Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 10, No. 4, 2022, 539-547.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Speech Emotion Recognition using Enriched Spectrogram and Deep Convolutional Neural Network Transfer Learning

Bibi Zahra Mansouri[1], Hamid Reza Ghaffary[1] and Ali Harimi[1,2*]

1. Electrical and Computer Engineering Department, Ferdows branch, Islamic Azad University, Ferdows, Iran.
2. Electrical and Computer Engineering Department, Shahrood branch, Islamic Azad University, Shahrood, Iran.

Article Info

Article History:
Received 28 August 2022
Revised 02 October 2022
Accepted 04 November 2022

DOI:10.22044/jadm.2022.12241.2372

Keywords:
Wideband and Narrowband Spectrogram, ResNet152, DCNN, Transfer Learning, Speech Emotion Recognition.

**Corresponding author: a.harimi@gmail.com (A. Harimi).*

Abstract

Speech emotion recognition (SER) is a challenging field of research that has attracted attention during the last two decades. Feature extraction has been reported as the most challenging issue in the SER systems. Deep neural networks could partially solve this problem in some other applications. In order to address this problem, we propose a novel enriched spectrogram calculated based on the fusion of wide-band and narrow-band spectrograms. The proposed spectrogram benefits from both high temporal and spectral resolution. Then we apply the resultant spectrogram images to the pre-trained deep convolutional neural network, ResNet152. Instead of the last layer of ResNet152, we add five additional layers to adopt the model to the present task. All the experiments performed on the popular EmoDB dataset are based on leaving one speaker out of a technique that guarantees the speaker's independence from the model. The model gains an accuracy rate of 88.97%, which shows the efficiency of the proposed approach in contrast to other state-of-the-art methods.

1. Introduction

Speech emotion recognition (SER) has received great attention during the last two decades. Speech signal coveys linguistic and paralinguistic information. While linguistic information refers to the contents of the speech, paralinguistic information addresses issues such as emotion, age, gender, illness, and drug consumption of the speaker. Currently, although machines can successfully recognize humans' speech, we are far from natural communication with machines. It is mostly due to the inability of machines to recognize humans' emotions and respond correspondingly.

From a classical viewpoint, SER can be considered as a pattern recognition problem with three major steps: feature extraction, feature selection/reduction, and classification. Feature extraction is the process of quantitating the original speech samples, in which each sample is represented by a set of quantities called the feature vector. Since the extraction of discriminative features from speech samples remains a challenging open problem, the brute force feature extraction technique is commonly used, which results in a huge number of unnecessary and noisy features. This arises the need for feature selection (or reduction) techniques to avoid the curse of dimensionality or overfitting. Finally, a classifier is trained to link the feature vectors (or equivalently the corresponding samples) to a target emotion label. In the testing phase, the trained classifier can predict the emotions of the unseen samples.

As the feature extraction task, fundamental frequency (pitch), intensity (energy), Mell Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) coefficients, Linear Prediction Coefficients (LPC), formants, harmonics' energy, and a variety of handcraft features are successfully applied to SER [1-11]. In this regard, the Interspeech 2009/2011 challenge introduced a set of standard features for SER [12, 13]. Moreover, several speech feature extraction tools such as Pratt, SPAC, and openSMILE have been introduced in the literature. Classifier models such as hidden Markov models [14, 15], neural

networks [16, 17], and support vector machines [18-20] have been used in various studies.

In the recent years, deep learning methods could address the feature extraction challenge by providing effective automatic feature learning schemes [21]. While in some studies low-level features are applied to deep neural networks (DNNs) for further analysis, in sophisticated end-to-end networks, raw signal is applied to DNNs and the network unifies the three stages feature extraction, feature reduction, and classification tasks within a single network [8]. In such a scheme, the network automatically learns to extract effective features. In this regard, the authors in [22, 23], proposed a 1-layer CNN with a Sparse Auto-Encoder (SAE) for learning the extraction of affective features from speech for SER. Also [24] combined a 2-layer CNN with a Long Short-Term Memory (LSTM) [25] to form an end-to-end SER system. In all the mentioned models, a 1-D convolution both in the frequency domain [23, 26, 27] or in the time domain [24] has been used. In contrast, while in Deep Convolutional Neural Networks (DCNNs), 2-D convolution is widely used [28, 29], these are much deeper than the mentioned 1-layer or 2-layer CNNs [28, 29]. These DCNNs consist of several layers with up to hundreds of millions of trainable parameters. However, training such a huge number of trainable parameters requires a large-scale dataset. On the other hand, in many cases, only a limited number of labeled samples are available, which leads to the overfitting of DNN. Transfer learning techniques can effectively solve this problem. in transfer learning, a network that has been pre-trained on a large-scale dataset is employed for a new dataset. While most preliminary layers of the pre-trained model are kept frozen, a limited number of the last layers are permitted to fine-tune based on the current small sample size dataset. Since the earlier layer of a DNN are similar in various tasks and only extract low-level features, it does not have a great impact on the model performance. Instead, the upper layers that are responsible for extracting high-level features and classification will adopted well to the present task during the fine-tuning procedure [30].

