ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/354391453

Stride Based Convolutional Neural Network for Speech Emotion Recognition

Conference Paper · August 2021

DOI: 10.1109/ICSIMA50015.2021.9526320

CITATIONS
8

READS
205

6 authors, including:

Taiba Majid Wani
Sapienza University of Rome
12 PUBLICATIONS **203** CITATIONS
SEE PROFILE

Teddy Surya Gunawan
International Islamic University Malaysia
318 PUBLICATIONS **2,895** CITATIONS
SEE PROFILE

Syed Asif Ahmad Qadri
National Tsing Hua University
12 PUBLICATIONS **225** CITATIONS
SEE PROFILE

Fatchul Arifin
Universitas Negeri Yogyakarta
51 PUBLICATIONS **252** CITATIONS
SEE PROFILE

All content following this page was uploaded by Teddy Surya Gunawan on 30 November 2021.

The user has requested enhancement of the downloaded file.

Stride Based Convolutional Neural Network for Speech Emotion Recognition

Taiba Majid Wani
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
wanitaiba1@gmail.com

Teddy Surya Gunawan
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
tsgunawan@iium.edu.my

Syed Asif Ahmad Qadri
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
syed17qadri@gmail.com

Hasmah Mansor
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
hasmahm@iium.edu.my

Fatchul Arifin
Electronics and Informatics Eng. Dept.
Universitas Negeri Yogyakarta
55281 Yogyakarta, Indonesia
fatchul@uny.ac.id

Yasser Asrul Ahmad
Electrical and Computer Engineering Dept.
International Islamic University Malaysia
53100 Kuala Lumpur, Malaysia
yasser@iium.edu.my

***Abstract*—Speech Emotion Recognition (SER) recognizes the emotional features of speech signals regardless of semantic content. Deep Learning techniques have proven superior to conventional techniques for emotion recognition due to advantages such as speed and scalability and infinitely versatile operation. However, since emotions are subjective, there is no universal agreement on evaluating or categorizing them. The main objective of this paper is to design a suitable model of Convolutional Neural Network (CNN) – Stride-based Convolutional Neural Network (SCNN) by taking a smaller number of convolutional layers and eliminate the pooling-layers to increase computational stability. This elimination tends to increase the accuracy and decrease the computational time of the SER system. Instead of pooling layers, deep strides have been used for the necessary dimension reduction. SCNN is trained on spectrograms generated from the speech signals of two different databases, Berlin (Emo-DB) and IITKGP-SEHSC. Four emotions, angry, happy, neutral, and sad, have been considered for the evaluation process, and a validation accuracy of 90.67% and 91.33% is achieved for Emo-DB and IITKGP-SEHSC, respectively. This study provides new benchmarks for both datasets, demonstrating the feasibility and relevance of the presented SER technique.**

Keywords—Speech Emotion Recognition (SER), Stride-based Convolutional Neural Networks (SCNN), Strides, Spectrograms.

I. Introduction

Human-Computer Interaction (HCI) is an attractive area of research due to its wide applications. Recently much work has been reported on Human-Computer Interaction (HCI), where the machine is trained to recognize human utterances or dialogues [1]. Even though much improvement has been reported in HCI, there is still an area where more work, more efforts have to be put in, and this thrust area recognizes emotions by a machine. Emotional expressions provide a depth of information about a person's mental state. Therefore, it has spawned a new study area known as automatic emotion recognition, whose primary purpose is to comprehend and recover desirable emotions.

Emotion categorization has long been a source of contention in psychiatry, affective science, and emotion studies. It is primarily focused on two common approaches: categorical (also known as discrete) and dimensional (also known as continuous) [2]. Emotions are defined using a discrete number of groups in the first approach. Axes define emotions in the second method, which is a synthesis of many psychological aspects. Several modalities have been investigated to distinguish emotional states, including facial expressions, physiological signs, and speech [3]. Speech is an important contact medium that is rich with emotions. The speaker's voice conveys a semantic message and knowledge about their emotional state. Speech signals are a good basis for affective computing because of their intrinsic advantages. Therefore, SER has piqued the attention of most researchers having basic goals to understand and recognize emotions from the speech signal. SER aims to deduce a speaker's emotional state from their speech. As a result, there has been a surge in research interest in the field [4].

