Deep Residual Local Feature Learning for Speech Emotion Recognition

Sattaya Singkul[1][0000−0001−7335−7105], Thakorn Chatchaisathaporn[2]
Boontawee Suntisrivaraporn[2], and Kuntpong Woraratpanya[1,⋆]

[1] Faculty of Information Technology,
King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
{59070173,kuntpong}@it.kmitl.ac.th
[2] Data Analytics, Siam Commercial Bank, Bangkok, Thailand
thakorn.chatchaisathaporn@scb.co.th, meng234@gmail.com

Abstract. Speech Emotion Recognition (SER) is becoming a key role in global business today to improve service efficiency, like call center services. Recent SERs were based on a deep learning approach. However, the efficiency of deep learning depends on the number of layers, i.e., the deeper layers, the higher efficiency. On the other hand, the deeper layers are causes of a vanishing gradient problem, a low learning rate, and high time-consuming. Therefore, this paper proposed a redesign of existing local feature learning block (LFLB). The new design is called a deep residual local feature learning block (DeepResLFLB). DeepResLFLB consists of three cascade blocks: LFLB, residual local feature learning block (ResLFLB), and multilayer perceptron (MLP). LFLB is built for learning local correlations along with extracting hierarchical correlations; DeepResLFLB can take advantage of repeatedly learning to explain more detail in deeper layers using residual learning for solving vanishing gradient and reducing overfitting; and MLP is adopted to find the relationship of learning and discover probability for predicted speech emotions and gender types. Based on two available published datasets: EMODB·and RAVDESS, the proposed DeepResLFLB can significantly improve performance when evaluated by standard metrics: accuracy, precision, recall, and F1-score.

Keywords: Speech Emotion Recognition · Residual Feature Learning · CNN Network · Log-Mel Spectrogram · Chromagram

1 Introduction

Emotional analysis has been an active research area for a few decades, especially in recognition domains of text and speech emotions. Even if text and speech emotions are closely relevant, both kinds of emotions have different challenges. One of the challenges in text emotion recognition is ambiguous words, resulting from omitted words [1,2]. On the other hand, one of the challenges in speech

[⋆] Corresponding author

emotion recognition is creating an efficient model. However, this paper focuses on only the recognition of speech emotions. In this area, two types of information, linguistic and paralinguistic, were mainly considered in speech emotion recognition. The linguistic information refers to the meaning or context of speech. The paralinguistic information implies the implicit message meaning, like the emotion in speech [3,4,5,6]. Speech characteristics can interpret the meaning of speech; therefore, behavioral expression was investigated in most of the speech emotion recognition works [7,8,9].

In recent works, local feature learning block (LFLB) [10], one of the efficient methods, has been used in integrating local and global speech emotion features, which provide better results in recognition. Inside LFLB, convolution neural network (CNN) was used for extracting local features, and then long short-term memory (LSTM) was applied for extracting contextual dependencies from those local features to learn in a time-related relationship. However, vanishing gradient problems may occur with CNN [11]. Therefore, residual deep learning was applied to the CNN by using skip-connection to reduce unnecessary learning and add feature details that may be lost in between layers.

Furthermore, the accuracy of speech recognition does not only rely on the efficiency of a model, but also of a speech feature selection [12]. In terms of speech characteristics, there are many distinctive acoustic features that usually used in recognizing the speech emotion, such as continuous features, qualitative features, and spectral features [13,12,14,15,16]. Many of them have been investigated to recognize speech emotions. Some researchers compared the pros and cons of each feature, but no one can identify which feature was the best one until now [3,4,17,18].

As previously mentioned, we proposed a method to improve the efficiency of LFLB [11] for deeper learning. The proposed method, deep residual local feature learning block (DeepResLFLB), was inspired by the concept of human brain learning; that is, 'repeated reading makes learning more effective,' as the same way that Sari [19] and Shanahan [20] were used. Responding to our inspired concept, we implemented a learning method for speech emotion recognition with three parts: Part 1 is for general learning, like human reading for the first time, Part 2 is for further learning, like additional readings, and the last part is for associating parts learned to decide types of emotions. Besides, the feature selection is compared with two types of distinctive features to find the most effective feature in our work: the normal and specific distinctive features are log-mel spectrogram (LMS), which is fully filtered sound elements, and MFCC deltas, delta-deltas, and chromagram (LMSDDC) are more clearly identify speech characteristics extracted based on the human mood.