The great success of transfer learning methods in image processing tasks in contrast to speech processing motivates the researchers to convert one-dimensional speech signals into two-dimensional images to benefit from such models [31]. Indeed, converting the speech signal to an image allows using a variety range of image processing tools for the analysis of speech signals. In this regard, in [8] the authors proposed a new transformation called CyTex to convert the speech signal to an image and employed pre-trained DCNNs such as AlexNet, InceptionV3, VGG16, and ResNet50 for SER. Authors in [32] provide a new transform called chaogram that is a color image derived from reconstructed phase space of speech and envisions the chaotic behavior of the signal. The extracted images were applied to a DCNN with a gray wolf optimization algorithm employed for the optimization of model hyper-parameters. In [31], the speech signal is converted to an image using spectrogram analysis. The spectrogram represents the signal energy based on time frames and frequency bands. In this representation, the intensity of each image pixel is proportional to the signal energy at that time window and frequency range. Although it is an effective method to convert the speech signal to an image, the researchers try to improve it using different strategies. In [33], a multi-feature fusion with spectrogram augmentation is applied to a DCNN to perform speech emotion recognition. The authors claim that their method improves the performance of SER through a noise removal and pitch tuning procedure. In another similar work, a multi-type features separating fusion learning framework has been used for SER [34]. They employed a DCNN to extract image information from the spectrogram of speech. Then they fused different types of features. One of the limitations of the spectrogram is that the quality of the spectrogram image regarding the time or frequency axes depends on the selected frame length of the speech signal. While longer time frames provide us with a high-frequency resolution (narrow band spectrogram), the shorter frames result in more time resolution (wide band spectrogram). Indeed, it is a trade-off between time resolution and frequency resolution based on the selected time frame. Another issue about the spectrogram transformation is the information loss. As we know, the spectrogram is a one-directional transformation; the spectrogram image cannot convert to the original speech signal since some data has been lost during the transformation process. Here, we propose to fuse the spectrogram images achieved by various window lengths to enrich the produced spectrogram image, which is the main contribution of this study. Next, we fine-tuned the modified pre-trained ResNet152 model using the proposed spectrogram images on the well-known public EMODB dataset.

The rest of the paper is organized as what follows. Our proposed method is detailed in Section 2. Experimental results and discussion are presented

in Section 3. The paper ends in Section 4 with a summary, main conclusions, and recommendations for possible future research works.

2. Proposed Method

The proposed SER system consists of two main steps. First, the one-dimensional speech signals are converted to enriched spectrogram images. Next, the generated RGB images are fed to the pre-trained ResNet152 model to fine-tune it for SER. Figure 1 shows the overall block diagram of the proposed method. As it can be seen in Figure 1, the procedure comprises two train and test phases. The process of preparing the enriched spectrogram images for DCNN is similar in both phases. The model is fine-tuned using train data (including samples and labels). The trained model is utilized to predict the target emotions from the input samples that have been kept unseen during the training phase. Each part of the system is described in the following.