On the other hand, recognizing emotions from the speech is difficult since people convey emotions in several ways, and the characteristics that different emotions are also ambiguous. Even humans have difficulty with the paralinguistic dilemma. Various Deep Neural Network-based SER models have been developed in recent years. One group of these models aims to detect significant features directly from raw sound recordings, while the other group uses only one specific representation of a sound file as input to their models, such as in [5]. With neural network architectures including Convolutional Neural Networks (CNNs) and different types of Recurrent Neural Networks (RNNs), Deep Learning (DL) has advanced, outperforming conventional approaches.

CNNs have been around for a while, but they have recently gained popularity for non-traditional uses, surpassing more complex models on several deep learning tasks [6]. DL with CNNs requires minimal preprocessing due to their convolutional layers and their ability to extract salient features and has also allowed networks to learn and extract features directly from raw audio signals. Badshah et al. [7] presented a technique for SER using Convolutional Neural Networks (CNNs) and spectrograms. The model was trained on Berlin Emotional Database (Emo-DB), and an accuracy of 84.3% was achieved. Huang et al. [8] proposed an Audio Word-based Embedding CNN model for emotion recognition. They used vector quantization to convert the low-level features to the audio words classification process. The evaluation process showed that the proposed system achieved 82.34% of accuracy. Qayyum et al. [9] proposed a unique CNN model fed with a raw speech from the SAVEE database to classify emotions. The model achieved 83.61% accuracy. Issa et al. [6] proposed a one-dimensional deep CNN that extracted five different features, Mel-frequency cepstral coefficients (MFCCs), spectral contrast features chromatogram, Mel-scale spectrogram, and Tonnetz representation. The features were

978-1-7281-7523-2/21/$31.00 ©2021 IEEE

extracted from three different databases, Berlin Emo-DB, RAVDESS, and IEMOCAP. The highest accuracy was achieved for Berlin Emo-DB with 95.71%, followed by RAVDESS with 86.1% and IEMOCAP with 64.3%.

In this paper, a Stride-based Convolutional Neural Network (SCNN) is presented. It uses deep strides in pooling layers to reduce feature maps' necessary dimension reduction, thus reducing the network's computational complexity. SCNN consists of 6 convolutional layers, 2 fully connected layers, and a softmax layer. We performed two experiments on two different databases, the Berlin Emotional Database (Emo-DB) and the Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC). SCNN is trained on clean spectrograms of four selected emotions, angry, happy, sad, and neutral, generated from the two databases.

II. Speech Emotion Databases

If the recognition systems are not well trained with an appropriate database, their accuracy and robustness can be easily afflicted. Consequently, good and relevant databases should be carefully selected.Acted/simulated emotions, natural/spontaneous emotions, and elicited emotions are the three primary categories of databases. We have used acted emotion databases in this study because they contain a lot of powerful emotional expressions. According to the literature, most experiments on SER have used emotionally acted speech, according to the literature [10]. The German Database (Berlin Emo-DB) [11] and the Hindi Database (IITKGP-SEHSC) [12] are the two emotional databases used in our study for classifying four discrete emotions, angry, sad, happy, and neutral.

A. Berlin Database of Emotional Speech

Berlin Emo-DB is an acted German speech dataset and is publicly available [11]. The database consists of audio files recorded by 5 males and 5 females. Each person recorded 10 German utterances (5 longer sentences and 5 shorter sentences) with 7 different emotions: anger, fear, happiness, sadness, disgust, boredom, and neutral. A total of 800 utterances were recorded. The sentences were recorded with a sampling rate of 48kHz and downsampled to 16kHz. For the evaluation process concerning the naturalness and recognizability of emotion displayed, 20 listeners were deployed. Only those sentences were further analyzed, which had recognition accuracy of greater than 60%. Therefore, 535 speech samples are present in the database, with 127 angry, 81 bored, 79 neutral, 71 happy, 69 scared, 62 sad, and 46 disgust speech utterances. For this paper, only 4 emotions have been considered: angry, happy, neutral, and sad, each with 60 utterances, which means a total of 240 utterances were taken from Berlin EMO-DB.

