Fractal based speech emotion detection using CNN

Saumya Borwankar, Manmohan Dogra

Abstract In our day to day life, speech is the primary medium of communication between humans and all the interpersonal communication that takes place is emotional. There is often a need to predict the emotion of the intended speech to help understand the emotional and psychological response and state of the person. Now machines are able to automate this task with the help of machine learning and so the task of speech emotion detection has seen many developments. In this paper we have looked at a different feature for the classification of speech emotion and we have analysed the results on 3 publicly available datasets namely Surrey Audio-Visual Expressed Emotion (SAVEE), Toronto emotional speech set (TESS) and Berlin Database of Emotional Speech (Emo-db). The accuracy of the model reaches around 97% which is better than previous approaches.

1 Introduction

Speech consist of 2 types of information, first being the text information and the second being the audio information. For a better human-machine interaction the machine needs to be able to understand the speech data and the intended emotion present in the speech data. In the case of educational assistant machine, it can help improve the academic skills and emotional ability of kids by taking in consideration all the necessary information of the conversation[17]. Teachers and parents can respond to problems in due time. In the case of driving if the car system is able to predict the emotion from the driver, it can help give early warnings if the driver

Saumya Borwankar
Department of Electronics and Communication, Nirma University,Ahmedabad e-mail: 17bec095@nirmauni.ac.in

Manmohan Dogra
Department of Computer Engineering, St. Francis Institute of Technology e-mail: mohanqwerty5@student.sfit.ac.in

sounds too angry or anxious, which in turn can help prevent accidents. Albornoz et al. in [2] has proposed a hierarchical classifier for the identification of a speaker, which has a high recognition accuracy and it doesn't consider the psychological information.

The introduction of machine learning in the domain of speech has solved many problems. Machine learning and deep learning algorithms are able to predict and classify different kinds of data that are fed to the models. One such deep learning method is known as convolutional neural networks. Convolutional neural networks (CNN) have recently gained attention for various audio processing applications, like audio denoising, source separation, music and speech transcription [15][11][21][5], speech enhancement etc. CNNs are also being used widely for the tasks of speech emotion detection [3][19][27]. Also depending on the theory of psychology 6 emotions are present namely sad, fear, happy, angry, surprise and neutral. The face also helps communicate the desired emotion, but as social conditions are concerned it becomes easy to control compared to speech expressions[9].

The most appropriate way of making an effective speech emotion recognition (SER) is the selection of features. Recent models were trained on various features like Mel frequency cepstrum coefficients (MFCC), Spectorgram etc. Susu et al. in [29] proposes a new set of features where they have included certain other features apart from MFCC like RMs energy, zero crossing rate, frequency, voicing probability. And they have used random forest for the classification task. Meaad et al. in [1] has made use of both audio and video data for the emotion detection and classified the emotion with the help of support vector machines (SVM)

The paper is divided into 7 sections. First is the Introduction, Section 2 contains various previous approaches and relevant work done in the field of SER. Section 3 contains information about the fractal dimension. Section 4 describes the various datasets used for evaluation. Followed by Section 5 has the implementation scenario. Section 6 has the results and discussions and lastly Section 7 covers the conclusion and future work.

2 Relevant Work

Many research methods have been proposed for the task of SER, and there are several machine learning approaches that have been coming up in recent years. Rieger et al. in [23] has proposed an approach that emphasizes the SER system based on the classification of spectral features using K-Nearest neighbors (KNN).

Tao et al. in [26] has proposed a work based on ensemble framework that is able to capture various aspects of characteristics related to emotions. The framework that is developed is evaluated on Multi modal emotion challenge (MEC) 2017 dataset. The dataset was gathered from chines TV programs and films, where the scenes are from real world.

Sidorov et al. in [25] has tried to find the most important features with the help of self adaptive multi objective genetic algorithm as a selection method for feature and a neural network classifier. The proposed approach was evaluated on English and German databases which were presented with 37 and 384 features. Sethu et al. [24] has examined the recent approaches and has provided us with a comparative analysis on various methods that are being used.

