



# Speech emotion recognition using data augmentation

V. M. Praseetha<sup>1</sup> · P. P. Joby<sup>1</sup>

Received: 28 January 2021 / Accepted: 8 August 2021 / Published online: 18 August 2021  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

## Abstract

Humans are considered as emotional beings and so the uttered speech reflect the human emotions. Human computer interaction can be done more effectively by automatically identifying the emotions from speech. Automatic speech emotion recognition is applied in many areas like computer gaming, call centre, speech therapy controlling robots etc. Emotion recognition can be considered as feature space to label space mapping. From the uttered speech, the different features are calculated. Then, to automatically recognize the emotions, the relationship between the emotions and the features are learned. The required preprocessing is done with the collected training samples and the features are extracted from the speech signals. The extracted feature vectors are stored in the database. When the input signal comes, the preprocessing and feature extraction are done and the extracted features are compared with the feature vectors in the database to determine the emotion in that speech signal. We have developed a deep learning model for speech emotion recognition with GRU which take the filterbank energies of the speech signals as input. To overcome the problem with the availability of database and to increase the number of input samples, we have applied data augmentation.

**Keywords** Human computer interaction · Speech emotion recognition · Biometric

## 1 Introduction

Speech or voice of a person is a behavioural biometric trait which can be used for many applications like speech recognition, speaker recognition, and voice to text conversion etc. The voice of a person is a signal which carries complex information and varies according to the size and shape of the vocal tract of a person. Speaker recognition is a very popular application which can be used for authentication based on the voice signals. This paper deals with another important application which is the automatic recognition of emotions from speech. Lot of machine learning approaches have been developed for automatic recognition of emotions from speech. The machine learning models can be trained with smaller data sets. But the problem is that there is a chance to lose the key features and it is not easy to make sure the selected features' quality. This may lead to less accuracy

for the models. In such situations we can apply deep learning where we can train our model with huge amount of data. Then the high level characteristics can be extracted and so good performance is assured in many complex tasks.

There are some issues which complicate the process of automatic speech emotion recognition (Chernykh & Prikhodko, 2017). They include:

- Emotions are mainly subjective: It is very difficult to define the notion of emotions (Devillers et al., 2005) as it is a complex social and psychological phenomena.
- Availability of database: The emotional databases are not easy to collect and we need to use artificial databases. So the database collection is expensive, complex and time consuming.

Automatic speech processing has gained more interest by the introduction of deep learning. Lot of studies have been done on various speech processing areas like speech recognition, speaker recognition etc. The emotional expressions of people may vary and the same emotion expressed by different people are different. This makes emotion recognition a difficult problem. Automatic speech emotion recognition

---

✉ V. M. Praseetha  
mpraseethav@gmail.com

P. P. Joby  
jobymone@gmail.com

<sup>1</sup> Department of Computer Science and Engineering, St. Joseph's College of Engineering & Technology, Palai, India

has a good scope for research since it can be applied in many areas of human machine interaction.

## 2 Related works

Most of the experiments use traditional machine learning techniques for recognizing the emotions from speech. They use classifiers like Gaussian Mixture Model (GMM) (Neiberg et al., 2006), Hidden Markov Model (HMM) (Mao et al., 2009; Ntalampiras & Fakotakis, 2012; Nwe et al., 2003; Schuller et al., 2003), support vector machine (SVM) (Hu et al., 2007; Zhou et al., 2006), artificial neural network (ANN) (Bhatti et al., 2004; Cowie et al., 2001; Nicholson et al., 2000; Wang et al., 2010) as well as the combination of any of these classifiers (Schuller et al., 2005; Wu & Liang, 2011). But deep learning approaches are the state-of-the-art techniques for this kind of application. Some of the researches conducted in this area using deep learning are briefed here.

A deep belief network (DBN) model with nonlinear features is explained in Kim et al. (2013) which gives an accuracy of 60% to 70%. A DBN-HMM model is proposed in Zheng et al. (2014) which in turn improves the performance. A robust and stable model for emotion recognition is explained in Mao et al. (2014) which uses a Convolutional Neural Network (CNN) classifier for recognition. A semi-CNN classifier, with an accuracy of 78% and 84% on SAVEE and Emo-DB databases respectively, is proposed in Huang et al. (2014). In Han et al., (2014), a DNN-Extreme Learning Machine (ELM) model is proposed by the authors which improves the accuracy of speech emotion recognition.

A neural network with spectrogram information is applied on Emo-DB database in Prasomphan (2015) to classify five emotions with an accuracy of 83:28%. An accuracy of 40% is obtained when Deep Convolutional Neural Network (DCNN) with spectrogram information is used (Zheng et al., 2015) on IEMOCAP database. RNN is applied on IEMOCAP database in Lee and Tashev (2015) to recognize the emotions and the accuracy obtained is 62%. The method used in Haoxiang (2020) uses the acoustic as well as lexical features for SER on IEMOCAP database. This method obtained an accuracy of 69:2%. Wang et al. proposed (Wang & Tashev, 2017) a DNN model for emotion recognition from speech. Each utterance is encoded into a vector of fixed length. The utterance level classification is done with a kernel extreme learning machine.