Our main contributions of this paper are as follows: (i) Deep residual local feature learning block (DeepResLFLB) was proposed. DeepResLFLB was arranged its internal network as LFLB, batch normalization (BN), activation function, normalization-activation-CNN (NAC), and deep layers. (ii) Learning sequences of DeepResLFLB were imitated from human re-reads. (iii) Speech emotion features, based on human mood determination factors such as LMS and LMSDDC, were applied and compared their performances.

2 Literature Reviews

Model efficiency is one of the important factors in SER. Many papers focused on learning methods of machine learning or deep learning. Demircan [17] introduced fuzzy c-mean as a preprocessing step to group and add characteristics before using machine learning. Venkataramanan [21] studied of using deep learning in SER. The findings of the study revealed that CNN outperformed the traditional machine learning. Also, Huang [15] showed that semi-CNN in SER can increase accuracy. Zhao [10] presented the use of CNN in conjunction with LSTM to extract and learn features. Zhao's method used a sequence of CNN in a block style, consisting of CNN, BN, activation function, and pooling, for local feature learning, and then used LSTM for extracting contextual dependencies in a time-related relationship. In this way, both local and global features are extracted and learned.

It is undeniable that the effectiveness of deep learning mainly depends on the data size for training [22]. Recently, Google brain research [23] proposed data augmentation, one of the efficient techniques that can increase the amount of data, by adding spectrogram characteristics, also known as "Spectrogram Augmentation." This augmentation consists of time warping to see more time shift patterns, time masking to reduce the overfitting rate of the model and improve the sound tolerance that may have characteristics of silence, frequency masking to reduce the overfitting rate and increase sound resistance from concealing characteristics of a specific wavelength. The spectrogram is a basic feature of sound that can lead to various specific features. Therefore, by using above methods, the model can learn more perspectives of the data.

Also, different features lead to different performances in speech emotion recognition. Among the features of speech, mel-frequency cepstral coefficient (MFCC) [21], which can be characterized by the frequency filter in the range of 20 Hz to 20 kHz, similar to human hearing, is widely used to obtain coefficients from the filtered sound. Recent research papers [21,17] used the difference of MFCC to get more specific details, but, in the aspect of MFCC, it has no time relationship. Therefore, many papers [21,10,15] used mel spectrogram (MS) instead. MS can respond to the time relationship, thus providing better results than just using MFCC. Besides, music can be looked different from speech; therefore, chromagram is widely used instead of MFCC, since it can provide better features than normal MFCC and MS.

Our work is different from the previously mentioned works in that the deep residual local feature learning block (DeepResLFLB) was redesigned from LFLB. This method helps reduce the chance of feature and updated losses caused by CNN model in the LFLB, especially in deeper layers. DeepResLFLB uses a repeated learning style that local features extracted from a bias frame with silent voice (see subsection 3.3) can be learned through a residual deep learning approach. Moreover, we extracted distinctive features based on a concept of determining human emotions, consisting of prosodic [24], filter bank [24,21], and glottal flow [25]. These three features in conjunction with our ResLFLB can improve learning efficiency.

3 The Proposed Model

To enable SER as efficiently as possible, the following factors: raw datasets, environments, and features are included in our system design. Based on such factors, a new designed framework, called DeepResLFLB, was proposed as shown in Fig. 1. This framework consists of five parts: (i) raw data preparation, (ii) voice activity detection, (iii) bias frame cleaning, (iv) feature extraction, and (v) deep learning.