Various deep learning techniques have been used to build SER systems, Like in [4] Jaybrata et al. has used MFCC features as the feature vector for input to a CNN-LSTM model. They were able to reach an average accuracy of around 80%. Mirsamadi et al. in her paper [20] used a BLSTM combined with an attention mechanism to form their network. Their results showed +3.1% in unweighted accuracy and +5.7% in weighted accuracy. On the other hand the authors in [7] used a 3D CRNN network and tested it on the IEMOCAP and Emo-DB dataset with a constraint of 4 emotions. They were able to reach an accuracy of 82.82% on the Emo-DB dataset and 64.74% accuracy on IEMOCAP dataset.

Fayek et al. in [10] describes the use of Deep Neural Networks for the SER system and they are able to achieve and accuracy of about 59.7% on the SAVEE dataset and about 60.53% on the eNTERFACE dataset. Eric et al. in [12] has come up with a new multi time scale convolution for the SER system and has made analysis using different sub sets of 4 datasets. They were able to reach a highest accuracy score of about 70.97% on the EMODB dataset. Puterka et al. in [22] has proposed a way though which spectrograms features can be applied to improve the function of SER, the authors also presented the speech features with the help of different segments in time domain. Lim et al. in [18] the authors have also worked on RNN and CNN for SER, they have used analysis of time-frequency by converting speech sample to a 2D input and evaluated the model on Emo-DB dataset and got an average accuracy of 92.02%, average F-1 score of 88.56% and an average recall of 88.42%.

Wunarso et al. in [28] extracted the speech duration, amplitude and approximate coefficients from the Indonesian speech database (I-SpeED) which is based on their native language with the help of SVM. Their results showed that they were able to get an average accuracy of 76.84% after their evaluation and analysis.

Zamil et al. in [30] the author has proposed MFCC feature analysis on the speech signal to identify the underlying emotions. The classifier used was Logistic Model Treel (LMT) to classify between different classes of emotion. The author used a voting algorithm on the frames that were classified to detect the emotions. RAVDESS and Emo-DB were the 2 datasets that were used to evaluate their algorithm. The performance of the best emotion was around 70%.

3 Fractals

The fractal conversion of the audio file is done according to the Katz, Higuchi methods for fractional dimension. The fractional dimensions are calculated on Matlab and are concatenated and stored as feature vectors. The Katz method[16] is able to calculate the fractal dimension of a sample data as follows. The average and the sum of euclidean distances are taken between successive points of the sample data. At the same time the maximum distance between any other point of the sample data and the first point are also calculated. So the feactional dimension of the sample becomes

$$D = \frac{\log(L/a)}{\log(d/a)} = \frac{\log(n)}{\log(n) + \log(d/L)'} \tag{1}$$

Where 'D' is the fractional dimension, 'L' is the sum of successive sample, 'a' is the average of the successive sample, 'd' is the maximum distance and 'n' is the ratio of 'L' and 'a'.

Higuchi's method [13] also helps us find the fractional dimension of the sample as it does so in the following way. First sub sets of the sample set(X) are created where N is the sample size.

$$X_k^m = \{X(m+ik)\}_{i=0}^{[(N-m)/k]} \tag{2}$$

$$k \in [1, k_{\max}], m \in [1, k] \tag{3}$$

Next the length of every subset of the sample is calculated with the formula.

$$L_m(k) = \frac{\left(\sum_{i=1}^{[(N-m)/k]} |X(m+ik) - X(m+(i-1)k)|(N-1)/([(N-m)/k]k)\right)}{k} \tag{4}$$

Finally the FD of the sample is calculated from.

$$\langle L(k) \rangle \propto k^{-D} \tag{5}$$

Finally both the fractional dimensions are concatenated to form a single feature vector. Figure 1 shows a sample fractal.

