JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 9(4), DECEMBER, 2021
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

Speech Emotion Recognition using a Novel Noise Augmentation Approach and Deep Learning

Shehu Mohammed Yusuf*, E. A. Adedokun, M. B. Mu'azu, I. J. Umoh
Department of Computer Engineering,
Ahmadu Bello University, Zaria

ABSTRACT
Speech emotion recognition (SER) based on deep learning methods are practically more suitable than existing conventional machine learning methods for categorizing emotions from speech. However, two limitations of existing speech emotion recognition (SER) frameworks are overfitting on scarce training inputs and poor robustness to environmental noise. This work proposes a novel augmentation scheme based on speech splitting and real environmental noise mix-up augmentation and, transfer learning by finetuning lower layers of pretrained networks to handle these aforementioned limitations. The speech splitting and real noise mix-up augmentation is adopted as a scheme to create additional speech samples required for training. Then, pretrained networks are adapted for speech emotion recognition and finetuned with the generated training datasets to develop a model robust to noisy environment. Thereby, improving the classification performance in the wild. The novelty of this work is in utilizing speech splitting and real noise modulation as an augmentation scheme to increase the number of speech training samples for speech emotion recognition. Also, this augmentation scheme combined with transfer learning makes the SER model more robust to speech degradation due to noise. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database was utilized to generate benchmark datasets. Findings showed that, under three environmental conditions of SNR 15dB, 20dB, and 40dB, the proposed SER framework achieved an improvement of 9%, 6% and 5%, respectively, as compared to the state-of-the-art method. The proposed augmentation scheme for SER using deep learning can be useful in characterizing emotions of speakers conversing virtually on online platforms.

ARTICLE INFO
Article History
Received: October, 2021
Received in revised form: November, 2021
Accepted: November, 2021
Published online: January, 2022

KEYWORDS
Noise augmentation, Speech emotion recognition, deep learning, Transfer learning, and Machine learning

INTRODUCTION

Emotion can be defined as the state of continuous mental activity that is strongly or weakly expressed or perceived by humans during conversations. Expressed or perceived emotions can be interpreted by humans in the form of cues which include sad, happy, boredom, anger, and calm [1]. Several attempts have been made by scientists and engineers to develop machines that predict emotions, just like humans, from various modalities, [2]. However, due to the type of human-computer interfaces required for applications such as online learning and medicine, it is expensive to recognize emotions from modalities such as faces and biophysical signals. Among the various modalities for emotion recognition, speech corpus is an effective and accessible data resource for emotion recognition because of its relatively cheap human-computer interface. But the challenge is how to design a speech emotion recognition system that exploits the rich emotion content from speech. Speech emotion recognition (SER) is the task of detecting the emotional aspect of speech regardless of the

Corresponding author: Shehu, M. Y. shekullahi@gmail.com *Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.*

semantic content, [2]. SER is an important module for Human-Computer interaction, [3, 4].