Eyben et al. (2009) and Stuhlsatz et al. (2011) in their works extracted the low level features from each frame to calculate the utterance level statistics. The low level features include pitch, MFCC, zero-crossing rate, energy,

voice probability etc. A deep neural network model is explained by Han et al. (2014) and a model using auto-encoder is explained by Ghosh et al. (2015) to learn the frame-level features for computing the utterance level statistics. A method which combines the CNN and LSTM is explained in Trigeorgis et al. (2016) which learns the best features from the representation of speech signals. Recurrent neural network (RNN) model with long-short term memory (LSTM) is used by Keren et al. (2016) for feature extraction from sequential data. In this method the average of the frame level prediction is used for finding the final prediction. The utterance-level label remains same for every frame for the LSTM training. This may lead to biasing towards majority classes. RNN with bidirectional long-short term memory (BLSTM) model is used by Lee et al. (2014) to extract the high level features of the emotional state. To overcome the problem of biasing, a sequence of random variables which are nothing but the label of each frame are trained. An evaluation of the performance of Gated Recurrent Unit (GRU) model on music modelling is explained by Junyoung et al. (2019). Among the two RNN models, GRU models are found simple in calculation and suitable for handling sequential data such as speech.

The performance of the existing state-of-the-art speech emotion recognition system is quite low. The main reason for low accuracy is the unavailability of the data. So we need a good model with high accuracy which can be used for real time applications.

### 2.1 Goal and contributions

Our goal is to develop a deep learning model which can recognize human emotions from speech in an effective manner. Towards this goal we have developed a GRU model which use filter bank energies. We have tested the model with the original database as well as the augmented database. Our GRU model give good results with augmented database.

Our contributions are given below:

- Recently proposed GRU network is used for implementing emotion recognition from speech.
- Dataset augmentation is done for increasing the size of the dataset.

We have used the Toronto Emotional Speech Set (TESS) which contains 2800 stimuli in total expressing the emotions like anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral. Our model is trained to recognize only five different emotions which are anger, fear, happiness, sadness and neutral. The proposed model gives more accuracy compared with the existing models.

## 2.2 Implementation

Deep learning models can perform well with large amount of input data. We have used the TESS (Toronto emotional speech set) database which contains only 2800 utterances. Data augmentation is a process by which we can create more and more datasets from the existing dataset by applying some transformation techniques without affecting class labels. In our experiment the speech signals are augmented by transformations like stretching and embedding noise. These transformations are done by varying the required parameters and thus new signals of the same class are generated. Thus data augmentation can increase the size of the training set. The increased size of the training set will help to fight against over-fitting of deep learning models and also, the model will be more generalized.

## 2.3 Feature extraction

Speech is a sequence of sounds. The shape and size of the vocal cavity determines the frequency or property of the voice that comes out of it. Figure 1 shows a sample speech signal with emotion ‘angry’.

Noise removal is an important pre-processing step when we deal with speech signals. The speech signal  $x(n)$  can be represented as  $x(n) = s(n) + d(n)$  where  $s(n)$  is the clean signal and  $d(n)$  is the background or ambient noise. To remove the noise and to get the clean signal the spectral subtraction procedure is done. The input signal is segmented and the speech and non-speech activity spectrums are computed using the FFT (Fast Fourier Transform). Voice activity detection is done on the segmented data by which the speech and non speech regions are classified. Noise spectrum estimation is done for the non speech data and spectral subtraction is done for the speech data.

The energy of the speech signal is increased by passing the signal through a filter. This is the pre-emphasis stage which gives the energized signal with more information. The

properties of speech vary with respect to time. So we consider small segments of speech known as frames assuming that the signal properties remain statistically unchanged for short time scales. Normally 20–40 ms time scale is used for framing. The characteristics or parameters are then extracted from each frame. If the frame size is too small it will be difficult to get enough samples for the spectral estimate and if the frame size is too long then the signal will be changing too much within the frame.

A windowing function is used after framing to reduce the data loss or discontinuities at the frame boundaries. The frame is shifted according to the window size such that some overlapping occur across frames. For example, if the frame size is 25 ms and the window size is 10 ms, in every 10 ms the properties of the speech are extracted and for a 1 s speech we will get 100 frames and speech vectors. The frames are then converted from time domain to frequency domain and the frequencies which are present in the frames are calculated by using FFT (Fast Fourier Transform). Thus the frequency spectrum of each frame is generated. Then the power spectrum which is also known as the periodogram is computed. The detailed diagrams of Filter bank energy calculation and spectrogram generation are shown in Figs. 2 and 3.

A set of triangular Mel filters which are equally spaced along the Mel-scale are used to estimate the energy that appears in various frequency regions. A Mel filter bank contains 20–30 Mel scale triangular filters. We have used 26 filters in the Mel filter bank. The normal frequency  $f$  can be converted to Mel scale  $m$  and vice versa by using the following equations.

$$m = 2595 \log_{10}(1 + f/700)$$

$$f = 700(10^{m/2595} - 1)$$

The GRU model use filter bank energies from 26 filters as features.

