org/uploads/2016/10/Lect-15.pdf>
- [25] M. Esfahanian, H. Zhuang, and N. Erdol, "Using local binary patterns as features for classification of dolphin calls," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. EL105–EL111, Jul. 2013, doi: 10.1121/1.4811162.
- [26] T. F. Quatieri, "Production and classification of speech sounds," in *Discrete-time speech signal processing*, Prentice-Hall, Inc, 2002.
- [27] P. Inkeaw, "Mel frequency cepstral coefficient MFCC." Practicalcryptography.com. [Online]. Available: <https://www2.cs.science.cmu.ac.th/courses/204371/files/MFCC.pdf>
- [28] S. Russell, I. S. Moskowitz, and B. Jalalian, "Context: separating the forest and the trees—Wavelet contextual conditioning for AI," in *Human-Machine Shared Contexts*, Elsevier, 2020, pp. 67–91. doi: 10.1016/B978-0-12-820543-3.00004-3.
- [29] K. Willems, "Pandas tutorial: DataFrames in Python," Radar al edition. [Online]. Available: <https://www.datacamp.com/tutorial/pandas-tutorial-dataframe-python> (accessed date: Dec,2022)
- [30] G. Huang, Z. Liu, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1978, pp. 1442–1446. doi: 10.48550/arXiv.1608.06993.





- [31] L. L. Mar, W. P. Pa, and T. L. Nwe, "BMISEC: corpus of burmese emotional speech," in *2023 IEEE Conference on Computer Applications (ICCA)*, IEEE, Feb. 2023, pp. 248–253. doi: 10.1109/ICCA51723.2023.10182163.
- [32] 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, May 2018, doi: 10.1371/journal.pone.0196391.
- [33] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech 2005*, ISCA: ISCA, Sep. 2005, pp. 1517–1520. doi: 10.21437/Interspeech.2005-446.
- [34] O. Mohamad Nezami, P. Jamshid Lou, and M. Karami, "ShEMO: a large-scale validated database for Persian speech emotion detection," *Language Resources and Evaluation*, vol. 53, no. 1, pp. 1–16, Mar. 2019, doi: 10.1007/s10579-018-9427-x.
- [35] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, Nov. 2003, doi: 10.1016/S0167-6393(03)00099-2.

BIOGRAPHIES OF AUTHORS



Lwin Lwin Mar     is a Ph. D candidate from University of Computer Studies, Yangon, Myanmar. She also received her B.C. Sc and M.C. Sc (Computer Science) from University of Computer Studies, Mandalay, Myanmar in 2005 and 2008, respectively. She is currently a Lecturer at Computer University, Pakokku, Myanmar. She is researching speech emotion classification for Burmese language. She can be contacted at email: lwinlwinmar@ucsy.edu.mm.



Win Pa Pa     got B.Sc. (Maths) degree from Mandalay University, M.I. Sc (Master of Information Science) and PhD (Information Technology) from University of Computer Studies, Yangon, Myanmar in 2001, 2004 and 2009. She is working as a professor and have been doing research at Natural Language Processing lab of UCSY since August, 2009. Her Ph. D thesis was on "Myanmar Word Segmentation" which is essential for Myanmar NLP and she is still doing research on Word Segmentation for accuracy. Her other research interests are Machine Translation and Speech synthesis. She can be contacted at email: winpapa@ucsy.edu.mm.