SER systems can broadly be grouped into the conventional and automatic frameworks. Conventional SER frameworks utilize handcrafted features, to determine the correlation between speakers' emotions. In such scenarios, feature extraction, modeling and classification are disjointly carried out. Several conventional frameworks have been developed for speech emotion recognition using Mel's frequency cepstral coefficients (MFCC), prosodic features and spectral features, [5, 6]. The works of [7] developed a SER system based on MFCC and hidden Markov model (HMM) classifier. In [8], three low-level acoustic features were hybridized and fed as input to a deep neural network to classify emotions from the IEMOCAP dataset split into 80% to 20% for training and testing, respectively. Although promising results were obtained, identifying the best acoustic features that model to characterize different emotions is a challenge. Real-time inference of emotional class labels cannot be achieved because feature calculation is time-consuming. Thus, the need for automatic feature generation procedure. Automatic SER techniques are end-to-end networks that utilize high-level discriminative features to model and classify emotions from speech, [9]. These techniques are deep learning, [10], pipelines that utilize spectrogram images obtained from speech to recognize emotions, [8, 11-13]. Lim *et al.* [14] utilized 2-D spectrograms, generated from short-time Fourier transform (STFT) as inputs to convolutional neural networks (CNN) concatenated with stacked long short-term memory (LSTM) network. Results showed the framework outperformed the CNN-based and LSTM-based SER framework. However, the framework is limited by high computational complexity. The work of [15] showed that utilizing spectrograms as input to a deep recurrent neural network (RNN) with attention mechanism outperformed the traditional support vector machines (SVM) method in predicting four emotional classes, namely sad, neutral, angry and happy emotions, of the IEMOCAP dataset, [16]. However, the model overfitted due to insufficient training samples. Ramet *et al.* [17] showed that an attention-based bidirectional long short-term memory (BLSTM) outperformed the state-of-the-art methods in classifying 32-D handcrafted frame-level emotional features extracted using the openSMILE toolkit, [18]. The work of [19] highlighted that the CNN-based methods are more generalizable in classifying emotion from speech samples in both single-corpus and cross-corpus modeling, as compared to LSTM-based methods. Padi *et al.* [20] reported that multi-window augmentation of 32-D audio features can improve the prediction accuracy deep learning frameworks utilized for SER. However, the model is not robust when applied to noisy samples. Thus, there is the need to leverage on speech augmentation to extract additional training samples to reduce overfitting and improve the performance of deep learning frameworks for SER *in the wild*. Recently, several noise augmentation techniques have been adopted to increase the training samples and enhance SER system robustness to environmental noise; noise *in the wild*. Pappagari *et al*. [21] proposed the COPYPASTE augmentation scheme to concatenate clean utterances with noise from the MUSAN corpus. Tiwari *et al*. [22] adopted multi-conditioning and data augmentation using a generative noise model to enhance the robustness of SER system to noise. This study leverages on the success of noise augmentation to develop a deep learning model, for predicting emotions from speech.

PROPOSED METHOD

Recently, the need for reducing model overfitting on scarce data and improve the prediction performance SER systems have made researchers propose data splitting and data augmentation as methods to generate additional training samples. However, there is still the need to further improve deep learning pipelines by developing new approach to data augmentation and adopting transfer learning to adapt them for real-world applications. In this study, speech splitting and real noise mix-up augmentation of speech samples are proposed as approaches to generate additional speech spectrograms; as depicted in the flowchart of Figure 1.

Corresponding author: Shehu, M. Y. ✉ shekullahi@gmail.com ✉ *Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.*

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 9(4), DECEMBER, 2021
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com

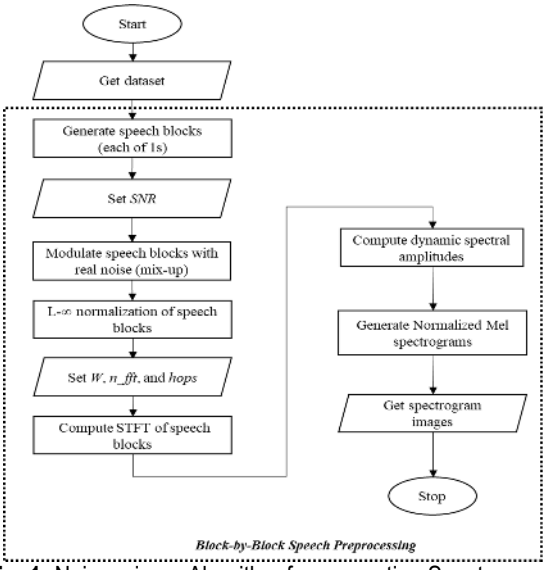


Fig. 1: Noise mix-up Algorithm for generating Spectrograms

From Figure 1, it can be depicted that the original dataset is split into 1s speech blocks the results modulated with real classroom environmental noise using equation 1.

$$\vec{y} = \vec{y}_{clean} + \lambda \vec{y}_{noise} \quad (1)$$

From equation 1, noisy speech block, $\vec{y}$, is generated by adding the clean speech block $\vec{y}_{clean}$, with real environmental noise, $\vec{y}_{noise}$. The factor λ, is set to control the amplitude of the collected environmental noise so that the output is of a desired signal-to-noise ratio (SNR).