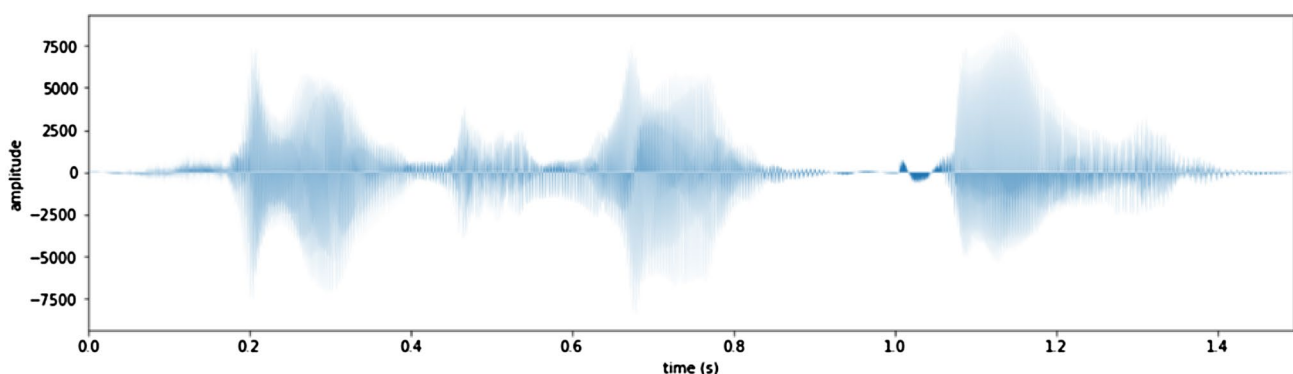


Fig. 1 Speech signal with emotion ‘angry’

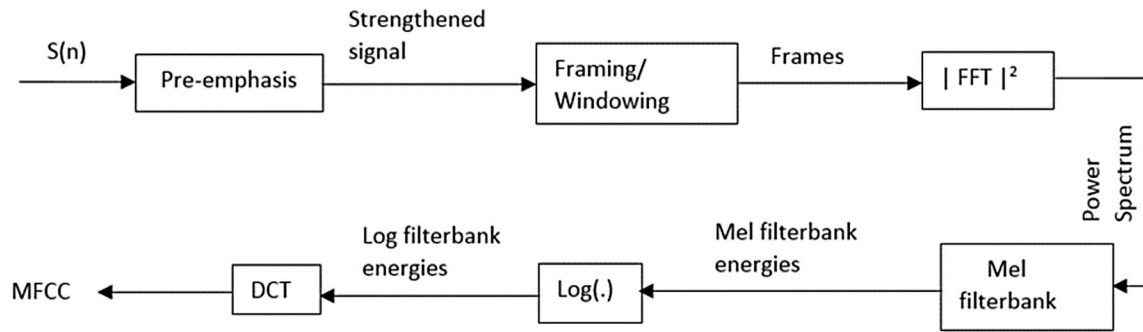


Fig. 2 Different stages in extraction of Filterbank energies

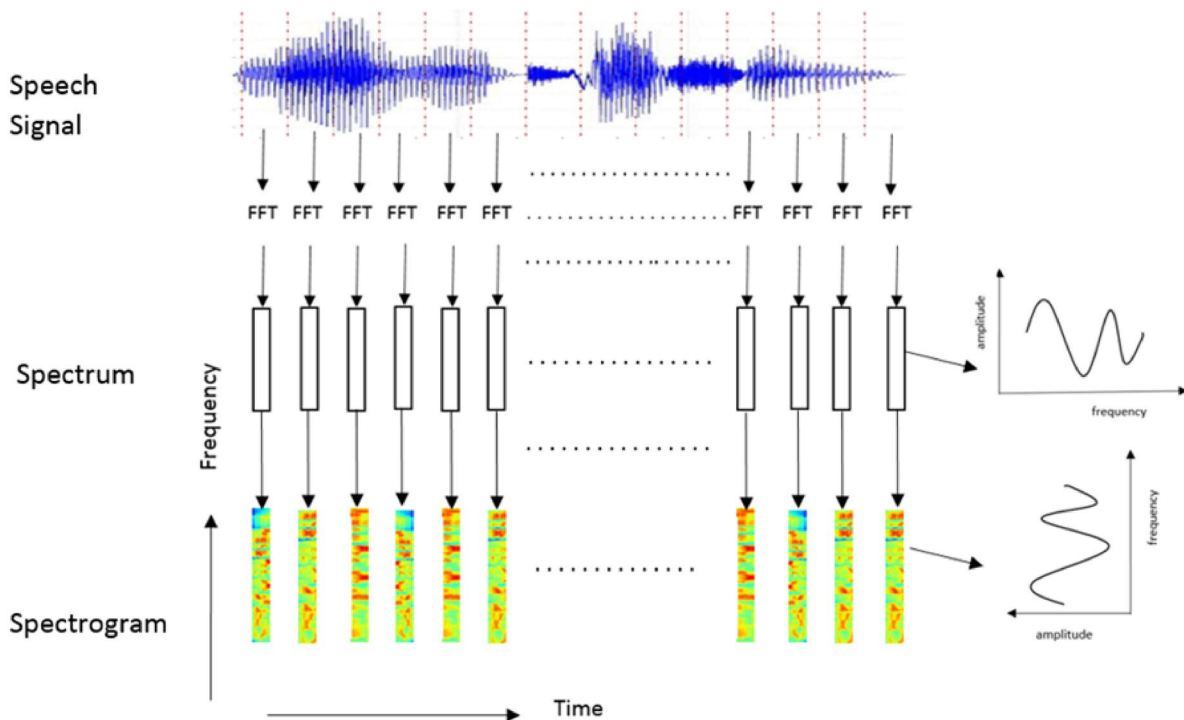
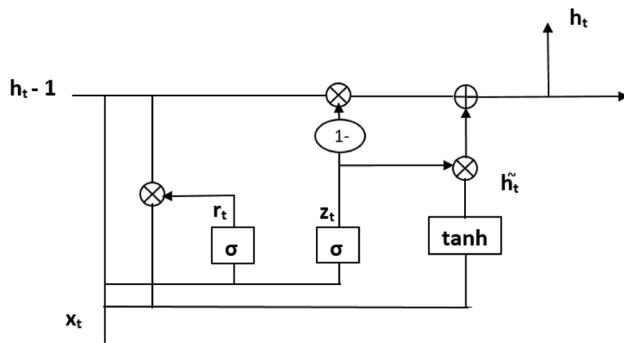