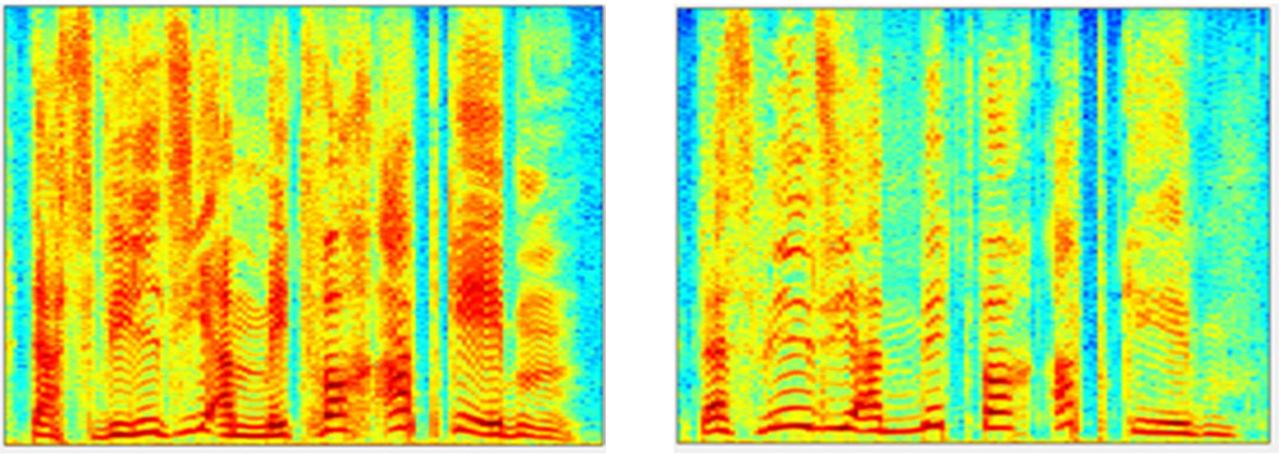
B. Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC)

The Hindi corpus is the Indian Institute of Technology Kharagpur Simulated Emotion Hindi Speech Corpus (IITKGP-SEHSC) [12]. In this corpus, simulated Hindi speech was introduced to analyze the emotions present in the speech signals. The emotions were recorded with the 15 Hindi texts 43 prompts taking into consideration. Every sentence is emotionally neutral in meaning. In one session, every artist must speak 15 different sentences in 8 basic emotions. The number of sessions considered for preparing the database is 10. Each emotion has 1500 utterances. Therefore the total number of utterances in the database is 12000 (15 textprompts×8 emotions×10 speakers×10 sessions). For this paper, only 4 emotions have been considered: angry, happy, neutral, and sad, each with 60 utterances that means a total of 240 utterances, is evaluated.

III. Methodology

SER is performed using time-frequency analysis, i.e., spectrograms. Initially, preprocessing of the audio signal like noise removal, normalization, and echo cancellation was carried, then short-time Fourier transform was applied, and spectrograms were constructed. Spectrograms were obtained from the two datasets mentioned in Section II.

The spectrogram is a basic tool to visualize sound and is often used as training input to CNNs for SER. It can be defined as an intensity plot of the Short-Time Fourier Transform (STFT). The intensity or volume is often represented on a logarithmic scale such as dB and is shown by the color - bright color means high volume and dark means low volume. The STFT is an overlapping sequence of FFTs, intersecting with 25-50%, and is important since the human system of hearing encodes audio data into a kind of spectrogram in the inner ear. The y-axis represents the frequency dimension (pitch), and the x-axis represents the time dimension. Figure 1 (a) and (b) represent the Neutral and Sad spectrogram, respectively.