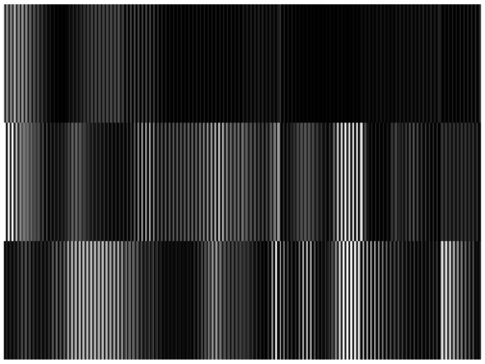


Fig. 1 Fractal Feature

4 Dataset

For the analysis of our proposed approach we have made use of 3 datasets that are available on kaggle and we have evaluated our model on different splitting of the dataset and combining different datasets.

4.1 SAVEE

Surrey Audio-Visual Expressed Emotion (SAVEE)[14] dataset has 4 english speaking males at the University of Surrey aged between 27 to 31 years. The dataset has 6 emotion labels namely surprise, sad, happy, fear, disgust, and anger. It consists of 480 utterances which were annotated. The dataset is publicly available. Every emotion label has 60 audio files each.

4.2 TESS

Toronto Emotional Speech Set[8] consist of 2 english speakers uttering around 2800 sentences which comes to a total of 1 hour 36 minutes. The dataset consist of 7 emotion labels namely sad, happy, disgusted, angry, surprise, neutral and fear. The augmentation of the data is done and the audio samples are taken at different intensity like 0dB, 5dB, 10dB, 15dB, 20dB, 25dB, 30dB, 100dB are taken after which the fractals of each are calculated. The dataset is publicly available. Every emotion label has around 3600 audio files after augmentation.

4.3 EMO-DB

German emotional corpus EMO-DB[6] consist of 10 sentences that cover a total of 7 emotion classes namely fear, anger, happy, disgust, sad, neutral and boredom. The dataset was gathered by the institute of communication science at technical university berlin. It is a very common dataset in the task of speech emotion detection. This dataset is also publicly available. Here angry emotion class has 127 audio files, boredom has 81 files, disgust has 46 files, fear has 69 files, happy has 71, neutral has 79, and sad has 62 audio files.

5 Implementation

The dataset was found to be imbalanced so as to remove this problem only 5 emotion labels were considered namely happy, sad, angry, neutral, and fear. To create a robust model the datasets were combined as to provide enough data for the CNN to actually outperform previous models. The results and various other experimentation details can be found in the next section.

After the feature set is made and stored the feature vector is loaded and split into training, testing and validation data. The split size is taken as 70%, 30% and 10% respectively. For the task of classification the use of various state-of-the art models were used like Resnet, AlexNet, VGG16, but the model that was able to perform best out of the three was ResNet. The Resnet architecture that was used consisted of residual blocks. The residual block help remove the vanishing gradient problem. The vanishing gradient problem was solved by the ResNet which is caused when the backpropagation takes place and the multiplication of a number between 0 and 1 causes the initial layers to disappear. Identity shortcut connection skips the training of one or more layers and creates a residual block. The identity mapping that ResNet introduces is just present to add the output from the previous layer to the next layer. This helps the architecture train very deep networks which was not possible before. The architecture outperforms ALexNet, VGG-16, Xception as the results with ResNet was consistent and most accurate with a specific database.

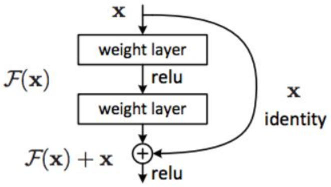
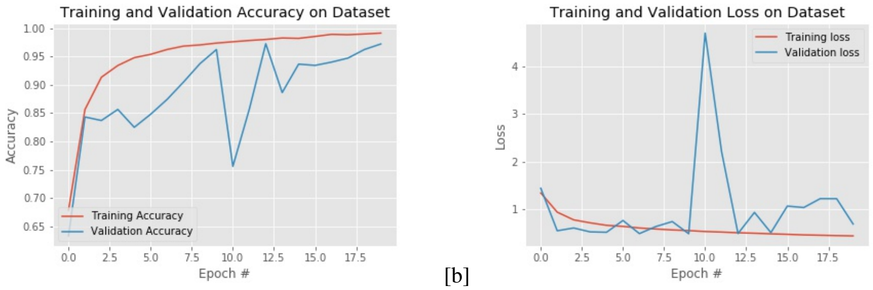

x
weight layer
relu
weight layer
$\mathcal{F}(\mathbf{x})$
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Fig. 2 Residual Block[22]