Fig. 3 Spectrogram generation

## 2.4 GRU model

Traditional RNN is suffering from the vanishing gradient problem. LSTM (Long Short Term Memory) and GRU (Gated Recurrent Unit) are two variants of RNNs to deal with the vanishing gradient problem. The GRU networks use fewer gates compared to LSTM and so it is faster. Each recurrent unit in GRU adaptively capture dependencies of different time scales and GRU is capable of dealing with speech signals efficiently. The information flow inside the unit is modulated by the update and reset gates. GRU does not have a separate memory cell but it keeps the existing content and add the new content on top of it. The GRU

model is suitable for speech emotion recognition as they can model long range dependencies and as speech emotions are with temporal dependency. The adaptive dependencies of each time step is captured using GRU and the full memory content can be exposed without any control. The training samples are passed through the network. The actual output and the obtained output are compared and the error is propagated back through the same path to adjust the variables. This process is repeated until the variables are well defined. When a new input comes, these variables are applied to make a prediction. It can control the information flow from the previous activation and the whole memory is exposed to the network. Figure 4 shows a GRU cell.



**Fig. 4** Gated recurrent unit

GRU network makes use of a reset gate  $r$  and an update gate  $z$ . The reset gate  $r_t$  at time  $t$  can be computed as

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (1)$$

When  $r_t = 0$ , the unit will forget the past. The update gate at time  $t$  controls the past state and decides about the unit's updating. It can be computed as

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (2)$$

The GRU activation  $h_t$  at time  $t$  can be computed as

$$h_t = \sigma(1 - z_t) * h_{t-1} + z_t * h_{\tilde{t}} \quad (3)$$

where  $h_{t-1}$  is the previous activation and  $h_{\tilde{t}}$  is the candidate activation which is computed as

$$h_{\tilde{t}} = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \quad (4)$$

The implementation difficulties of the complex deep networks can be lowered by using the APIs provided by the modern and powerful deep learning platforms like TensorFlow (Abadi et al., 2015), Theano (Bergstra et al., 2010) and Torch (Jacob, 2019). Our model is built with TensorFlow which is a second-generation interface for deploying machine learning algorithms. TensorFlow is a python based framework for implementing machine learning and is provided by Google (Abadi et al., 2015). TensorFlow models are very flexible and so we can execute these models on devices varying from mobile devices to large distributed systems (Wongsuphasawat et al., 2017). The computations with tensorflow are expressed as a data flow like model and then they are mapped to various hardware platforms. The training of the neural network can be scaled for larger deployments through parallelism. The high-level components of the tensorflow model are visualized as tensorflow graph which gives an overview of their relationships and the nested structure of the model. The visualization helps to understand the similarities and differences between various components, the details of the operations etc.

**Table 1** Various parameters used for the GRU Model

Parameter	GRU model with original database	GRU model with augmented database
Features used	Filterbank Energies	Filterbank Energies
Emotions	5 classes	5 classes
Training samples	1500	4500
Validation samples	300	1500
Testing samples	100	1369
Learning rate	0.001	0.001
Optimizer	Adam	Adam
Epochs	100	100

We have used filterbank energies as parameters for the GRU models.

### 3 Data augmentation

#### 3.1 Data augmentation based on tempo perturbation

In this method the duration of the audio signal is stretched without affecting the shape of its spectral envelope (Kanda et al., 2013). The time domain audio segment is decomposed into short analysis blocks and the perturbed output is constructed by relocating the analysis blocks along the time axis.