(a) Neutral

(b) Sad

Fig. 1. Spectrograms of Neutral and Sad Speech Signals

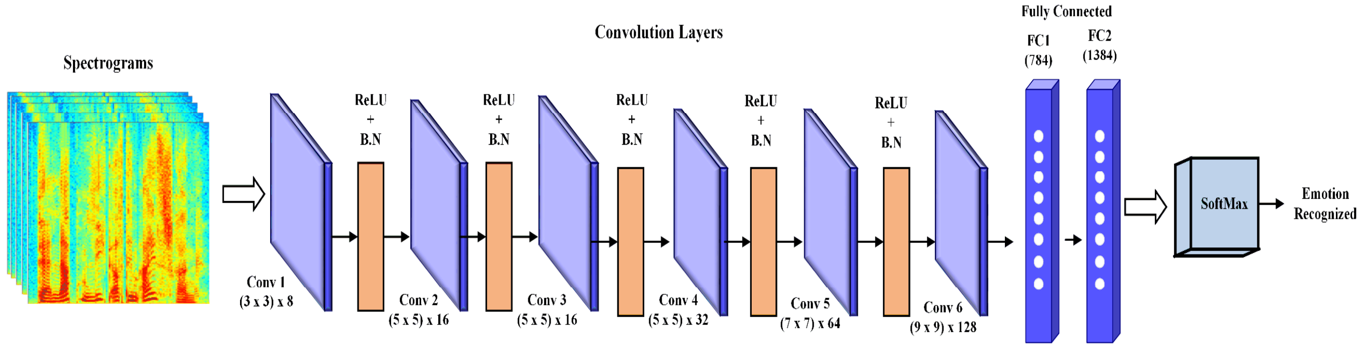


Fig. 2. The Proposed Architecture of Stride-Based Convolution Neural Network

The training method used is Stride-based Convolutional Neural Networks (SCNN). Spectrograms generated from the speech signals are input SCNN. The proposed model of SCNN architecture consists of an input image layer, six convolutional layers, two fully connected layers, and a SoftMax unit. Each layer of convolution layer is followed by a residual unit of ReLU and batch normalization.

Figure 2 represents the proposed architecture of SCNN. The first convolutional layer consists of 8 kernels of size (3×3) stride (2×2) pixels with zero paddings to efficiently modify the data's dimensionality. Kernels of scale (5×5) 16 with stride (2×2) are used in the second and third convolutional layers, respectively. Convolutional layer 4 has a kernel size of (7×7) 32, and layer 5 has a kernel size of (7×7) 64. The final convolutional layer comprises (9×9) 128 kernels with a stride of (2×2) pixels. After the final convolutional layer, there are two connected layers of 784 and 1384 neurons. The second fully connected layer is followed by a 25% dropout ratio to resolve overfitting. Finally, the features are fed into the SoftMax unit's activation function for the classification process.

IV. Results and Discussion

Two experiments were carried out. Firstly, Stride-based Convolutional Neural Networks (SCNN) was trained over the Berlin-Emo database, and the performance was evaluated. Next, SCNN was trained on the IITKGP-SEHEC database.

A. Experimental Setup