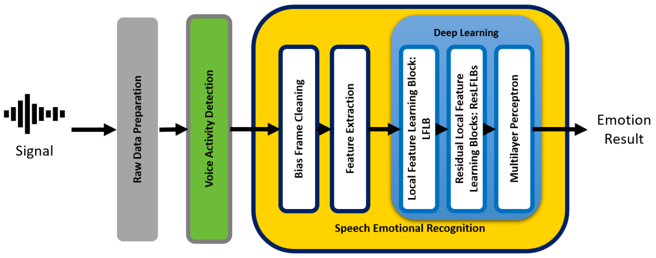

Fig. 1. A deep residual local feature learning framework.

3.1 Raw Data Preparation

Due to the complex nature of datasets and the difference of languages like EMODB in Berlin German and RAVDESS in English, the representation of the original datasets may not be enough for training a model based on deep learning. Therefore, increasing a variety of data to see more new dimensions or characteristics is essential. Responding to this, various data augmentation techniques, including noise adding, pitch tuning, and spectrogram, were used in this work to make the model more robust to noise and unseen voice patterns.

3.2 Voice Activity Detection

Although both datasets, EMODB and RAVDESS, were produced in a closed environment, quite a bit of noise, they were found that noise remains at the starting and stopping points of sound records. Indeed, noise is not related to speakers' voice, so it could be removed. Here, voice activity detection [26] was used to detect only voice locations, i.e., excluding noise locations. As a result of the voice activity detection, selected frames can be efficiently analyzed and classified male and female voices by energy-base features.

3.3 Bias Frame Cleaning

Bias frame cleaning is used as a postprocessing of voice activity detection; that is, each frame segmented by the voice activity detection is identified its loudness through Fourier transform (FT). If FT coefficients of a segmented frame are zero, that frame is identified as no significant information for emotional analysis, so it is rejected.

3.4 Feature Extraction

Model performance of deep learning mainly depends on features. The good features usually gain more model performance. Thus, this paper focuses on efficient extraction of human emotion features. Naturally, speech signals always contain human emotions. In other words, we can extract human emotions from speech signals. Here, we briefly describe three important components of speech signals: glottal flow, prosody, and human hearing. Glottal flow can be viewed as a source of speech signals [25]. It mainly produces fundamental frequencies [27] or latent sounds within the speech. Prosody is vocal frequencies, which are produced from the air pushed by the lung [25]. It contains important characteristics, such as intonation, tone, stress, and rhythm. On the other hand, for human hearing, MFCC is one of the analytical tools that can mimic the behavior of human ears by applying cepstral analysis [28]. Based on our assumption of extracting better emotion features, two important factors are included for feature extraction design: (i) the wide band frequencies of speech signals are regarded as much as possible to cover important features of speech emotions, and (ii) time-frequency processing is used for extracting speech emotions. Here, log-mel spectrogram (LMS) was used as time-frequency representation for emotion features. Two additional features extracted based on MFCC were delta and delta-delta. Furthermore, chromagram feature [29,30] was extracted as one of the emotion features. Fig. 2 shows our emotion feature extraction. As a result, four features, LMS, delta, delta-delta, and chromagram were used as emotional representation.

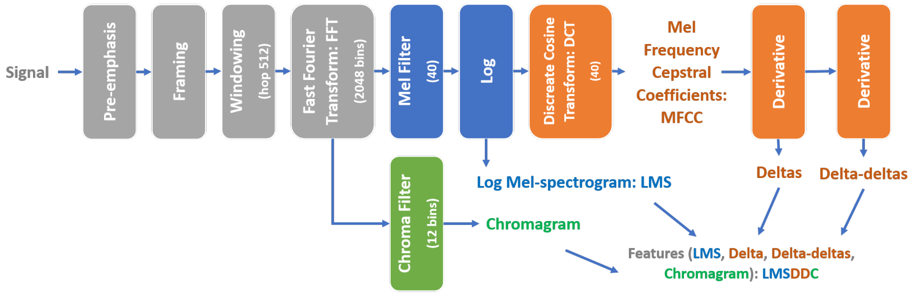

Fig. 2. Feature extraction of LMS and LMSDDC

3.5 Deep Learning

Inspired by learning characteristics of human brain activity, i.e., the more repeated reading, the more comprehension. It is similar with Shanahan's definition, called "repeatedly reads" or "re-reads" [20]. Responding to the use of re-reading theory for improving the accuracy of SER, we designed a feature learning method as shown in Fig. 3, consisting of three sections: (i) main feature learning (MFL), (ii) sub-feature learning (SFL), and (iii) extracted relation of feature distribution (ERFD).