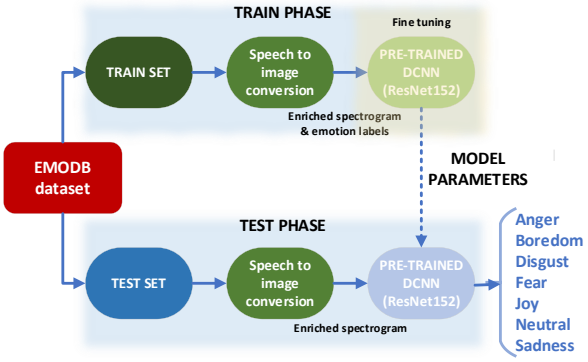
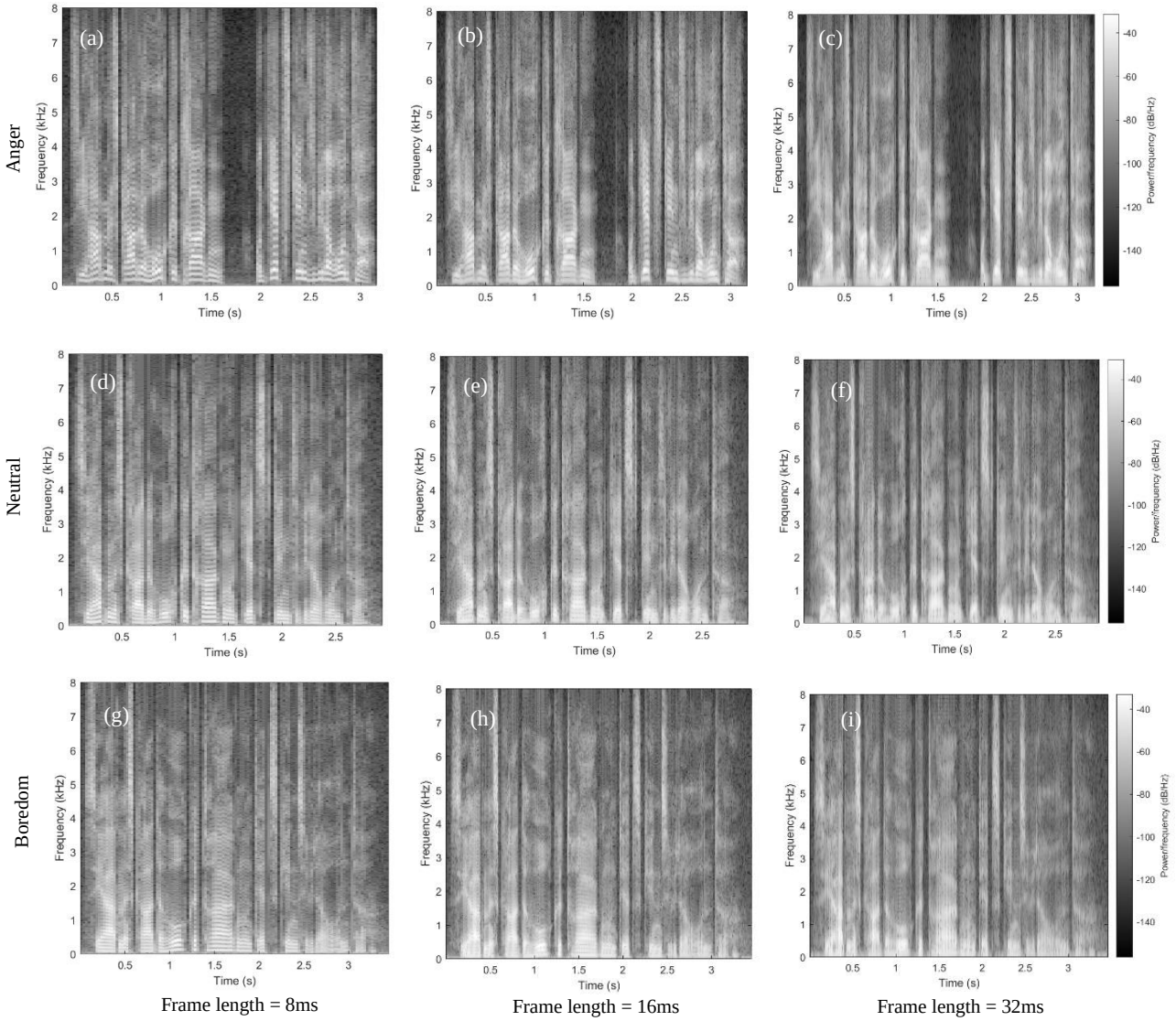


Figure 1. Overall framework of the proposed SER system.



Frame length = 8ms | Frame length = 16ms | Frame length = 32ms

Figure 2. Spectrograms of a sentence from the EMODB dataset uttered by a speaker under three different emotions (first row: anger (03b02Wb), second row: neutral (03b03Nb), third row: boredom (03b02La)), calculate with three different frame length (first column: 8ms, second column: 16ms, and third column: 32ms). (a,b,c) sentence 03b02Wb, (d,e,f) sentence 03b03Nb, and (g, h, i) sentence 03b02La, while the first item determined by 8ms window, the second and third ones determined by 16ms, and 32ms window lengths, respectively.

2.1. Enriched spectrogram

The log Mel-spectrogram is an effective method to decompose the one-dimensional speech signal based on its frequency components in short time intervals and represent it as an image. The length of the frames has a significant impact on the quality of the resultant image. In this regard, short-time frames result in wide-band spectrogram images with the high temporal and low spectral resolution, and wise versa. Figure 2 illustrates spectrogram images of a speech sentence uttered by a speaker under different emotions calculated by three different time frames.

As it can be seen in Figure 2, the narrow band spectrograms that are calculated by shorter frames provide higher frequency resolution images than wide band spectrograms calculated by longer frames that provides higher time resolution images. Indeed, the frequency resolution is determined based on the sampling frequency of the signal, the FFT length, the frame size, and the window type. The sampling rate of audio signals used in this work is set to 16khz. The order of FFT is set to 1024, which provide spectrogram images with 512 pixels in height. The rectangular window has been used in all the experiments. The frame length of 8ms, 16ms, and 32ms with 50% overlap has been used for various wide-band, mid-band, and narrowband spectrograms. Although these images are different in some features, all of them seem to be informative. Here, in order to benefit from all information provided in these spectrograms, we propose to treat them as the three R, G, and B color channels of an RGB image. Consequently, the enriched spectrogram is constructed based on three spectrograms obtained by three different time frames (8ms, 16ms, and 32ms).