The software utilized was MATLAB (2018b). The training was performed on a single i5-8250 CPU with 8GB RAM. The SCNN model was trained and evaluated using 2 cross-fold validation. The network was trained with the feature selection from a feature extraction algorithm (SCNN). Four emotions, anger, happy, neutral, and sad, were considered for the experimentation process. Each is assigned with a label, as shown in Table I. The data were divided into training, validation, and testing. 70% of the dataset was used for training, and 30% was used for validation/testing. A total of 480 spectrograms were generated. The model was tested with 100 epochs and 500 iterations. The classification results have been derived from the confusion matrix and graphs obtained from the Neural Network toolbox.

TABLE I. Emotions Considered and Its Labelling

Emotion	Label
Anger	1
Happiness	2
Neutral	3
Sad	4

B. Experiment on Berlin Emo-DB

The first experiment, Stride-based Convolutional Neural Network (SCNN), was trained on Emo-DB for 100 epochs and 500 iterations. The architecture was run on 100 epochs to allow the algorithm to run until the model error had been sufficiently minimized. Figure 3 shows the Confusion Matrix of the validation data (a) and the total data (b) at 100 epochs. Figure 4 shows the training process for 100 epochs in 1 minute and 55 seconds. The accuracy obtained for training and validation is 95% and 90.67%, respectively.

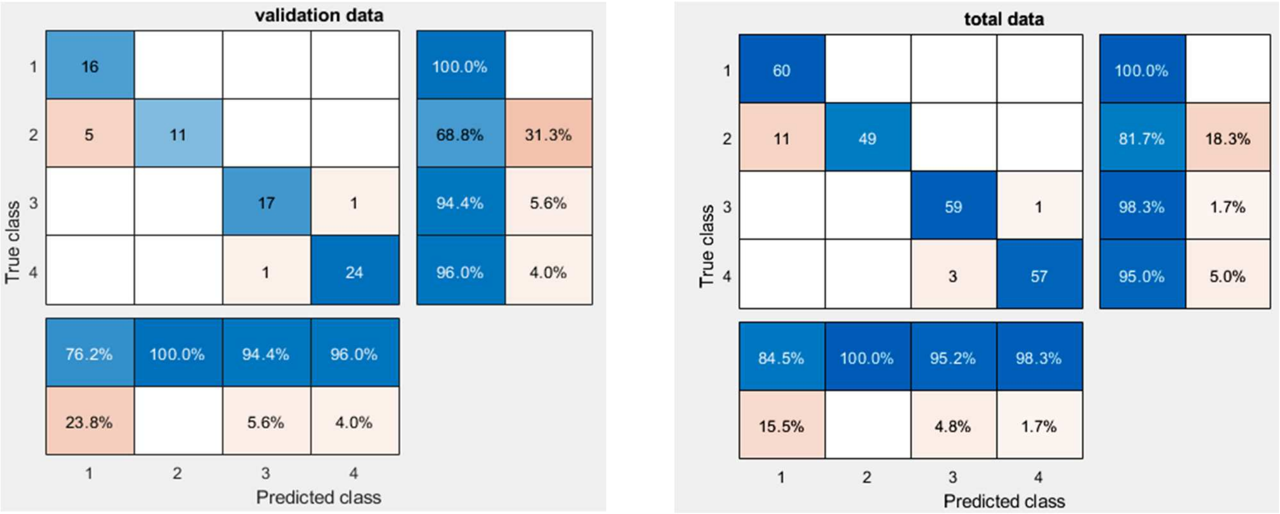
Validation accuracy is less than the training accuracy, and experimentally it is always so. The reason being that the model is already familiar with training data, while the validation data is the collection of new data points of which the model is unfamiliar. Hence, when the model interacts with validation data, accuracy is less than that of training data.