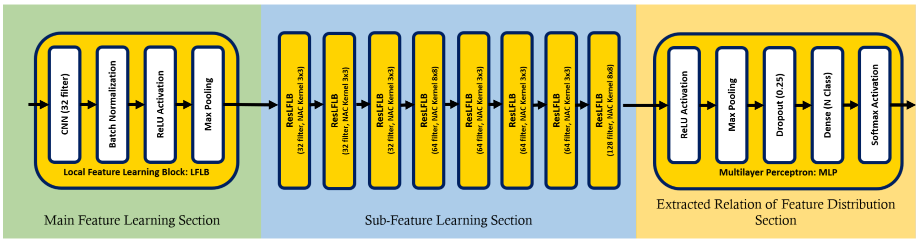
CNN (32 filter)
Batch Normalization
ReLU Activation
Max Pooling
Local Feature Learning Block: LFLB
Main Feature Learning Section
ResLFLB (32 filter, NAC Kernel 3x3)
ResLFLB (32 filter, NAC Kernel 3x3)
ResLFLB (32 filter, NAC Kernel 3x3)
ResLFLB (64 filter, NAC Kernel 8x8)
ResLFLB (64 filter, NAC Kernel 3x3)
ResLFLB (64 filter, NAC Kernel 3x3)
ResLFLB (64 filter, NAC Kernel 3x3)
ResLFLB (128 filter, NAC Kernel 8x8)
Sub-Feature Learning Section
ReLU Activation
Max Pooling
Dropout (0.25)
Dense (N Class)
Softmax Activation
Multilayer Perceptron: MLP
Extracted Relation of Feature Distribution Section

Fig. 3. A feature learning structure based on the re-reading theory.

Main Feature Learning (MFL) Section was designed based on behavior of human brain, like first reading that a human brain starts to learn things. We designed MFLS similar with the LFLB procedure to learn locally basic information as the following steps: (i) 2D-CNN was used to extract necessary features; (ii) BN was applied to enhance learning efficiency of a model; (iii) activation functions converted data to suit for the learning model; and (iv) pooling was for reducing feature size and increased learning speed.

Sub-Feature Learning (SFL) Section was a further learning process that plays a role in assembling repeated reading for deeper learning. In general, LFLB may be at risk of a vanishing gradient problem that affects learning efficiency. Therefore, we have improved the LFLB's efficient by means of residual deep learning, or also known as skipping connections, to skip deeper learning layers that are unnecessary and add more feature details after passing each learning layer; this can avoid the vanishing gradient problem.

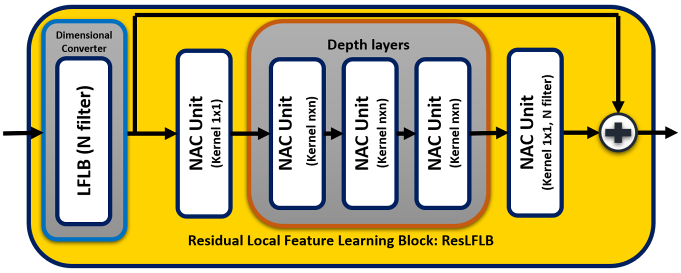

Fig. 4. A residual local feature learning block structure.