2.2. DCNN model

Here, we employed the pre-trained DCNN model, ResNet152 [35] with some modifications. This network has been trained on the ImageNet dataset that consists of various sorts of images including texture images. Due to the similarity of spectrogram images to texture images, it is expected that using the transfer learning method, this network can extract features from spectrograms as well as images of the ImageNet dataset. The architecture of the proposed DCNN model is illustrated in Figure 3.

As we know, the last layer of each DNN model performs the classification task. When we use pre-trained models for a new task, the model should be customized. Hence, we should make some changes in the last layer of the ResNet152 to customize and fine-tune the model on the current dataset for speech emotion recognition. To this end, as shown in Figure 3, we removed the last fully connected layer of ResNet152, and instead, we added a sequence of dropout and linear fully connected layers. In this regard, five layers are added as follows: one dropout layer with $p = 0.4$ that helps to avoid overfitting, followed by a flatten and a fully connected layer. Then normalization is performed by a batch normalization layer, and finally, a SoftMax layer performs the classification task. During the training, the first three blocks of the ResNet152 are kept frozen, while other layers are permitted to train based on spectrogram images. ResNet152 is deeper and at the same time less complex (due to the fewer trainable parameters) than networks such as VGG-16/19. The reason for employing this network among so many other available networks is the results reported by [8] that compared several networks. [8] concluded that the ResNet152 is the best choice for SER compared to VGG-16, ResNet50, and ResNet101.

3. Experiments and Results

All the experiments have been performed on the EMODB dataset. Algorithms have been implemented on a laptop (ASUS TUF GAMING A17) with CPU-AMD RYZEN9 4900H, NVIDIA-GTX 1660Ti, RAM-16GB DDR4 specifications. All the codes have been written with Python 3.8, and simulations have been performed on the Spyder IDE in the Anaconda environment. The pre-trained DCNN models have been implemented on the TensorFlow framework by the high-level library, Keras Chollet (2021).

3.1. Database

The Berlin emotional speech corpus (Emo-DB) [36] is a well-known publicly available standard database. Due to its wide use in SER, it allows the researchers to compare their results with each other. This dataset consists of 535 speech samples expressed by ten German speakers (five males and five females) under seven basic emotions (anger/127 samples, boredom/81 samples, disgust/46 samples, fear/69 samples, joy/71 samples, neutral/79 samples, and sadness/ 62 samples). All the samples were recorded at a 16 kHz sampling rate with a 16-bit resolution.

3.2. Experimental results

All the experiments have been performed using the Leave One Speaker Out (LOSO) technique.

As we know, emotions are highly correlated to the identity of the speaker. However, in order to enforce the algorithm to learn the relationship of the speech and the emotions independently of the identity of the speaker, it is common to train the system with some speakers and test it with other speakers that do not have any shared sample with train data.

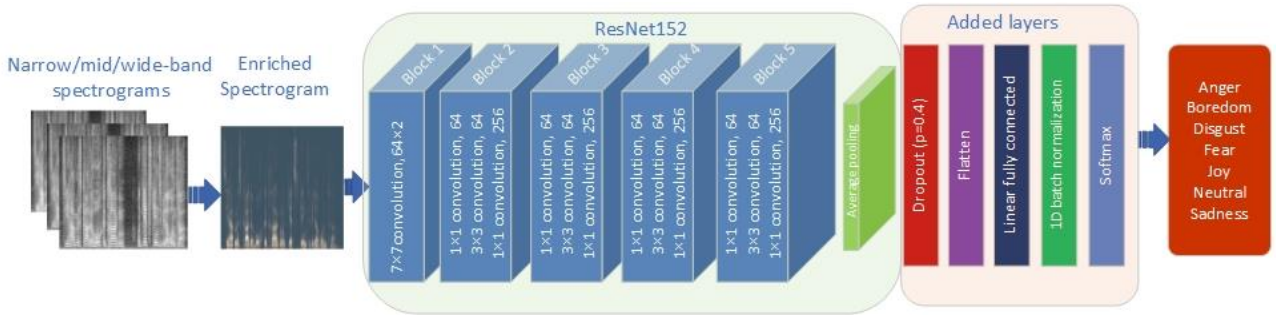


Figure 3. Architecture of the proposed DCNN model.