C. Experiments on the IITKGP-SEHSC database

In the second experiment, SCNN was trained on the IITKGP-SEHSC database for 100 epochs and 500 iterations. Figure 5 shows the Confusion Matrix of the validation data and the total data at 100 epochs. Figure 6 shows the training process for 100 epochs in 1 minute and 33 seconds. The accuracy obtained for training and validation is 100% and 91.33%, respectively. The validation and training accuracy of the SCNN architecture for both databases are shown in Table II.

TABLE II. Performance of The Proposed SCNN

Database	Validation Accuracy (%)	Training Accuracy (%)	Computational Time (s)
Berlin Emo-DB	90.67	95.00	1.55
IITKGP-SEHSC	91.33	100.00	1.33



validation data

True class	1	2	3	4		
1	16				100.0%	
2	5	11			68.8%	31.3%
3			17	1	94.4%	5.6%
4			1	24	96.0%	4.0%
	76.2%	100.0%	94.4%	96.0%		
	23.8%		5.6%	4.0%		

Predicted class

(a) Validation Confusion Matrix

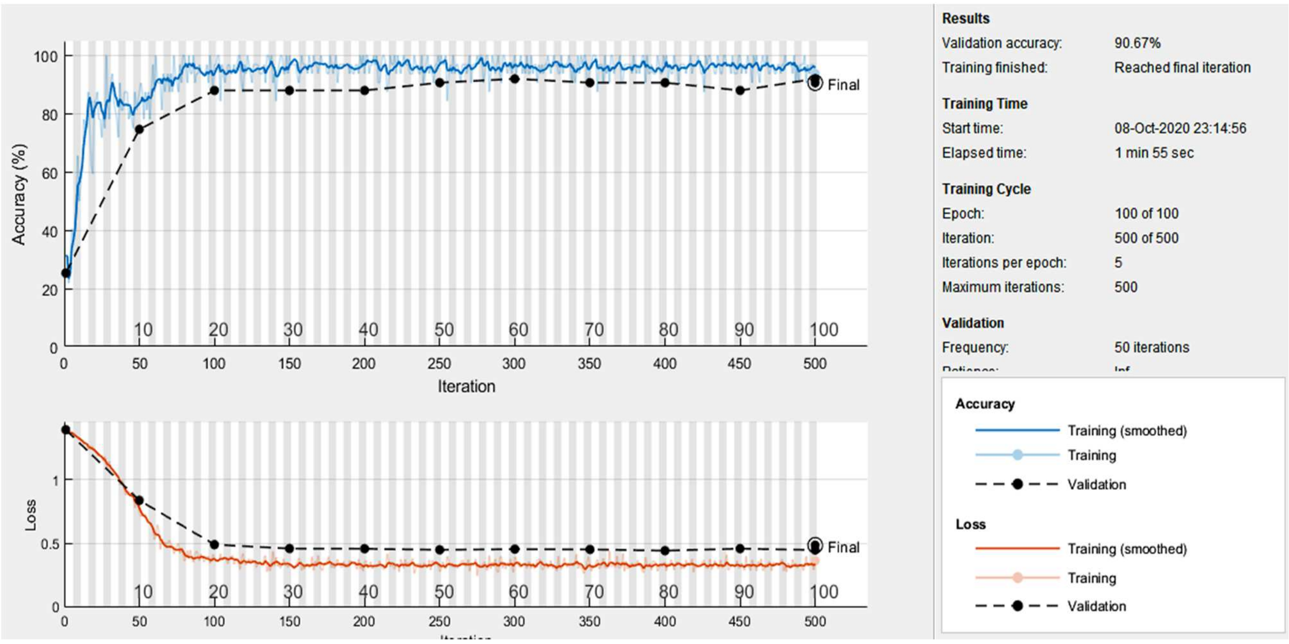
total data

True class	1	2	3	4		
1	60				100.0%	
2	11	49			81.7%	18.3%
3			59	1	98.3%	1.7%
4			3	57	95.0%	5.0%
	84.5%	100.0%	95.2%	98.3%		
	15.5%		4.8%	1.7%		

Predicted class

(b) Overall Confusion Matrix

Fig. 3. Validation and Overall Confusion Matrix for EmoDB



Results

Validation accuracy: 90.67%

Training finished: Reached final iteration

Training Time

Start time: 08-Oct-2020 23:14:56

Elapsed time: 1 min 55 sec

Training Cycle

Epoch: 100 of 100

Iteration: 500 of 500

Iterations per epoch: 5

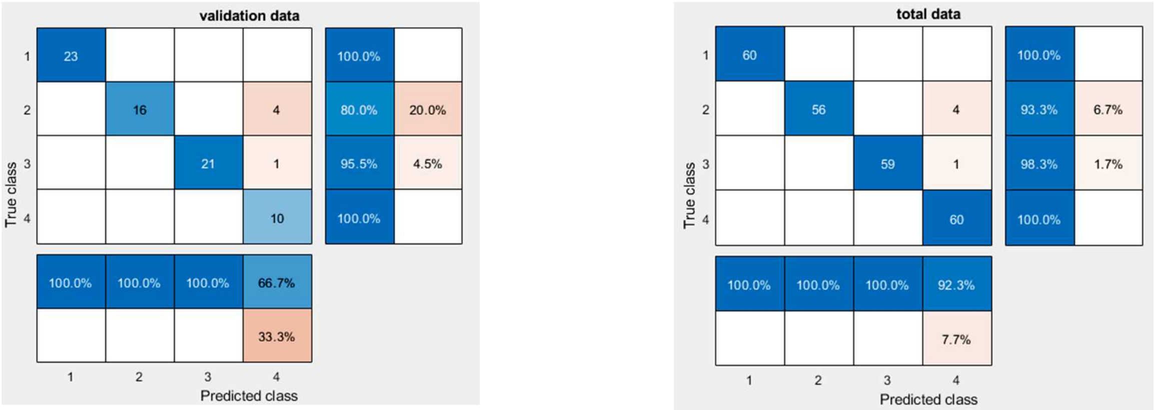
Maximum iterations: 500

Validation

Frequency: 50 iterations

Patience: Inf

Fig. 4. The training process for optimum SCNN architecture for EmoDB at 100 Epochs



validation data

True class	1	2	3	4		
1	23				100.0%	
2		16		4	80.0%	20.0%
3			21	1	95.5%	4.5%
4				10	100.0%	
	100.0%	100.0%	100.0%	66.7%		
				33.3%		