For sub-feature learning, there are two important sections: (i) A normal LFLB was used as a preprocessing phase of block to transform input data dimensions into a suitable form for the next section. (ii) Depth layers learned features in more depth by using the network sequence as normalization-activation-CNN (NAC) [31], which is a sequence of sub-layers in residual deep learning. With this structure, it can reduce test error. In addition, more deeper learning layers may be at risk of a vanishing gradient problem; therefore, the skipping connection was added at an input of this section to compensate features lost in deeper layers. This arrangement of two sections is called residual local feature block (ResLFLB) as shown in Fig. 4. The LFLB also known as a dimensional converter and the NAC final layer had the same output filters to determine the number of output filters of ResLFLB. The kernel size in the first and last layer in deep layers were one. Furthermore, in between the first and last learning layers, the bottleneck design, the compression/decompression strategy, was applied to achieve higher learning performance.

Extracted Relation of Feature Distribution (ERFD) Section was implemented with multilayer perceptron (MLP) to extract the relationship of learning results. A sequence of processes are as follows: (i) Activation ReLU transformed the data to the suitable relationship; (ii) Max pooling reduced data size and extracted important data based on maximum values; (iii) Dropout reduced the overfitting of the model; (iv) Dense layers obtained relationships in which neurons equals the number of classes needed to predict; and (v) Activation softmax determined the probability of predicting the emotions.

4 Experiments and Discussion

The proposed DeepResLFLB and LMSDDC were evaluated with two main objectives: (i) classification performance and (ii) model performance. In classification performance, four metrics, accuracy, precision, recall, and F1-score as defined by (1), (2), (3), and (4), respectively, were used for evaluation. In model performance, validation loss was used as monitoring vanishing gradient problems and the number of CNN layer parameters was used as indicating resource-consuming. The experiments were conducted in comparison with three models: the normal ML with fuzzy c-mean [17], traditional LFLB [10], and DeepResLFLB, and two different features: LMS and LMSDDC. Note that, in Tables 2 and 4, Dermircan's method was excluded from those experiments due to a mismatch of feature dimensions.

$$Accuracy = \frac{true\,positive + true\,negative}{number\,of\,data} \tag{1}$$

$$Precision = \frac{true\,positive}{total\,predicted\,positive} \tag{2}$$

$$Recall = \frac{true\,positive}{total\,actual\,positive} \tag{3}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

Dataset preparation Two available published datasets: Berlin emotional database (EMODB) [32] and Ryerson audio-visual database (RAVDESS) [33] were used to evaluate speech emotion performance of our and baseline methods. Two key factors of both selected datasets are the difference in data size and language vocalization that can prove the performance of test methods. EMODB is a German speech in Berlin with 535 utterances and RAVDESS is English speech with 1440 utterances. EMODB dataset was recorded by male and female voices and contained seven different emotions: happiness, sadness, angry, neutral, fear, boredom, and disgust while RAVDESS dataset has one more emotion than EMODB, that is the calm emotion. Here, each dataset was divided into three subsets: 80% for training set, 10% for validation set, and 10% for test set.

Parameter settings for learning model All learning models were set up with the following parameter settings. Learning rate (LR) is very important in deep learning, when compared to step rate to find the minimum gradient. Generally, a high LR may make it over the minimum point. On the other hand, a low LR may take a long time to reach the goal. Here, we choose Adam optimizer for our experiments. Its learning rate and maximum epoch were set to 0.001 and 150, respectively. In addition, plateau strategy was used for reducing the LR and for avoiding overstepping the minimum point. In this case, we set the minimum LR to 0.00001. Batch size of models was set to 10. Finally, if an error value tends to increase, the early stopping criteria is active and then take the model weight with the previous minimum error.

Table 1. A performance comparison of DeepResLFLB and baseline methods with LMS feature, tested on Berlin EMODB dataset.

Method	Accuracy	Precision	Recall	F1-score
Demircan [17]	0.6755±0.0351	0.7549±0.0375	0.6295±0.0346	0.6295±0.0346
1D-LFLB [10]	0.7577±0.0241	0.7609±0.0224	0.7574±0.0318	0.7514±0.0275
2D-LFLB [10]	0.8269±0.0214	0.831±0.0228	0.824±0.0215	0.8233±0.0.0233
DeepResLFLB	**0.8404±0.0225**	**0.8481±0.0225**	**0.8298±0.0236**	**0.8328±0.0244**

Table 2. A performance comparison of DeepResLFLB and baseline methods with LMSDDC feature, tested on Berlin EMODB dataset.