The Resnet architecture that is used is a compact version of the original ResNet-50. Each residual block in our proposed approach has a set of 3 convolutional layers. The first convolution layer has 64 filters followed by a stack of 3 residual blocks, with every layer in the residual block having 32, 64, and 128 filters in that order, with dimension reduction applied. After which 4 other sets of residual blocks, where the 3 convolution layers will have 64,64 and 256 filters respectively, with dimension reduction applied. Finally 6 sets of residual blocks where each convolution layer will have 128,128, and 512 filters, with dimension reduction applied for the last time. Then an average pooling layer is applied with the last layer being a softmax function layer. The optimizer that is used is the stochastic gradient descent with a learning rate of 0.1 and momentum at 0.9. At the time of compilation of the model the loss function is set to categorical cross entropy as there are 5 emotion classes for classification, and the system parameters are updated after each epoch. The model has been trained for 20 epochs, with a batch size of 64.

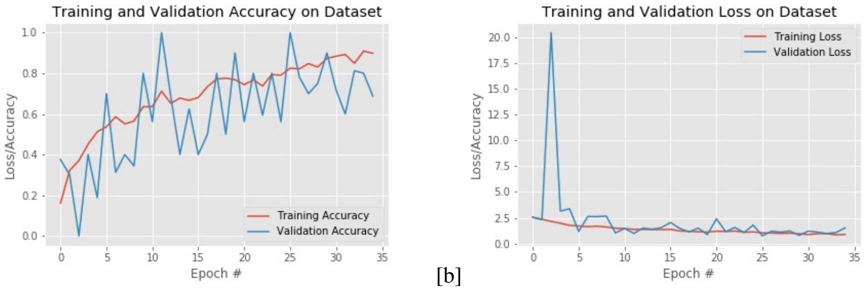
6 Results

So as the aim of the paper is to evaluate the new method for feature selection the classification was done on every dataset individually and then taken 2 at a time and lastly for the creation of a robust SER system all the datasets are combined and fed to the resnet for the classification. The results of the approach are shown in Table 1. the evaluation paramets such as accuracy, recall, and F1 are taken into cosideration.



[a] [b]

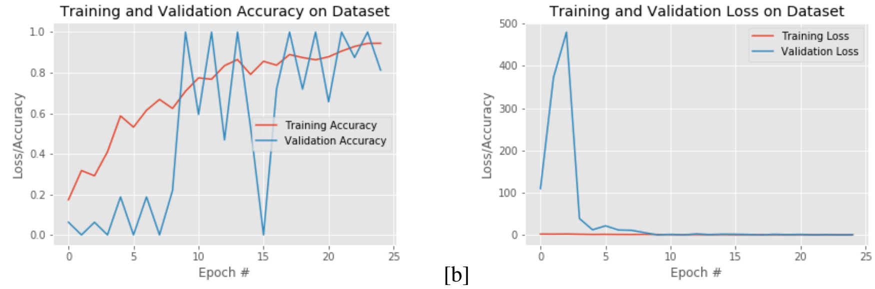
Fig. 3 Training and Validation (a) Accuracy and (b) Loss on All dataset combined



[a] [b]