This task can easily be performed on the Emo-DB dataset since it comprises speech samples from 10 speakers. Therefore, the experiments have been performed in 10 independent trials. In each trial, while samples of 9 speakers have been used for training, the samples of the remaining speaker kept unseen for testing. The procedure continued until all ten speakers participated in the test phase.

Moreover, in order to evaluate the proposed enriched spectrograms, experiments have been performed on simple spectrograms to estimate how the fusion of such simple spectrograms can improve the classification performance. Tables 1 to 3 show the results achieved by wide-band (8ms time frame), mid-band (16ms time frame), and narrow-band (32ms time frame) spectrograms in terms of their confusion matrices, respectively.

Table 1. Confusion matrix of wide-band spectrogram (8ms time frame).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Recall (%)
Anger	99	4	1	9	11	0	3	77.95
Boredom	0	69	0	0	0	5	7	85.19
Disgust	0	3	40	2	0	1	0	86.96
Fear	2	1	1	62	3	0	0	89.86
Joy	5	1	0	0	65	0	0	91.55
Neutral	0	3	1	0	0	73	2	92.41
Sadness	3	2	0	1	0	3	53	85.48
Precision	90.83	83.13	93.02	83.78	82.28	89.02	81.54	
Accuracy = 86.17								

Table 2. Confusion matrix of mid-band spectrogram (16ms time frame).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Recall (%)
Anger	102	3	5	7	8	0	2	80.31
Boredom	0	68	1	0	0	6	6	83.95
Disgust	0	2	41	2	0	1	0	89.13
Fear	2	3	0	61	2	0	0	88.41
Joy	2	1	0	1	67	0	0	94.37
Neutral	0	1	1	0	0	75	2	94.94
Sadness	4	1	0	1	0	2	54	87.10
Precision	92.73	86.08	85.42	84.72	87.01	89.29	84.38	
Accuracy = 86.73								

Table 3. Confusion matrix of narrow-band spectrogram (32ms time frame).

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Recall (%)
Anger	105	1	3	7	9	0	1	82.68
Boredom	1	63	2	1	1	7	6	77.78
Disgust	0	2	41	2	0	1	0	89.13
Fear	1	2	0	62	4	0	0	89.86
Joy	3	1	0	1	66	0	0	92.96
Neutral	0	2	1	0	0	74	2	93.67
Sadness	5	1	0	2	1	1	52	83.87
Precision	91.30	88.73	87.23	82.67	81.48	89.16	85.25	
Accuracy = 86.54								

Table 4. Confusion matrix of proposed enriched spectrogram.

Emotion	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness	Recall (%)
Anger	106	1	3	7	7	1	2	83.46
Boredom	0	67	3	0	0	5	6	82.72
Disgust	0	1	43	0	0	0	1	93.48
Fear	1	1	0	62	2	2	0	89.86
Joy	1	0	0	2	68	0	0	95.77
Neutral	1	1	0	0	2	74	1	93.67
Sadness	2	1	0	2	1	0	56	90.32
Precision	95.50	93.06	87.76	84.93	85.00	90.24	84.85	
Accuracy = 88.97								

Table 4 shows the confusion matrix achieved using the proposed enriched spectrogram. In these tables, the columns show the recognized emotions, while rows represent the true emotions. The last column of the table shows the recall that is determined as the correctly recognized samples of each class divided by the total number of samples within the class. Indeed, it is the recognition rate of each class. The precision is a measure to calculate the reliability of the classification result of each class. It is calculated as the correctly classified samples of each class divided by the total number of samples assigned to that class by the classifier, and accuracy or recognition rate as the main criteria for evaluating each method is determined as the total number of correctly classified samples divided by the total number of samples within the dataset.

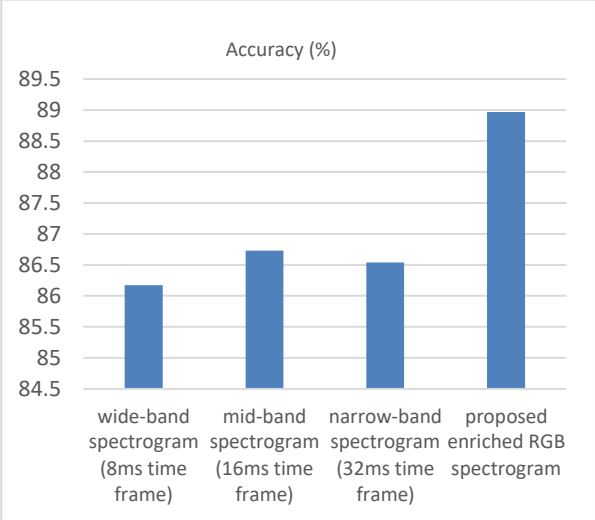


Figure 4. Accuracy achieved by wide-band spectrogram (8ms time frame), mid-band spectrogram (16ms time frame), narrow band spectrogram (32ms time frame), and the proposed enriched RGB spectrogram.