Method	Accuracy	Precision	Recall	F1-score
Demircan [17]	-	-	-	-
1D-LFLB [10]	0.8355±0.0186	0.8385±0.0170	0.8313±0.0205	0.8322±0.0198
2D-LFLB [10]	0.8754±0.0232	0.8802±0.0237	0.8733±0.0237	0.8745±0.0226
DeepResLFLB	**0.8922±0.0251**	**0.8961±0.0212**	**0.8856±0.0322**	**0.8875±0.0293**

Result discussion Based on dataset preparation and parameter setup for learning models, all experiments were used 5-fold validation. Tables 1 and 2 shows performance comparison between LMS and LMSDDC features, respectively, tested on EMODB dataset. It can be seen that LMSDDC feature (Table 2) provided the improvement of accuracy, precision, recall, and F1-score, when compared with LMS feature (Table 1). In the same way, when the same learning models with different features, LMS and LMSDDC, were tested on RAVDESS dataset,

Table 3. A performance comparison of DeepResLFLB and baseline methods with LMS feature, tested on RAVDESS dataset.

Method	Accuracy	Precision	Recall	F1-score
Demircan [17]	0.7528±0.0126	0.7809±0.0109	0.7422±0.0076	0.7479±0.0114
1D-LFLB [10]	0.9487±0.0138	0.9491±0.0134	0.948±0.0123	0.9478±0.0133
2D-LFLB [10]	0.9456±0.0128	0.9438±0.0136	0.946±0.0129	0.9442±0.0135
DeepResLFLB	**0.9602±0.0075**	**0.9593±0.0072**	**0.9583±0.0066**	**0.9584±0.0071**

Table 4. A performance comparison of DeepResLFLB and baseline methods with LMSDDC feature, tested on RAVDESS dataset.

Method	Accuracy	Precision	Recall	F1-score
Demircan [17]	-	-	-	-
1D-LFLB [10]	0.9367±0.0225	0.9363±0.0196	0.9352±0.0218	0.9347±0.0217
2D-LFLB [10]	0.9466±0.0159	0.9475±0.0171	0.9441±0.0162	0.9449±0.0168
DeepResLFLB	**0.949±0.0142**	**0.9492±0.0143**	**0.9486±0.016**	**0.9484±0.0154**

Table 5. A comparison of a number of parameters in DeepResLFLB and LFLB models, tested on EMODB and RAVDESS datasets.

Method	EMODB		RAVDESS	
	LMS	LMSDDC	LMS	LMSDDC
2D-LFLB [10]	260544	262272	260544	262272
DeepResLFLB	**156068**	**163268**	**156074**	**164608**

as shown in Tables 3 and 4, the evaluation results were comparable, not much of an improvement. One of the main reasons is that RAVDESS has less speech variation than EMODB, as reported by Breitenstein research [34]. The less variation of speech leads to the lower quality of features. This made no difference in quality of LMS and LMSDDC features. These results have proved that the LMSDDC feature extracted with three components of human emotions: glottal flow, prosody, and human hearing, usually provided wider speech band frequencies, can improve the speech emotion recognition, especially with high speech variation datasets.

When considering the efficiency of the learning model, Tables 1, 2, 3, and 4 show that DeepResLFLB outperforms the baselines with the highest accuracy, precision, recall, and F1-score. This achievement proved that a learning sequence of DeepResLFLB, imitated from “repeatedly reading” concept of human, is efficient. In addition, DeepResLFLB can avoid a vanishing gradient problem and reduce resource-consuming. Fig. 5 shows that DeepResLFLB had better validation loss and generalization; it can be seen from the graph that has less fluctuation than 2D-LFLB, and Table 5 reports that DeepResLFLB still used fewer parameters than the baseline model around 40%. These results have proved that DeepResLFLB used residual deep learning by arranging its internal network as LFLB, BN, activation function, NAC, and deep layers can solve vanishing gradient and resource-consuming. Besides, when regarding resource-consuming between LMS and LMSDDC, LMSDDC parameters were slightly more than a baseline.