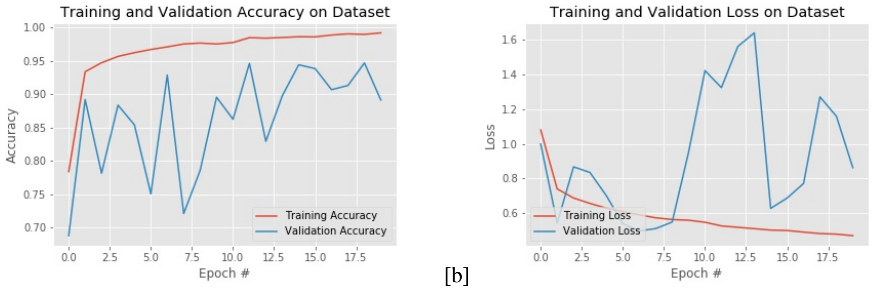
Fig. 4 Training and Validation (a) Accuracy and (b) Loss on EMODB dataset

As we can see from Table 1 the model was able to perform better when it was given more training data. The model was able to reach 97% accuracy which is considered good as compared to previous approaches.



[a] [b]

Fig. 5 Training and Validation (a) Accuracy and (b) Loss on SAVEE dataset



[a] [b]

Fig. 6 Training and Validation (a) Accuracy and (b) Loss on TESS dataset

Dataset	Classes	Accuracy(%)	F1	Recall
EMODB	7	78%	0.76	0.80
SAVEE	7	77%	0.78	0.78
TESS	5	88%	0.88	0.87
SAVEE+TESS+EMO	5	**97%**	**0.97**	**0.97**

Table 1 Evaluation parameters

Method	Dataset	Accuracy	F1	Recall
[12] Multi-time Scale Convolution	TESS	53.05%	-	-
[12] Multi-time Scale Convolution	EMODB	70.97%	-	-
[10] Deep neural networks	SAVEE	59.7%	-	-
[18] Recurrent Neural Network and CNN	EMODB	92.02%	0.8856	0.8842

Table 2 Comparison of different methods

Table 2 tells us how different methods perform on different datasets. The results clearly show that the use of fractals as a feature for the task of SER can help achieve nice and robust results.

7 Conclusion

In previous approaches, various features were used for the task of speech emotion detection. In this paper we proposed a new feature vector, the Fractal feature vector and the evaluation on three datasets SAVEE, TESS, EMODB were done and the performance of this feature outperforms previous features. Furthermore the classification of the fractal features can be improved in the future work and other classification architectures can be looked at. The analysis showed that the fractal features are effective in recognizing emotions from speech data. The proposed approach was able to achieve 97% accuracy when evaluated on all the 3 datasets together.

References

1. Abdul-Hadi, M.H., Waleed, J.: Human speech and facial emotion recognition technique using svm. In: 2020 International Conference on Computer Science and Software Engineering (CSASE), pp. 191–196. IEEE (2020)
2. Albornoz, E.M., Milone, D.H., Rufiner, H.L.: Spoken emotion recognition using hierarchical classifiers. Computer Speech & Language **25**(3), 556–570 (2011)
3. Badshah, A.M., Ahmad, J., Rahim, N., Baik, S.W.: Speech emotion recognition from spectrograms with deep convolutional neural network. In: 2017 international conference on platform technology and service (PlatCon), pp. 1–5. IEEE (2017)
4. Basu, S., Chakraborty, J., Aftabuddin, M.: Emotion recognition from speech using convolutional neural network with recurrent neural network architecture. In: 2017 2nd International Conference on Communication and Electronics Systems (ICCES), pp. 333–336. IEEE (2017)
5. Bittner, R.M., McFee, B., Salamon, J., Li, P., Bello, J.P.: Deep salience representations for f0 estimation in polyphonic music. In: ISMIR, pp. 63–70 (2017)
6. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B.: A database of german emotional speech. In: Ninth European Conference on Speech Communication and Technology (2005)
7. Chen, M., He, X., Yang, J., Zhang, H.: 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters **25**(10), 1440–1444 (2018)
8. Dupuis, K., Pichora-Fuller, M.K.: Toronto emotional speech set (TESS). University of Toronto, Psychology Department (2010)
9. Emerich, S., Lupu, E., Apatean, A.: Emotions recognition by speechand facial expressions analysis. In: 2009 17th European Signal Processing Conference, pp. 1617–1621. IEEE (2009)
10. Fayek, H.M., Lech, M., Cavedon, L.: Towards real-time speech emotion recognition using deep neural networks. In: 2015 9th international conference on signal processing and communication systems (ICSPCS), pp. 1–5. IEEE (2015)
11. Fu, S.W., Tsao, Y., Lu, X.: Snr-aware convolutional neural network modeling for speech enhancement. In: Interspeech, pp. 3768–3772 (2016)
12. Guizzo, E., Weyde, T., Leveson, J.B.: Multi-time-scale convolution for emotion recognition from speech audio signals. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6489–6493. IEEE (2020)
13. Higuchi, T.: Approach to an irregular time series on the basis of the fractal theory. Physica D: Nonlinear Phenomena **31**(2), 277–283 (1988)
14. Jackson, P., Haq, S.: Surrey audio-visual expressed emotion (savee) database. University of Surrey: Guildford, UK (2014)
15. Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., Weyde, T.: Singing voice separation with deep u-net convolutional networks (2017)