Then output of equation 1 is normalised to set the amplitude of these noisy speech blocks in the range of -1 to 1 using equation 2.

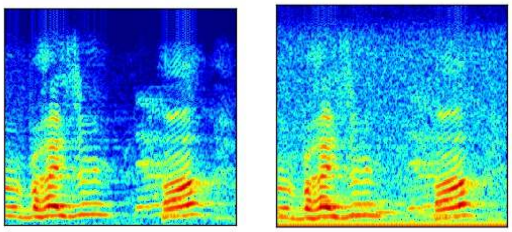
$$\vec{y}_{normalized} = \frac{\vec{y}}{\left\|\vec{y}\right\|_\infty} = \frac{\vec{y}}{\max\left\{\left|y_i\right| : i = 1, 2, \ldots, n\right\}} \quad (2)$$

$\vec{y}$ is a speech block which is normalised to produce another vector, $\vec{y}_{normalized}$, by dividing each of the n noisy speech blocks by the infinity norm the speech blocks, $\left\|\vec{y}\right\|_\infty$. Then, each normalized speech block is short time Fourier transformed (STFT) using Equation 2, [23].

$$X_i(f, m) = \left|STFT\left\{x_i\right\}(f, m)\right|^2 \quad (3)$$

X_i, f and, m is the speech signal, frequency and, window position, respectively.

The next step is to obtain the spectrogram by plotting the power spectrum of the transformed signal as against frequency and time. Figure 2 shows some plots saved as an image.



(a) (b)

Fig. 2: 2-D Visuals of (a) Clean speech spectrogram (b) Noisy speech spectrogram

Corresponding author: Shehu, M. Y. shekullahi@gmail.com *Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.*

As shown in Figure 2, high frequency magnitudes are captured by light colours and vice versa. Also, it can be seen that the noisy speech spectrogram in Figure 2b has distortions introduced in the form of higher frequency components as compared to the spectrogram of the clean speech block. The *Librosa* [24] and *Matplotlib* [25] Python packages were utilized in implementing the noise mix-up augmentation scheme and for generating speech spectrograms.

Subsequently, a transfer learning [26] model is developed for the SER task. Here, the lower convolutional layer and three dense layers of pretrained AlexNet [27] model are finetuned to transfer learned features of previous layers to quickly learn new features for SER task; as presented in Figure 3.

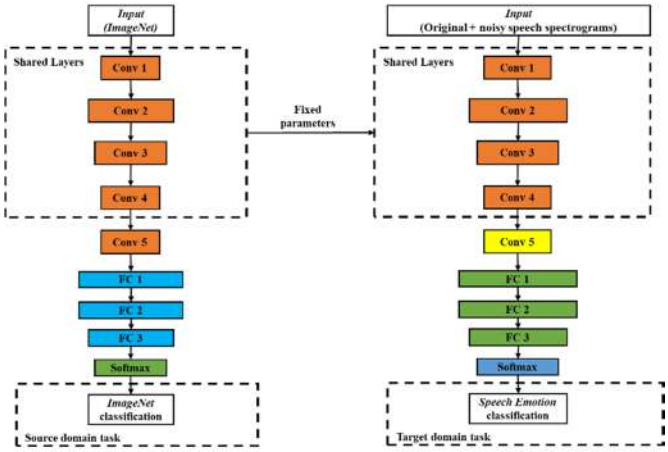


Fig. 3: Proposed Transfer Learning Framework

EXPERIMENTAL SETUP

This work investigated the merits of utilizing real environmental noise as an augmentation scheme and transfer learning for SER by applying them on the popular and publicly available IEMOCAP database. The database contains about 12 hours of speech conversations among 10 actors. The speech conversations contains 5 sessions in which 2 speakers acted scripted and improvised scenarios in each session, [11]. Two datasets were generated from the improvised IEMOCAP scenario namely; the AHSN and the AESN dataset. The AHSN dataset contains the anger, happy, sad and neutral emotional speech samples, [19]. And the AESN dataset contains the anger, excitement, sad and neutral emotional speech samples, [20]. The works of [21] and [22] were chosen as the baseline models for which the inspiration of multi-window data augmentation and deep learning framework were obtained, respectively. Speech samples were broken into 1s block and then converted to spectrograms using the *Librosa* and *Matplotlib* Python packages. The parameters set for the proposed augmentation scheme are as follows (Table 1).