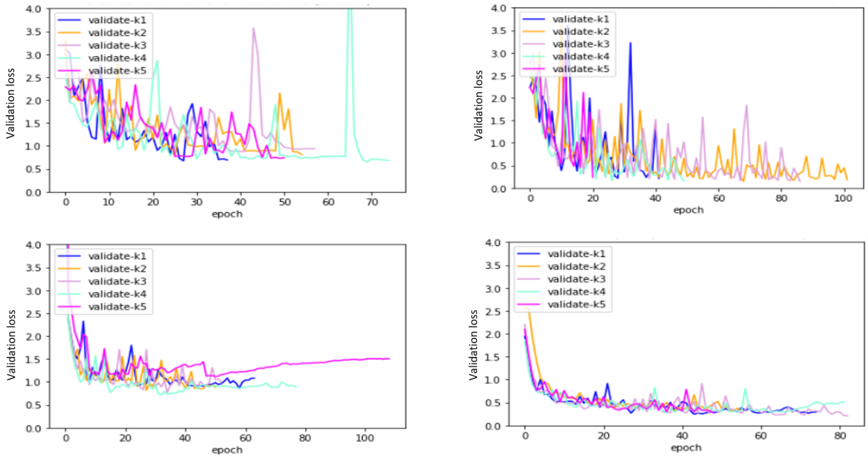

Fig. 5. Validation loss of learning models: conventional LFLB tested on EMODB (top-left), DeepResLFLB tested on EMODB (bottom-left), conventional LFLB tested on RAVDESS (top-right), and DeepResLFLB tested on RAVDESS (bottom-right). Note that only LMS feature was used to test in this experiment.

5 Conclusion

This paper has described a DeepResLFLB model and LMSDDC feature for speech emotion recognition. The DeepResLFLB was redesigned from LFLB based on the 'repeatedly reads' concept while the LMSDDC was emotional feature extracted from speech signals based on human glottal flow and human hearing. Performance of our model and emotional feature was tested on two well-known databases. The results show that the DeepResLFLB can perform better than baselines and use fewer resources in learning layers. In addition, the proposed LMSDDC can outperform conventional LMS.

Although DeepResLFLB presented in this paper have provided better performance in speech emotion recognition, many aspects still can be improved, especially activation function. In future work, we will apply different kinds of activation function in each section of neural network; this will improve the performance of DeepResLFLB.

Acknowledgments

We would like to thank Science Research Foundation, Siam Commercial Bank, for partial financial support to this work.