16. Katz, M.J.: Fractals and the analysis of waveforms. Computers in biology and medicine **18**(3), 145–156 (1988)
17. Khan, Z.A., Sohn, W.: Abnormal human activity recognition system based on r-transform and kernel discriminant technique for elderly home care. IEEE Transactions on Consumer Electronics **57**(4), 1843–1850 (2011)
18. Lim, W., Jang, D., Lee, T.: Speech emotion recognition using convolutional and recurrent neural networks. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–4. IEEE (2016)
19. Mao, Q., Dong, M., Huang, Z., Zhan, Y.: Learning salient features for speech emotion recognition using convolutional neural networks. IEEE transactions on multimedia **16**(8), 2203–2213 (2014)
20. Mirsamadi, S., Barsoum, E., Zhang, C.: Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2227–2231. IEEE (2017)
21. Palaz, D., Collobert, R., et al.: Analysis of cnn-based speech recognition system using raw speech as input. Tech. rep., Idiap (2015)
22. Puterka, B., Kacur, J.: Time window analysis for automatic speech emotion recognition. In: 2018 International Symposium ELMAR, pp. 143–146. IEEE (2018)
23. Rieger, S.A., Muraleedharan, R., Ramachandran, R.P.: Speech based emotion recognition using spectral feature extraction and an ensemble of knn classifiers. In: The 9th International Symposium on Chinese Spoken Language Processing, pp. 589–593. IEEE (2014)
24. Sethu, V., Epps, J., Ambikairajah, E.: Speech based emotion recognition. In: Speech and Audio Processing for Coding, Enhancement and Recognition, pp. 197–228. Springer (2015)
25. Sidorov, M., Brester, C., Minker, W., Semenkin, E.: Speech-based emotion recognition: Feature selection by self-adaptive multi-criteria genetic algorithm. In: LREC, pp. 3481–3485 (2014)
26. Tao, F., Liu, G., Zhao, Q.: An ensemble framework of voice-based emotion recognition system. In: 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp. 1–6. IEEE (2018)
27. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: 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), pp. 5200–5204. IEEE (2016)
28. Wunarso, N.B., Soelistio, Y.E.: Towards indonesian speech-emotion automatic recognition (i-spear). In: 2017 4th International Conference on New Media Studies (CONMEDIA), pp. 98–101. IEEE (2017)
29. Yan, S., Ye, L., Han, S., Han, T., Li, Y., Alasaarela, E.: Speech interactive emotion recognition system based on random forest. In: 2020 International Wireless Communications and Mobile Computing (IWCMC), pp. 1458–1462. IEEE (2020)
30. Zamil, A.A.A., Hasan, S., Baki, S.M.J., Adam, J.M., Zaman, I.: Emotion detection from speech signals using voting mechanism on classified frames. In: 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), pp. 281–285. IEEE (2019)