#### 3.2 Data augmentation based on speed perturbation

Speed perturbation is done by resampling the audio signal in time domain (Ko et al., 2015). When a perturbation factor  $\alpha$  is applied along the time axis of the audio segment  $x(t)$ . This will lead to change in audio duration and perturbation in the spectral envelope.

#### 3.3 Performance of GRU models with original and augmented dataset

The parameters used for building the GRU model are given in Table 1.

In the first experiment we have used the original TESS database for training and testing. Out of 2800 utterances in the database we have used 1900 utterances of five different classes. The emotions we consider in this experiment are angry, happy, sad, fear and neutral. Out of the 1900 utterances, 1500 are used for training the model. 300 are used for validation and 100 are used for testing the model. The model is trained for 100 epochs and the training accuracy, training

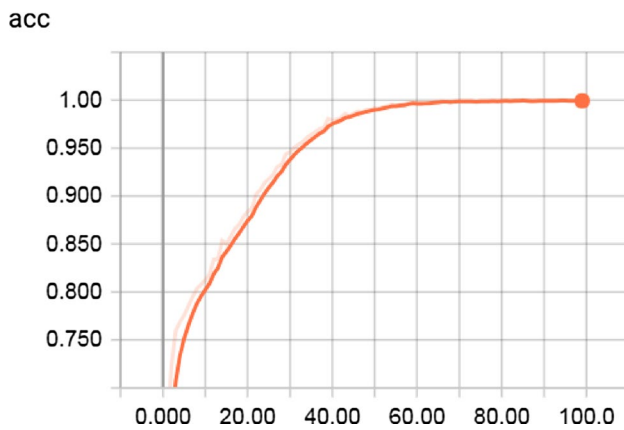


Fig. 5 Training accuracy of the GRU model

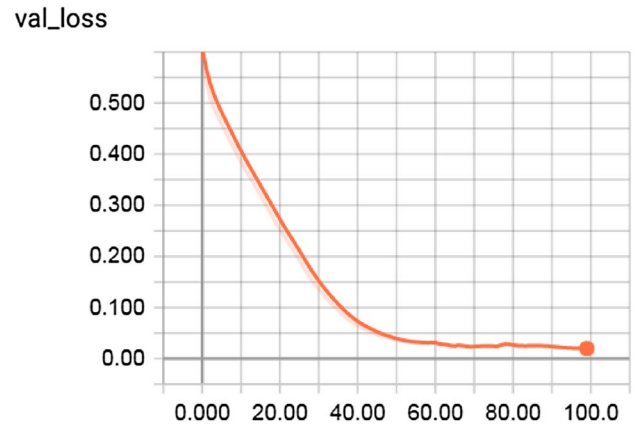


Fig. 8 Validation loss of the GRU model with original dataset

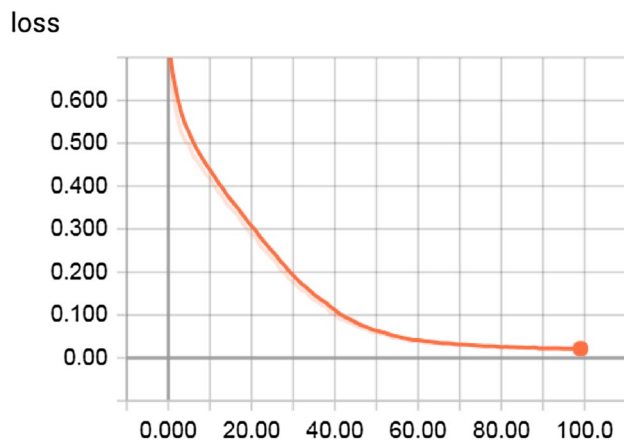


Fig. 6 Training loss of the GRU model with original dataset with original dataset

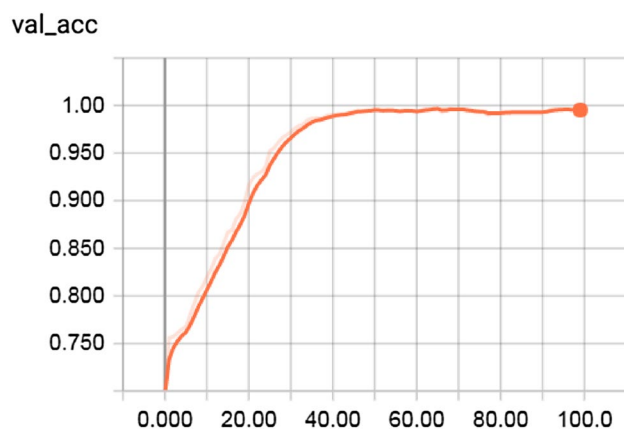


Fig. 7 Validation accuracy of the GRU model with original dataset

Table 2 Summary of GRU model with original data set

Emotion	Precision	Recall	F1-score	Support
Fear	0.95	0.95	0.95	20
Happy	0.95	0.90	0.92	20
Angry	0.86	0.95	0.90	20
Neutral	1.00	0.90	0.95	20
Sad	0.90	0.95	0.92	20
Avg/total	0.93	0.93	0.93	100

loss, validation accuracy and validation loss are given in Figs. 5, 6, 7, and 8.