As we can see in Tables 1 to 3, the accuracy rate achieved by the mid-band spectrogram is slightly higher than that of the wide-band and narrow-band spectrograms. However, the proposed enriched spectrogram shows a significant improvement in accuracy rate in comparison to the simple spectrograms calculated by a fixed time frame length. Figure 4 compares the results achieved by these four spectrograms.

Table 5. Comparison between the results of our proposed method and some of the state-of-the-art methods on the same dataset and experiment conditions. (LLD and HSF stand for low-level descriptors and high-level statistical functions, respectively.)

Method	Aaccuracy (%)	Input
[37](2015)	73.75	LLDs and HSFs
[38](2021)	77.5	GA-Optimized feature set
[39] (2014)	80.10	time-frequency vocal feature
[40] (2015)	81.74	Local Hu moments
[41](2020)	82.00	Mel-Spectrogram
[42](2019)	82.32	LLDs and HSFs
[43](2018)	82.41	LLDs and HSFs
[44](2020)	82.82	LLDs and HSFs
[45] (2020)	86.10	Spectrogram, LLDs, HSFs
[46] (2016)	87.21	LLDs and HSFs
Ours	88.97	Enriched RGB spectrogram

As it can be seen in Figure 4, the fusion of the three spectrograms achieved with three different time frame lengths can enrich the spectrogram which results in a remarkable increase in accuracy rate. Table 5 compares the results achieved using the proposed SER system with some of the state-of-the-art recent studies in this field.

As it is illustrated in Table 5, thanks to the new information added to the proposed enriched spectrogram by the fusion of three simple spectrograms, the proposed method gains a higher accuracy rate than other reviewed methods. It should be noted that we compare our results with only speaker-independent methods. As speaker-dependent schemes, there are several methods that reported much higher accuracy rates in the literature. For example, the recognition rate of 89.16 has been reported for the classification of seven emotions on the berlin dataset using a model based on multi-level local binary and local ternary patterns [47]. The accuracy rate of 90.21% was achieved using acoustic and deep features [48]. Authors in [49] reached an accuracy rate of 95.33% using the Mel-spectrogram of speech. Also in [8], a recognition accuracy of 95.83 is reported using Mel-spectrogram, while with the same network but a novel speech-to-image transformation called CyTex they reached to an accuracy of 96.14%.

4. Conclusion

Converting one-dimensional speech signals to images is a beneficial task that paves the path for employing many image-processing tools for speech-processing applications. In this work, we proposed an enriched spectrogram as an informative two-dimensional representation of speech signals for the recognition of emotion from speech. As we know, in speech emotion recognition both time and frequency features are important, and indeed they are complements of each other. The proposed enriched spectrogram calculated by fusion of wide-, mid-, and narrow-band spectrograms enables us to benefit from both time and frequency high resolution. According to our experiments, among the three frame lengths of 8ms, 16ms, and 32ms, the best and worst accuracies achieved by spectrograms were calculated based on 16ms and 8ms frame lengths, respectively. However, a combination of these three spectrograms improves the recognition rate significantly. Investigating in confusion matrices shows that most of the error rate occurred for anger samples misclassified as fear and joy. It seems that a tandem classifier can be employed as an error correction stage to reduce the overall error rate. We consider it in our future plans. As the other future work, we plan to design a network that fuses several spectrograms and make a single grayscale high-quality spectrogram that is high resolution both in time and frequency.

Acknowledgment

This work was supported by the Technology Grant Program (Javaneh) of the Ministry of Science, Research, and Technology, Tehran, Iran, and Semnan Science and Technology Park, Semnan, Iran [grant number 160002000771].

References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition,* vol. 44, no. 3, pp. 572-587, 2011.

[2] E. H. Kim, K. H. Hyun, S. H. Kim, and Y. K. Kwak, "Improved emotion recognition with a novel speaker-independent feature," *IEEE/ASME transactions on mechatronics,* vol. 14, no. 3, pp. 317-325, 2009.

[3] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, "Formant position based weighted spectral features for emotion recognition," *Speech Communication,* vol. 53, no. 9-10, pp. 1186-1197, 2011.

[4] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," *Artificial Intelligence Review,* vol. 43, no. 2, pp. 155-177, 2015.

[5] A. Harimi, A. AhmadyFard, A. Shahzadi, and K. Yaghmaie, "Anger or joy? Emotion recognition using nonlinear dynamics of speech," *Applied Artificial Intelligence,* vol. 29, no. 7, pp. 675-696, 2015.

[6] A. Shahzadi, A. Ahmadyfard, A. Harimi, and K. Yaghmaie, "Speech emotion recognition using nonlinear dynamics features," *Turkish Journal of Electrical Engineering & Computer Sciences,* vol. 23, 2015.