Table 1: Parameter setting for MWA of Spectrogram Algorithm

Parameter	Setting
Window type	Hamming
Window size	16*ms*
SNR	15, 20, 40 dB
n_fft	512
$hops$	62
Sample rate	16000 Hz

Corresponding author: Shehu, M. Y. ✉ shekullahi@gmail.com ✉ *Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.*

The RGB images generated, using the proposed augmentation algorithm with the set parameters, are of dimension 257 x 259 pixels based on the "jet" colourmap, [2]. The summary statistics for the AHSN dataset before and after applying the proposed augmentation scheme is presented in Table 2.

Table 2: Summary statistics of the AHSN dataset

Emotion	No. of Utterances	No. of original spectrograms	Overall generated spectrograms
Anger	292	1,197	4,788
Happy	282	1,022	4,088
Sadness	591	2,616	10,464
Neutral	1,081	3,859	15,436
Total	2,246	8,694	34,776

From above, the total of 34,776 speech spectrograms were generated. Table 3 shows the summary statistics for the AESN dataset.

Table 3: Summary statistics of the AESN dataset

Emotion	No. of Utterances	No. of original spectrograms	Overall generated spectrograms
Anger	292	1,197	4,778
Excitement	664	2,215	8,860
Sadness	591	2,616	10,464
Neutral	1,081	3,859	15,436
Total	2,628	9,887	39,548

As shown in Table 3, a total of 39,548 spectrograms were generated. The 5-fold cross-validation was applied to the training set; made up of 80% of the original spectrograms and the additional noisy speech spectrograms generated. However, four separate test sets were made from the remaining 20% of the original clean spectrograms generated and, 15dB, 20dB, and 40dB noisy spectrograms; respectively.

To achieve fine-tuning of the proposed transfer leaning framework, the channels of spectrograms images were first transformed to a 256 x 256 RGB images. Subsequently, the image channels were normalized using the mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225), respectively. To adapt the pretrained AlexNet for SER, the last dense, FC8, SoftMax and output layers were adjusted to have 4 outputs equivalent to the number of emotional classes to be differentiated. Finally, the last convolutional layer, Conv5, and all the dense layers were finetuned on the labelled emotional training set. Table 4 presents the set values of the tuned hyperparameters.

Table 4: Hyperparameter setup for finetuning

Hyperparameter	Set Value
Optimizer	SGD
Momentum	0.9
Batch size	128
Weight decay	0.0004
Maximum epochs	6
Initial learning rate	0.0004
Weight learning rate	0.01
Bias learning rate	0.01
Gamma	0.2

Corresponding author: Shehu, M. Y. ✉ shekullahi@gmail.com ✉ Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.

The overall deep learning framework was implemented using the Pytorch [28] application programing interface. And NVIDIA Ge Force RTX 2080 Ti GPU was utilized to accelerate the training process. However, testing was carried out on the best model using X64-based Intel (R) Core (TM) i7-7500U CPU processor @ 5.6GHz with installed memory of 12GB on a 64-bit Windows operating system.