Predicted class

(a) Validation Confusion Matrix

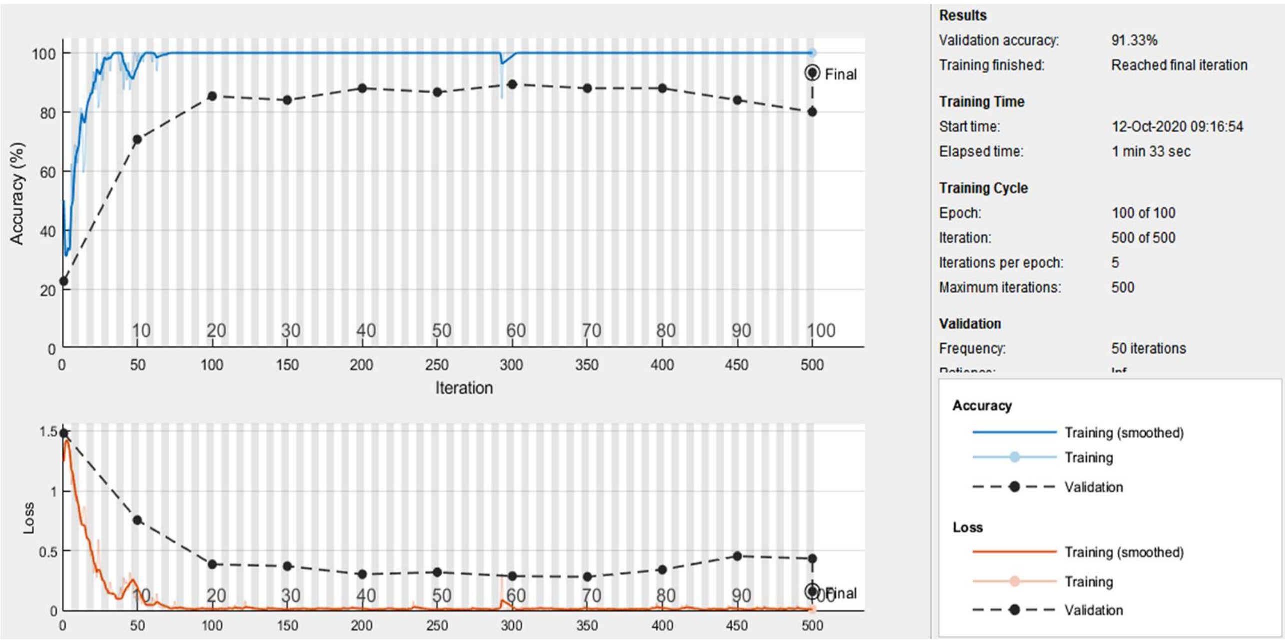
total data

True class	1	2	3	4		
1	60				100.0%	
2		56		4	93.3%	6.7%
3			59	1	98.3%	1.7%
4				60	100.0%	
	100.0%	100.0%	100.0%	92.3%		
				7.7%		

Predicted class

(b) Overall Confusion Matrix

Fig. 5. Validation and Overall Confusion Matrix for IITKGP-SEHSC



Results
Validation accuracy: 91.33%
Training finished: Reached final iteration

Training Time
Start time: 12-Oct-2020 09:16:54
Elapsed time: 1 min 33 sec

Training Cycle
Epoch: 100 of 100
Iteration: 500 of 500
Iterations per epoch: 5
Maximum iterations: 500

Validation
Frequency: 50 iterations
Patience: Inf

Fig. 6. The training process for optimum SCNN architecture for IITKGP-SEHSC at 100 Epochs

D. Discussion

Fundamentally, CNNs consist of convolutional layers, pooling layers, and fully connected layers. This study implemented Stride-based Convolution Neural Networks (SCNNs), a different and more efficient architecture to classify emotions. The model was trained on the spectrograms formulated from the speech signals of two databases, the Berlin EmoDB and IITKGP-SEHSC database. Only four emotions are considered for the research angry, happy, neutral, and sad. SCNN has the same architecture as that of CNN, except that it excludes pooling layers. Instead, 2×2 strides are used for the required reduction in the dimensionality of features. It is done by decreasing the number of pixels of the output of convolutional layers. The network extracts some of the features, which are the most activated pixels, preserves these values, and discards the lower pixel values. Omitting the use of pooling layers decreases the computational load of the architecture, decreases the number of parameters, and hence less computational time is required.

For EmoDB, the validation accuracy achieved by the SCNN model was 90.67% at 100 epochs in 1 minute 55 seconds. The model surpassed the work done in [7], where the proposed model of CNN achieved an overall accuracy of 84.3%. For the IITKGP-SEHSC database, the validation accuracy of 91.33% was obtained using the SCNN model at 100 epochs with 1 minute and 33 seconds of computational time. Till now, CNN has not been implemented on IITKGP-SEHSC in any literature. The presented work is the first to implement CNN on this database. However, other traditional and deep learning classifiers have been evaluated for the IITKGP-SEHSC. Recently in [13], the authors implemented a bagged Support Vector Machine (SVM) for emotion recognition on IITKGP-SEHSC. The highest accuracy obtained was 84.11% which is sufficiently lower than the results obtained in the presented work.

Table III shows the benchmarking with other research in terms of the database used and accuracy achieved. It clearly shows the efficacy of the presented architecture of SCNN. Due to its less complex architecture, it has surpassed both traditional classifications like GMM, SVM, KNN, and Deep learning classifications like RNN and set new benchmarks for both the databases indicating the potential of the proposed system for SER using spectrograms of speech signals.