[7] A. Harimi, H. S. Fakhr, and A. Bakhshi, "Recognition of emotion using reconstructed phase space of speech," *Malaysian Journal of Computer Science,* vol. 29, no. 4, pp. 262-271, 2016.

[8] A. Bakhshi, A. Harimi, and S. Chalup, "CyTex: Transforming speech to textured images for speech emotion recognition," *Speech Communication,* vol. 139, pp. 62-75, 2022/04/01/ 2022, doi: https://doi.org/10.1016/j.specom.2022.02.007.

[9] H. Marvi, Z. Esmaileyan, and A. Harimi, "Estimation of LPC coefficients using Evolutionary Algorithms," *Journal of AI and Data Mining,* vol. 1, no. 2, pp. 111-118, 2013, doi: 10.22044/jadm.2013.115.

[10] A. Harimi, A. Shahzadi, A. Ahmadyfard, and K. Yaghmaie, "Classification of emotional speech using spectral pattern features," *Journal of AI and Data Mining,* vol. 2, no. 1, pp. 53-61, 2014, doi: 10.22044/jadm.2014.150.

[11] E. Kalhor and B. Bakhtiari, "Multi-Task Feature Selection for Speech Emotion Recognition: Common Speaker-Independent Features Among Emotions," *Journal of AI and Data Mining,* vol. 9, no. 3, pp. 269-282, 2021, doi: 10.22044/jadm.2021.9800.2118.

[12] B. Schuller, S. Steidl, and A. Batliner, *The Interspeech 2009 Emotion Challenge*. 2009, pp. 312-315.

[13] B. Schuller, A. Batliner, S. Steidl, F. Schiel, and J. Krajewski, *The interspeech 2011 speaker state challenge*. 2011, pp. 3201-3204.

[14] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Transactions on Multimedia,* vol. 14, no. 1, pp. 142-156, 2011.

[15] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).* 2003, vol. 2: Ieee, pp. II-1.

[16] M. Bejani, D. Gharavian, and N. M. Charkari, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks," *Neural Computing and Applications,* vol. 24, no. 2, pp. 399-412, 2014.

[17] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural computing & applications*, vol. 9, no. 4, pp. 290-296, 2000.

[18] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, p. 104886, 2019.

[19] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *2004 IEEE international conference on acoustics, speech, and signal processing*, 2004, vol. 1: IEEE, pp. I-577.

[20] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6-9, 2010.

[21] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2528-2536, 2016.

[22] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech Emotion Recognition Using CNN," presented at the Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA, 2014. [Online]. Available: https://doi.org/10.1145/2647868.2654984.

[23] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203-2213, 2014, doi: 10.1109/TMM.2014.2360798.

[24] G. Trigeorgis *et al.*, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 20-25 March 2016 2016, pp. 5200-5204, doi: 10.1109/ICASSP.2016.7472669.

[25] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[26] D. Guiming, W. Xia, W. Guangyan, Z. Yan, and L. Dan, "Speech recognition based on convolutional neural networks," in *2016 IEEE International Conference on Signal and Image Processing (ICSIP)*, 13-15 Aug. 2016 2016, pp. 708-711, doi: 10.1109/SIPROCESS.2016.7888355.

[27] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 801-804.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017, doi: 10.1145/3065386.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in *Computer Vision – ECCV 2014*, Cham, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., 2014// 2014: Springer International Publishing, pp. 346-361.

[30] F. Chollet, *Deep learning with Python*. Manning New York, 2018.

[31] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576-1590, 2017.

[32] M. Falahzadeh, F. Farokhi, A. Harimi, and R. Sabbaghi, "Deep Convolutional Neural Network and Gray Wolf Optimization Algorithm for Speech Emotion Recognition," *Circuits, Systems, and Signal Processing*, pp. 1-44, 08/25 2022, doi: 10.1007/s00034-022-02130-3.

[33] S. Jothimani and K. Premalatha, "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons & Fractals*, vol. 162, p. 112512, 2022/09/01/ 2022, doi: https://doi.org/10.1016/j.chaos.2022.112512.

[34] X. Xu, D. Li, Y. Zhou, and Z. Wang, "Multi-type features separating fusion learning for Speech Emotion Recognition," *Applied Soft Computing*, vol. 130, p. 109648, 2022/11/01/ 2022, doi: https://doi.org/10.1016/j.asoc.2022.109648.

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[36] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Ninth european conference on speech communication and technology*, 2005.

[37] S. M S, A. Elampulakkadu, T. Deepa, C. Shameema, and S. Rajan, *Emotion recognition from audio signals using Support Vector Machine*. 2015, pp. 139-144.