The performance of the MWA-SER system was evaluated using accuracy (Acc), average precision ($AvgPr$), average recall ($AvgRc$) and, average f1score ($Avgf_1$) metrics as shown below, [26].

$$Acc = \frac{\#TP+\#TN}{\#TP+\#FP+\#TN+\#FN} \quad (3)$$

$$pr_i = \frac{\#TP}{\#TP+\#FP} \quad (4)$$

$$rc_i = \frac{\#TP}{\#TP+\#FN} \quad (5)$$

$$f1_i = \left\{\frac{2*pr_i*rc_i}{pr_i+rc_i}\right\} \quad (8)$$

$$AvgPr = \frac{1}{n}\cdot\sum_{i=1}^{n} pr_i \quad (6)$$

$$AvgRc = \frac{1}{n}\cdot\sum_{i=1}^{n} rc_i \quad (7)$$

$$Avgf_1 = \frac{1}{n}\cdot\sum_{i=1}^{n} f1_i \quad (8)$$

pr_i and rc_i represents the precision and recall of a recognized emotional term *i*, respectively. n is the overall count of emotional classes. #TP and #TN denotes the overall count of the positive and negative terms of recognized speech emotion which are recognized correctly, respectively. #FP and #FN denotes the overall number of the positive and negative terms of speech which are misclassified, respectively. Since the emotional classes are unbalanced, the weighted average (WA_Q) of each metrics was computed as follows, [2];

$$WA_Q = \frac{Q_1\times\#c_1+\cdots+Q_i\times\#c_i}{\#c_1+\cdots+\#c_i} \quad (9)$$

Q_i denotes the either the precision, recall or f1 score for the i^{th} class and, $\#c_i$ represents its class size.

EXPERIMENTAL RESULTS

The first experiment was conducted to determine how robust is the deep learning structure when trained on clean speech samples and tested on speech samples from the wild (noisy samples). In developing the AHSN model, 6955 and 1739 speech were randomly selected from the AHSN dataset as clean speech training and tests sets, respectively. Furthermore, 3 independent real world test samples with namely; 15dB, 20dB and, 40dB test samples, each of size 1739, were utilized. Also, a total of 7909 and 1978 speech were randomly selected from the AESN dataset as clean speech training and tests sets, respectively. Similarly, 3 independent real world test samples with namely; 15dB, 20dB and, 40dB test samples, each of size 1978, were utilized. During this experiment, only the last convolutional layer, Conv5, and dense layers were finetuned on the original training dataset. The output layer of the original AlexNet framework was adapted to produce 4 classes. Unweighted accuracy, weighted average precision, weighted average recall, and weighted average f1-score utilized as metrics. Table 5 presents the results for training on clean speech datasets.

Table 5: Prediction performance (%) of clean speech SER model

Dataset	Test Set	UAcc	WavgPr	WavgR	Wavgf1
AHSN	Clean speech samples	**65.27**	**61.18**	**65.27**	**61.57**
	15dB noisy samples	55.43	54.56	55.43	49.81
	20dB noisy samples	54.46	55.24	54.46	50.19
	40dB noisy samples	61.70	58.48	61.70	57.74
AESN	Clean speech samples	**59.4**	**58.82**	**59.40**	**58.90**
	15dB noisy samples	48.46	51.64	48.46	45.12
	20dB noisy samples	50.95	52.94	50.95	48.17
	40dB noisy samples	55.85	55.15	55.85	54.24

Note: Bold indicates the best performance.

Corresponding author: Shehu, M. Y. ✉ shekullahi@gmail.com ✉ Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.

From Table 5, can be seen that the developed models performed best on the clean test dataset. However, the models performed poorly in three real world conditions (SNR of 15dB, 20dB, and 40dB). Also, it can be seen that the performances of these models in the real world improved as the SNR of the speech increased. Therefore, the AHSN and AESN models developed from clean speech, as training samples, are not robust to predicting emotions in the wild.

The second experiment was conducted to determine how robust is the proposed deep learning pipeline when trained on both clean speech samples and speech samples *in the wild* and tested on speech samples *in the wild* (noisy samples). To develop the robust SER model from the AHSN dataset, the training sample of size 27,820, were randomly selected from the overall generated AHSN speech spectrogram set. And, to develop the robust SER model from the AESN dataset, the training sample of size 31,636, were randomly selected from the overall generated AHSN speech spectrogram set. Similarly, 4 independent test samples were utilized as in the case of the first experiment. Table 6 presents the test classification performance of the proposed robust model developed by finetuning AlexNet with the AHSN and AESN datasets; respectively.