The precision, recall and f1-score obtained with the GRU model for various emotions are given in Table 2. The precision is the the number of positive class predictions that actually belong to the positive class and is calculated as  $\frac{TP}{(TP+FP)}$ . Recall is the sensitivity or true positive rate. Recall is calculated as  $\frac{TP}{(TP+FN)}$ . F1-score is the harmonic mean of the precision and recall. It is calculated as  $2 * \frac{Precision * Recall}{Precision + Recall}$ .

For testing the model total 100 speech utterances, 20 from each class is given to the model. 19 samples are correctly classified for the emotion ‘fear’ and one is classified wrongly to the class ‘angry’. The classification accuracy for the emotion ‘fear’ is found to be 95%. 18 samples are correctly classified for the emotion ‘happy’ and two are classified wrongly as ‘angry’. The classification accuracy for the emotion ‘happy’ is found to be 90%. 19 samples are correctly classified for the emotion ‘angry’ and one is classified wrongly as ‘fear’. The classification accuracy for the emotion ‘angry’ is found to be 95%. 18 samples are correctly classified for the emotion ‘neutral’ and two are classified wrongly as ‘sad’. The classification accuracy for the emotion ‘neutral’ is found to be 90%. 19 samples are correctly classified for the emotion ‘sad’ and one is

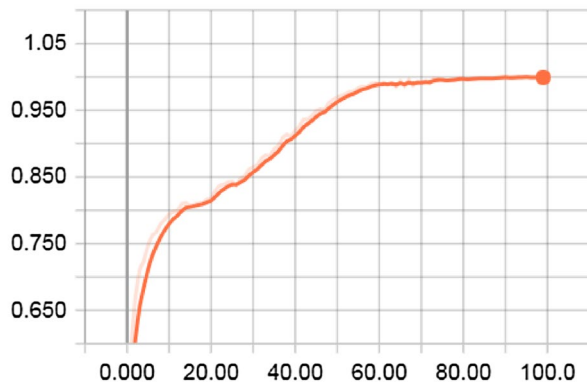
**Table 3** Confusion matrix of GRU Model with original training and testing sets

Emotion	Fear	Happy	Angry	Neutral	Sad
Fear	19	0	1	0	0
Happy	0	18	2	0	0
Angry	1	0	19	0	0
Neutral	0	0	0	18	2
Sad	0	1	0	0	19

**Table 4** Confusion matrix in the numerical form for the GRU Model with original training set and test set

Emotion	Fear	Happy	Angry	Neutral	Sad
Fear	95.00	0	5.00	0	0
Happy	0	90.00	10.00	0	0
Angry	5.00	0	95.00	0	0
Neutral	0	0	0	90.00	10.00
Sad	0	5.00	0	0	95.00

acc



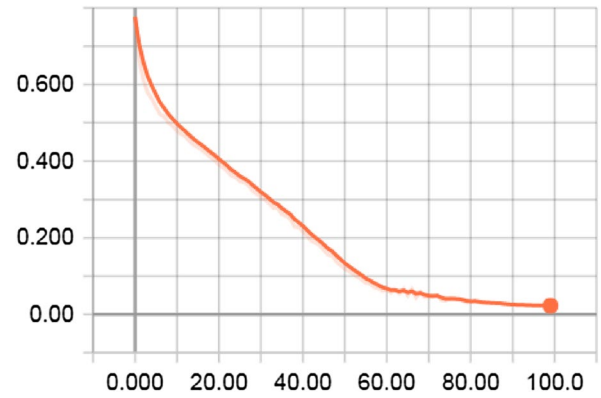
**Fig. 9** Training accuracy of the GRU model with augmented training set

classified wrongly as ‘happy’. The classification accuracy for the emotion ‘sad’ is found to be 95%. The confusion matrix of the GRU model with original data set is given in Tables 3 and 4.

In the second experiment we have used the augmented data set. Out of 7369 speech samples, 4500 samples of the augmented data set are used for training the model. 1500 samples are used for validation and 1369 samples are used for testing. The model is trained for 100 epochs and the training accuracy, training loss, validation accuracy and validation loss are given in Figs. 9, 10, 11, 12.

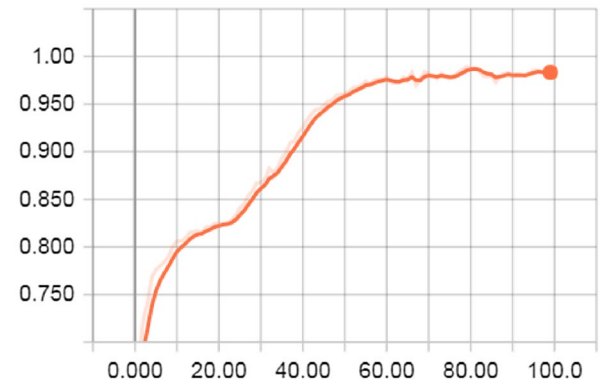
The precision, recall and f1-score obtained with the GRU model with augmented data set for various emotions are given in Table 5.