References

1. Singkul, S., Khampingyot, B., Maharattamalai, N., Taerungruang, S., Chalothorn, T.: Parsing thai social data: A new challenge for thai nlp. In: 2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). pp. 1–7 (2019)
2. Singkul, S., Woraratpanya, K.: Thai dependency parsing with character embedding. In: 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE). pp. 1–5 (2019)
3. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition **44**(3), 572–587 (2011)
4. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review **43**(2), 155–177 (2015)
5. Zhang, Z., Coutinho, E., Deng, J., Schuller, B.: Cooperative learning and its application to emotion recognition from speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**(1), 115–126 (2014)
6. Guidi, A., Vanello, N., Bertschy, G., Gentili, C., Landini, L., Scilingo, E.P.: Automatic analysis of speech f0 contour for the characterization of mood changes in bipolar patients. Biomedical Signal Processing and Control **17**, 29–37 (2015)
7. Gunes, H., Piccardi, M.: Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications **30**(4), 1334–1345 (2007)
8. Bong, S.Z., Wan, K., Murugappan, M., Ibrahim, N.M., Rajamanickam, Y., Mohamad, K.: Implementation of wavelet packet transform and non linear analysis for emotion classification in stroke patient using brain signals. Biomedical signal processing and control **36**, 102–112 (2017)
9. Yuvaraj, R., Murugappan, M., Ibrahim, N.M., Sundaraj, K., Omar, M.I., Mohamad, K., Palaniappan, R.: Detection of emotions in parkinson's disease using higher order spectral features from brain's electrical activity. Biomedical Signal Processing and Control **14**, 108–116 (2014)
10. Zhao, J., Mao, X., Chen, L.: Speech emotion recognition using deep 1d & 2d cnn lstm networks. Biomedical Signal Processing and Control **47**, 312–323 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Wu, S., Falk, T.H., Chan, W.Y.: Automatic speech emotion recognition using modulation spectral features. Speech communication **53**(5), 768–785 (2011)
13. He, L., Lech, M., Maddage, N.C., Allen, N.B.: Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech. Biomedical Signal Processing and Control **6**(2), 139–146 (2011)
14. Pérez-Espinosa, H., Reyes-Garcia, C.A., Villaseñor-Pineda, L.: Acoustic feature selection and classification of emotions in speech using a 3d continuous emotion model. Biomedical Signal Processing and Control **7**(1), 79–87 (2012)
15. Huang, Z., Dong, M., Mao, Q., Zhan, Y.: Speech emotion recognition using cnn. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 801–804 (2014)
16. Huang, Y., Wu, A., Zhang, G., Li, Y.: Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition. IET Signal Processing **9**(4), 341–348 (2015)

17. Demircan, S., Kahramanli, H.: Application of fuzzy c-means clustering algorithm to spectral features for emotion classification from speech. Neural Computing and Applications **29**(8), 59–66 (2018)
18. Sun, Y., Wen, G., Wang, J.: Weighted spectral features based on local hu moments for speech emotion recognition. Biomedical signal processing and control **18**, 80–90 (2015)
19. Sari, S.W.W.: The influence of using repeated reading strategy towards student's reading comprehension. Proceeding 1st Annual International Confrence on Islamic Education and Language: The Education and 4.0 Industrial Era in Islamic Perspective p. 71 (2019)
20. Shanahan, T.: Everything you wanted to know about repeated reading. Reading Rockets. https://www.readingrockets.org/blogs/shanahan-literacy/everything-you-wanted-know-about-repeated-reading (2017)
21. Venkataramanan, K., Rajamohan, H.R.: Emotion recognition from speech (2019)
22. Soekhoe, D., Putten, P., Plaat, A.: On the impact of data set size in transfer learning using deep neural networks. pp. 50–60 (2016)
23. Park, D.S., Chan, W., Zhang, Y., Chiu, C.C., Zoph, B., Cubuk, E.D., Le, Q.V.: Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779 (2019)
24. Jagini, N.P., Rao, R.R.: Exploring emotion specific features for emotion recognition system using pca approach. In: 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). pp. 58–62 (2017)
25. Degottex, G.: Glottal source and vocal-tract separation. Ph.D. thesis (2010)
26. Doukhan, D., Carrive, J., Vallet, F., Larcher, A., Meignier, S.: An open-source speaker gender detection framework for monitoring gender equality. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5214–5218. IEEE (2018)
27. Doval, B., d'Alessandro, C., Henrich, N.: The spectrum of glottal flow models. Acta acustica united with acustica **92**(6), 1026–1046 (2006)
28. Wang, Y., Guan, L.: Recognizing human emotional state from audiovisual signals. IEEE transactions on multimedia **10**(5), 936–946 (2008)
29. Robinson, K., Patterson, R.D.: The stimulus duration required to identify vowels, their octave, and their pitch chroma. The Journal of the Acoustical Society of America **98**(4), 1858–1865 (1995)
30. Wakefield, G.H.: Chromagram visualization of the singing voice. In: International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (1999)
31. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)
32. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of german emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
33. Livingstone, S.R., Russo, F.A.: 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 **13**(5) (2018)
34. Breitenstein, C., Lancker, D.V., Daum, I.: The contribution of speech rate and pitch variation to the perception of vocal emotions in a german and an american sample. Cognition & Emotion **15**(1), 57–79 (2001)