Table 6: Prediction performance (%) of the proposed robust SER model

Dataset	Test Set	UAcc	WavgPr	WavgR	Wavgf1
AHSN	Clean speech samples	**65.84**	**63.15**	**65.84**	63.66
	15dB noisy samples	63.34	60.41	63.34	61.21
	20dB noisy samples	64.85	61.83	64.85	62.85
	40dB noisy samples	**65.84**	62.81	**65.84**	**63.88**
AESN	Clean speech samples	61.63	61.18	61.63	61.33
	15dB noisy samples	61.38	60.67	61.38	60.77
	20dB noisy samples	61.48	60.82	61.48	60.97
	40dB noisy samples	**62.81**	**62.43**	**62.81**	**62.49**

Note: Bold indicates the best performance.

As shown in Table 6, our proposed SER models outperformed earlier models developed from clean speech datasets (in Table 5) on all metrics. Figure 4 further illustrates the plot of performance in terms of weighted average f1-scores for the AHSN based models.

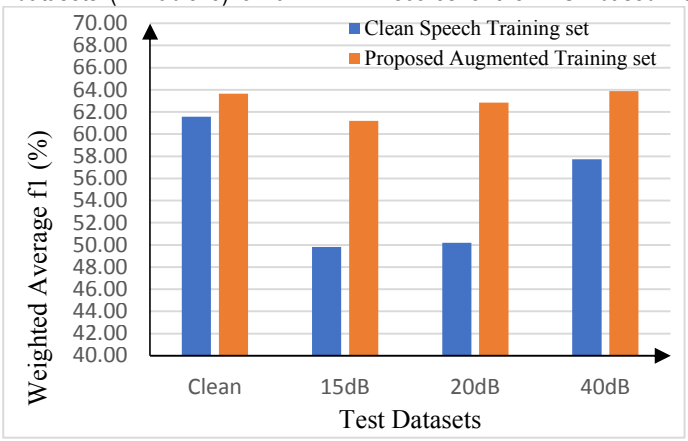


Fig. 4: Clean speech based Vs. proposed augmentation scheme on the AHSN dataset

From Figure 4, it can be seen that propped SER model when applied in the wild at an SNR of 40dB condition outperformed other test conditions; including the model and scenarios of

Corresponding author: Shehu, M. Y. shekullahi@gmail.com *Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.*

the first experiment. Similar observation can be noticed when the AESN dataset was utilized; Figure 5.

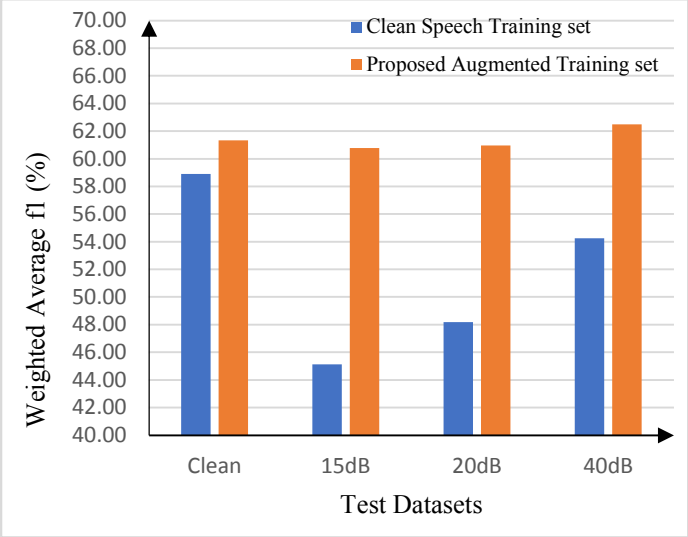


Fig. 5: Clean speech based Vs. proposed augmentation scheme on the AESN dataset

Therefore, the proposed SER framework, based on real-world noise augmentation and transfer learning, is more robust and suitable for emotion detection in the wild than frameworks that do not consider environmental noise.