TABLE III. BENCHMARKING WITH RECENT WORK

Study	Classification	Database	Accuracy
[14]	Concatenated CNN And RNN	EmoDB	89.10
[15]	Parallelized convolutional recurrent neural network (PCRN)	EmoDB	86.44
Present Work	**Proposed SCNN**	**EmoDB**	**90.67**
[16]	LDA, SVM, KNN	IITKGP-SEHSC	90.60
[17]	GMM	IITKGP-SEHSC	75.59
Present Work	**Proposed SCNN**	**IITKGP-SEHSC**	**91.33**

V. Conclusion

The Speech Emotion Recognition (SER) system aims to establish accurate and consistent methods for recognizing emotions. This study presented a deep learning technique, Stride-Convolutional Neural Network (CNN), for effective emotion recognition from speech. Initially, both the considered speech signals of two speech databases, EmoDB and IITKGP-SEHSC, were converted into spectrograms as they are rich in acoustic and semantic features and the quality of high time resolution. A total of 480 spectrograms were formulated about 4 classes, happy, sad, angry, and neutral. Then, these spectrograms were given as an input to SCNN architecture that performed feature extraction and classification process. For EmoDB, an accuracy of 90.67% was achieved, and 91.33% of validation accuracy was obtained for IITKGP-SEHSC. Thus, SCNN can be implemented in real-time speech processing with different spectral features like MFCC, LPCC, LPC, and TEO to classify emotions in future works.

ACKNOWLEDGMENT

This work was supported by the Malaysian Ministry of Education through Fundamental Research Grant, FRGS19-076-0684, under Grant FRGS/1/2018/ICT02/UIAM/02/4. In addition, the authors would like to express their gratitude to International Islamic University Malaysia and Universitas Negeri Yogyakarta, which have provided facilities for this research.

REFERENCES

[1] T. Ozseven, Human-computer interaction, Nova Science Publishers. 2019.

[2] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognit., 2011, doi: 10.1016/j.patcog.2010.09.020.

[3] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "Emotion recognition involving physiological and speech signals: A comprehensive review," in Studies in Systems, Decision and Control, 2018.

[4] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," IEEE Access, 2021, doi: 10.1109/access.2021.3068045.

[5] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp.1-4, 2016, doi: 10.1109/APSIPA.2016.7820699.

[6] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," Biomed. Signal Process. Control, 2020, doi: 10.1016/j.bspc.2020.101894.

[7] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in International Conference on Platform Technology and Service (PlatCon), pp. 1-5, 2017, doi: 10.1109/PlatCon.2017.7883728.

[8] K. Y. Huang, C. H. Wu, Q. B. Hong, M. H. Su, and Y. R. Zeng, "Speech emotion recognition using convolutional neural network with audio word-based embedding," in 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 265-269, 2018, doi: 10.1109/ISCSLP.2018.8706610.

[9] A. Bin Abdul Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition," in IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), pp. 122-125, 2019, doi: 10.1109/SPICSCON48833.2019.9065172.

[10] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: a review," Int. J. Speech Technol., 2018, doi: 10.1007/s10772-018-9491-z.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in Ninth European conference on Speech Communication and Technology, 2005.

[12] S. G. Koolagudi, R. Reddy, J. Yadav, and K. S. Rao, "IITKGP-SEHSC : Hindi speech corpus for emotion analysis," in International Conference on Devices and Communications (ICDeCom), pp. 1-5, 2011, doi: 10.1109/ICDECOM.2011.5738540.

[13] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," Knowledge-Based Syst., 2019, doi: 10.1016/j.knosys.2019.104886.

[14] K. Aghajani and I. Esmaili Paeen Afrakoti, "Speech emotion recognition using scalogram based deep structure," Int. J. Eng. Trans. B Appl., 2020, doi: 10.5829/IJE.2020.33.02B.13.

[15] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized Convolutional Recurrent Neural Network with Spectral Features for Speech Emotion Recognition," IEEE Access, 2019, doi: 10.1109/ACCESS.2019.2927384.

[16] D. Lingampeta and B. Yalamanchili, "Human Emotion Recognition using Acoustic Features with Optimized Feature Selection and Fusion Techniques," in International Conference on Inventive Computation Technologies (ICICT), pp. 221-225, 2020, doi: 10.1109/ICICT48043.2020.9112452.

[17] M. Kumar and J. Yadav, "Speech Emotion Recognition Using Vowel Onset and Offset Points," in Recent Advances in Mathematics, Statistics and Computer Science, pp. 444-454, 2016, doi: 10.1142/9789814704830_0041.

View publication stats