loss



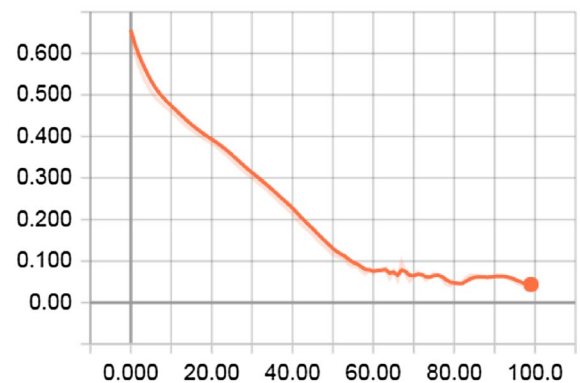
**Fig. 10** Training loss of the GRU model with augmented training set

val\_acc



**Fig. 11** Validation accuracy of the GRU model with augmented training set

val\_loss



**Fig. 12** Validation loss of the GRU model with augmented training set

**Table 5** Summary of GRU Model with augmented training set and original test set

Emotion	Precision	Recall	F1-score	Support
Fear	0.96	1.00	0.98	274
Happy	0.97	0.99	0.98	274
Sad	1.00	0.93	0.97	273
Angry	1.00	0.99	0.99	274
Neutral	0.97	0.95	0.94	274
Avg/total	0.97	0.97	0.97	1369

**Table 6** Confusion matrix of GRU Model with augmented training set and original test set

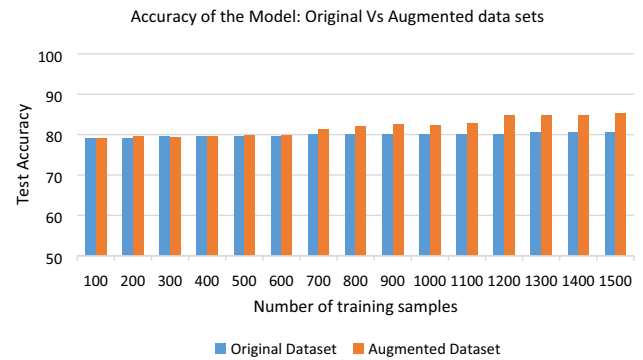
Emotion	Fear	Happy	Angry	Neutral	Sad
Fear	273	0	0	0	1
Happy	1	271	0	1	1
Sad	0	2	255	0	16
Angry	1	1	0	272	0
Neutral	8	6	0	0	260

**Table 7** Confusion matrix in the numerical form for the GRU Model with augmented training set and original test set

Emotion	Fear	Happy	Angry	Neutral	Sad
Fear	99.64	0	0	0	0.36
Happy	0.36	98.91	0	0.36	0.36
Sad	0	0.73	93.41	0	5.86
Angry	0.36	0.36	0	99.27	0
Neutral	2.92	2.19	0	0	94.89

Out of 274 samples, 273 samples are correctly classified for the emotion ‘fear’ and one is classified wrongly to the class ‘neutral’. The classification accuracy for the emotion ‘fear’ is found to be 99:64%. Out of 274 samples, 271 samples are correctly classified for the emotion ‘happy’. The classification accuracy for the emotion ‘happy’ is found to be 98:91%. Out of 273 samples, 255 samples are correctly classified for the emotion ‘sad’. The classification accuracy for the emotion ‘sad’ is found to be 93:41%. Out of 274 samples, 272 samples are correctly classified for the emotion ‘angry’. The classification accuracy for the emotion ‘angry’ is found to be 99:27%. Out of 274 samples, 260 samples are correctly classified for the emotion ‘neutral’ and the classification accuracy for the emotion ‘neutral’ is found to be 94:89%. The confusion matrix of the GRU model with original data set is given in Tables 6 and 7.

The accuracy of the model for the original and augmented datasets are shown in Fig. 13.

**Fig. 13** Accuracy of the GRU model

From the results it is clear that the GRU model gives better results with augmented dataset. An average accuracy of 93% is obtained when the model is trained with 5000 samples of the augmented data set.

## 4 Conclusion

Automatic recognition of speech emotions has gained much importance nowadays since it can be used in many areas like human machine interaction, translation of one language to another, gaming etc. It can also be applied to provide better customer service. Traditional machine learning techniques are found to be inefficient as the emotions are to be identified only from the speech signals. Such complicated and challenging problems can be solved by using deep learning techniques which use automatic feature learning on a large amount of data. We have implemented two different deep learning networks for automatic recognition of emotions from speech. From our studies we conclude that the GRU model with filter bank energies is very much suitable for automatic emotion recognition from speech and it gives better results. We have overcome the limitation in the availability of big data set by augmenting the data set. The GRU model has been applied to the original as well as the augmented data set. By the experiment it has been understood that GRU models perform very well with sequential data such as speech signals. The model gave good performance when trained with huge number of inputs. So data augmentation is found to be helpful to increase the efficiency of the model.