Figure 6 presents the weighted average f1 performance of the proposed SER system in comparison to state-of-the-art SER techniques on the AHSN dataset.

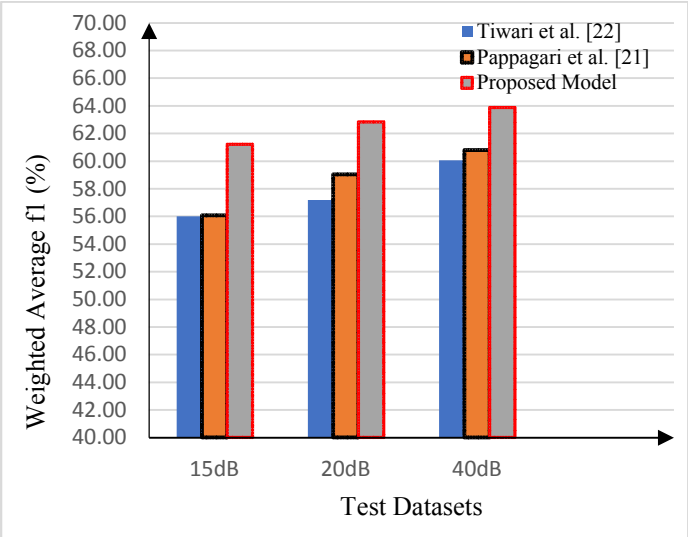


Fig. 6: Comparison between proposed, Tiwari *et al.* [22], and Pappagari *et al* [21] methods

Corresponding author: Shehu, M. Y. shekullahi@gmail.com *Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.*

Observations from Figure 6 shows that the proposed method, when applied on the 15dB test condition, outperformed the state-of the-art method by an improvement of 9% in terms of weighted average f1 score. Also, compared with the state-of-the-art method, a 6% improvement was obtained for the 20dB test condition. Similarly, a 5% improvement was obtained for the proposed method when compared to the state-of-the-art method on the 40dB test condition. Thus, the speech splitting and noise mix-up spectrogram augmentation scheme combined with transfer learning can improve the classification performance of the SER model.

CONCLUSION

This work developed and tested a novel deep learning model based on speech splitting and noise mix-up augmentation scheme for the recognition of emotional cues from speech records. The work highlights the advantage of speech spectrogram augmentation of the original training samples. Findings show that the proposed deep learning-based SER system, which utilizes real noise augmentation of spectrograms, has a better predictive performance than the state-of-the-art deep learning methods that utilize handcrafted features. The proposed SER system achieved the state-of-the-art results by improving the predictive performance by 9%, 6% and 5%, on the 15dB, 20dB, and 40dB AHSN datasets, respectively; in terms of weighted average f1 scores. The proposed SER system generally outperformed models developed from only clean speech samples. The main contribution of our work is the use a novel spectrogram augmentation scheme and transfer learning to develop an SER model robust to degradation due to noise in a classroom. Thus, making the proposed framework have some desirable advantages. First, findings show that the proposed method is more accurate than previous deep learning methods in modelling and predicting emotional cues in classroom environment. Also, the proposed augmentation scheme is another approach for dealing with the problem of data scarcity. However, the proposed SER framework is affected by data imbalance because some speech emotional classes have a lot of speech blocks while others have few numbers of speech blocks. This can cause poor mining of speech emotional features which results in the misclassification of some speech samples.

Consequently, in the future, there is a need to further improve the predictive performance of the proposed SER method by exploring methods of reducing data imbalance. Also, other deep learning methods can be explored with the proposed augmentation scheme to capture rich emotional features. In addition, there is the need to further study the robustness of the proposed model in predicting emotions in real-time and in other real-world scenarios. The developed SER model can be utilized by teachers in predicting the emotions of their students during online classroom interactions.