[38] S. Kanwal and S. Asghar, "Speech Emotion Recognition Using Clustering Based GA-Optimized Feature Set," *IEEE Access*, vol. 9, pp. 125830-125842, 2021, doi: 10.1109/ACCESS.2021.3111659.

[39] L. Zão, D. Cavalcante, and R. Coelho, "Time-Frequency Feature and AMS-GMM Mask for Acoustic Emotion Classification," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 620-624, 2014, doi: 10.1109/LSP.2014.2311435.

[40] H. Tao, R. Liang, C. Zha, X. Zhang, and L. Zhao, "Spectral Features Based on Local Hu Moments of Gabor Spectrograms for Speech Emotion Recognition," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 8, pp. 2186-2189, 2016, doi: 10.1587/transinf.2015EDL8258.

[41] M. Lech, M. N. Stolar, C. Best, and R. S. Bolia, "Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding," in *Frontiers in Computer Science*, 2020.

[42] S. Sekkate, M. Khalil, A. Abdellah, and S. Jebara, "An Investigation of a Feature-Level Fusion for Noisy Speech Emotion Recognition," *Computers*, vol. 8, p. 91, 12/13 2019, doi: 10.3390/computers8040091.

[43] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech Emotion Recognition: Methods and Cases Study," in *ICAART*, 2018.

[44] F. Daneshfar and S. J. Kabudian, "Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm," *Multimedia Tools Appl.*, vol. 79, no. 1–2, pp. 1261–1289, 2020, doi: 10.1007/s11042-019-08222-8.

[45] D. Issa, M. F. Demirci, and A. Yazıcı, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control.*, vol. 59, p. 101894, 2020.

[46] A. Shirani and A. R. N. Nilchi, "Speech Emotion Recognition based on SVM as Both Feature Selector and Classifier," *International Journal of Image, Graphics and Signal Processing*, vol. 8, pp. 39-45, 2016.

[47] Y. Ü. Sönmez and A. Varol, "A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns," *IEEE Access*, vol. 8, pp. 190784-190796, 2020, doi: 10.1109/ACCESS.2020.3031763.

[48] M. B. Er, "A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features," *IEEE Access*, vol. 8, pp. 221640-221653, 2020, doi: 10.1109/ACCESS.2020.3043201.

[49] Z. Zhao *et al.*, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97515-97525, 2019.

تشخیص احساس از روی گفتار به کمک غنی سازی طیف نگاره و انتقال یادگیری در شبکه عصبی عمیق کانولوشنی

بی بی زهرا منصوری[۱]، حمیدرضا غفاری[۱] و علی حریمی[۱،۲*]

[۱] گروه مهندسی برق و کامپیوتر، واحد فردوس، دانشگاه آزاد اسلامی، فردوس، ایران.

[۲] گروه مهندسی برق و کامپیوتر، واحد شاهرود، دانشگاه آزاد اسلامی، شاهرود، ایران.

ارسال ۲۰۲۲/۰۸/۲۸؛ بازنگری ۲۰۲۲/۱۰/۰۲؛ پذیرش ۲۰۲۲/۱۱/۰۴

چکیده:

تشخیص احساس از روی گفتار یکی از حوزه های پژوهشی چالش برانگیز است که در دو دهه ی اخیر توجه بسیاری از پژوهشگران را به خود جلب نموده است. استخراج ویژگی بعنوان یکی از مهمترین چالش های سیستم‌های تشخیص احساس از روی گفتار معرفی شده است. شبکه‌های عصبی عمیق تا حدودی در حل این مشکل در سایر حوزه های پژوهشی موفق بوده‌اند. به منظور حل این مشکل، در این تحقیق یک بازنمایی جدید از طیف نگاره با استفاده از همجوشی طیف نگاره پهن باند و باریک باند معرفی کردیم. طیف نگاره پیشنهادی از مزایای رزولوشن بالا هم در بعد زمان و هم در بعد فرکانس بهره می‌برد. طیف نگاره پیشنهادی به یک شبکه عصبی عمیق کانولوشنی از قبل آموزش دیده، ResNet152، اعمال می‌شود. به جای لایه-ی آخر شبکه مذکور، پنج لایه‌ی جدید به شبکه اضافه کردیم تا مدل با کاربرد مورد نظر انطباق داده شود. همه ی آزمایش‌ها با استفاده از پایگاه داده‌ی معرف برلین و با تکنیک مستقل از گوینده انجام شدند. مدل پیشنهادی به دقت ۸۸,۹۷٪ دست پیدا کرد که بیانگر کارآمد بودن روش پیشنهادی نسبت به سایر روش‌های جدید مقایسه شده می‌باشد.

کلمات کلیدی: طیف نگاره پهن باند و باریک باند، ResNet152، شبکه عصبی عمیق کانولوشنی، انتقال یادگیری، تشخیص احساس از روی گفتار.