## References

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., & Ghemawat, S. (2015). Tensorflow: Large-scale machine learning on



- heterogeneous systems. *Software Available from Tensorflow. Org*, 1(2), 2015.
- Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., & Bengio, Y. (2010). Theano: A CPU and GPU math compiler in python. In *Proc. 9th Python in Science Conf*, vol. 1, pp. 3–10.
- Bhatti, M. W., Wang, Y., & Guan, L. (2004). A neural network approach for human emotion recognition in speech. In *2004 IEEE International Symposium on Circuits and Systems* (IEEE Cat. No. 04CH37512) (Vol. 2, pp. II-181). IEEE.
- Chernykh, V., & Prikhodko, P. (2017). Emotion recognition from speech with recurrent neural networks. <http://arxiv.org/abs/1701.08071>.
- Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001). Emotion recognition in human–computer interaction. *IEEE Signal Processing Magazine*, 18(1), 32–80.
- Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4), 407–422.
- Eyben, F., Wöllmer, M., & Schuller, B. (2009). OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In *ACII 2009. 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, (pp. 1–6). IEEE.
- Ghosh, S., Laksana, E., Morency, L. P., & Scherer, S. (2015). Learning representations of affect from speech. <http://arxiv.org/abs/1511.04747>.
- Han, K., Yu, D., & Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth annual conference of the International Speech Communication Association*.
- Haixiang, W. (2020). Emotional analysis of Bogus statistics in social media. *Journal of Ubiquitous Computing and Communication Technologies (UCCT)*, 2(03), 178–186.
- Hu, H., Xu, M. X., & Wu, W. (2007). GMM supervector based SVM with spectral features for speech emotion recognition. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07* (Vol. 4, pp. IV-413). IEEE.
- Huang, Z., Dong, M., Mao, Q., & Zhan, Y. (2014). Speech emotion recognition using CNN. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 801–804).
- Jacob, I. J. (2019). Capsule network based biometric recognition system. *Journal of Artificial Intelligence*, 1(02), 83–94.
- Kanda, N., Takeda, R., & Obuchi, Y. (2013). Elastic spectral distortion for low resource speech recognition with deep neural networks. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 309–314). IEEE.
- Keren, G., & Schuller, B. (2016). Convolutional RNN: an enhanced model for extracting features from sequential data. In *2016 International Joint Conference on Neural Networks (IJCNN)*, (pp. 3412–3419). IEEE.
- Kim, Y., Lee, H., & Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 3687–3691). IEEE.
- Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.
- Lee, J., & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*.
- Manoharan, S. (2019). Study on Hermitian graph wavelets in feature detection. *Journal of Soft Computing Paradigm (JSCP)*, 1(01), 24–32.
- Mao, X., Chen, L., & Fu, L. (2009). Multi-level speech emotion recognition based on HMM and ANN. In *2009 WRI World Congress on Computer Science and Information Engineering* (Vol. 7, pp. 225–229). IEEE.
- Mao, Q., Dong, M., Huang, Z., & Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8), 2203–2213.
- Neiberg, D., Elenius, K., & Laskowski, K. (2006). Emotion recognition in spontaneous speech using GMMs. In *Ninth International Conference on Spoken Language Processing*.
- Nicholson, J., Takahashi, K., & Nakatsu, R. (2000). Emotion recognition in speech using neural networks. *Neural Computing & Applications*, 9(4), 290–296.
- Ntalampiras, S., & Fakotakis, N. (2012). Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Transactions on Affective Computing*, 3(1), 116–125.
- Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, 41(4), 603–623.
- Prasomphan, S. (2015). Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. In *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)* (pp. 73–76). IEEE.
- Schuller, B., Rigoll, G., & Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*. (Vol. 2, pp. II-1). IEEE.
- Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., & Rigoll, G. (2005). Speaker independent speech emotion recognition by ensemble classification. In *IEEE International Conference on Multimedia and Expo. ICME 2005*, (pp. 864–867). IEEE.
- Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., & Schuller, B. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5688–5691). IEEE.
- Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2016). 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.
- Wang, S., Ling, X., Zhang, F., & Tong, J. (2010). Speech emotion recognition based on principal component analysis and back propagation neural network. In *2010 International Conference on Measuring Technology and Mechatronics Automation* (Vol. 3, pp. 437–440). IEEE.
- Wang, Z. Q., & Tashev, I. (2017). Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (pp. 5150–5154). IEEE.
- Wongsuphasawat, K., Smilkov, D., Wexler, J., Wilson, J., Mane, D., Fritz, D., Krishnan, D., Viégas, F. B., & Wattenberg, M. (2017). Visualizing dataflow graphs of deep learning models in tensorflow. *IEEE Transactions on Visualization and Computer Graphics*, 24(1), 1–12.
- Wu, C. H., & Liang, W. B. (2011). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1), 10–21.
- Zheng, W. L., Zhu, J. Y., Peng, Y., & Lu, B. L. (2014). EEG-based emotion classification using deep belief networks. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, (pp. 1–6). IEEE.

- Zheng, W. Q., Yu, J. S., & Zou, Y. X. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 827–831). IEEE.
- Zhou, J., Wang, G., Yang, Y., & Chen, P. (2006). Speech emotion recognition based on rough set and SVM. In *2006 5th IEEE International Conference on Cognitive Informatics* (Vol. 1, pp. 53–61). IEEE.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.