REFERENCES

[1] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *2015 9th international conference on signal processing and communication systems (ICSPCS)*, 2015, pp. 1-5: IEEE.

[2] M. Lech, M. Stolar, C. Best, and R. Bolia, "Real-Time Speech Emotion Recognition Using a Pre-Trained Image Classification Network: Effects of Bandwidth Reduction and Companding," *Frontiers in Computer Science,* vol. 2, p. 14, 2020.

[3] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion,* vol. 49, pp. 69-78, 2019.

[4] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access,* vol. 7, pp. 117327-117345, 2019.

[5] P. Shen, Z. Changjun, and X. Chen, "Automatic speech emotion recognition using support vector machine," in *Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology*, 2011, vol. 2, pp. 621-625: IEEE.

Corresponding author: Shehu, M. Y. ✉ shekullahi@gmail.com ✉ *Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.*

[6] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech communication,* vol. 53, no. 5, pp. 768-785, 2011.

[7] G. Vyas, M. K. Dutta, K. Riha, and J. Prinosil, "An automatic emotion recognizer using MFCCs and Hidden Markov Models," in *2015 7th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2015, pp. 320-324: IEEE.

[8] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech (Singapore)*, 2014, pp. 1-5.

[9] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," ed, 2020, pp. 1459–1462.

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

[11] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks,* vol. 92, pp. 60-68, 2017.

[12] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*, 2017, pp. 1-5: IEEE.

[13] M. Stolar, M. Lech, R. S. Bolia, and M. Skinner, "Acoustic Characteristics of Emotional Speech Using Spectrogram Image Classification," *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS), Cairns, Australia,* pp. 1-5, 2018.

[14] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016, pp. 1-4: IEEE.

[15] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227-2231: IEEE.

[16] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources evaluation*, vol. 42, no. 4, p. 335. doi: doi: 10.1007/s10579-008-9076-6

[17] G. Ramet, P. N. Garner, M. Baeriswyl, and A. Lazaridis, "Context-aware attention mechanism for speech emotion recognition," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 126-131: IEEE.

[18] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835-838.

[19] J. Parry *et al.*, "Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition," in *INTERSPEECH*, 2019, pp. 1656-1660.

[20] S. Padi, D. Manocha, and R. D. Sriram, "Multi-Window Data Augmentation Approach for Speech Emotion Recognition," *arXiv preprint arXiv:.09895,* 2020.

[21] R. Pappagari, J. Villalba, P. Żelasko, L. Moro-Velazquez, and N. Dehak, "CopyPaste: An Augmentation Method for Speech Emotion Recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6324-6328: IEEE.

[22] U. Tiwari, M. Soni, R. Chakraborty, A. Panda, and S. K. Kopparapu, "Multi-

Corresponding author: Shehu, M. Y. ✉ shekullahi@gmail.com ✉ *Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.*

Conditioning and Data Augmentation Using Generative Noise Model for Speech Emotion Recognition in Noisy Conditions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7194-7198: IEEE.

[23] 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.

[24] P. Raguraman, R. Mohan, and M. Vijayan, "Librosa based assessment tool for music information retrieval systems," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 109-114: IEEE.

[25] N. Ari and M. Ustazhanov, "Matplotlib in python," in *2014 11th International Conference on Electronics, Computer and Computation (ICECCO)*, 2014, pp. 1-6: IEEE.

[26] S. M. Yusuf, F. Zhang, M. Zeng, and M. Li, "DeepPPF: A deep learning framework for predicting protein family," *Neurocomputing,* vol. 428, pp. 19-29, 2021.

[27] P. Sun, W. Feng, R. Han, S. Yan, and Y. Wen, "Optimizing network performance for distributed dnn training on gpu clusters: Imagenet/alexnet training in 1.5 minutes," *arXiv preprint arXiv:.06855,* 2019.

[28] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:.01703,* 2019.

Corresponding author: Shehu, M. Y. ✉ shekullahi@gmail.com ✉ Department of Computer Engineering, Ahmadu Bello University, Zaria, Kaduna State.