Journal of Computer Science

Original Research Paper

Deep Learning Models for Speech Emotion Recognition

Praseetha, V.M. and Sangil Vadivel

Department of Computer Science,
Birla Institute of Technology and Science Pilani, International Academic City, Dubai, UAE

Article history
Received: 11-08-2018
Revised: 01-10-2018
Accepted: 24-11-2018

Corresponding Author:
Praseetha, V.M.
Department of Computer Science, Birla Institute of Technology and Science Pilani, International Academic City, Dubai, UAE
Email: praseethasunil@gmail.com

Abstract: Emotions play a vital role in the efficient and natural human computer interaction. Recognizing human emotions from their speech is truly a challenging task when accuracy, robustness and latency are considered. With the recent advancements in deep learning now it is possible to get better accuracy, robustness and low latency for solving complex functions. In our experiment we have developed two deep learning models for emotion recognition from speech. We compare the performance of a feed forward Deep Neural Network (DNN) with the recently developed Recurrent Neural Network (RNN) which is known as Gated Recurrent Unit (GRU) for speech emotion recognition. GRUs are currently not explored for classifying emotions from speech. The DNN model gives an accuracy of 89.96% and the GRU model gives an accuracy of 95.82%. Our experiments show that GRU model performs very well on emotion classification compared to the DNN model.

Keywords: Deep Learning, Neural Network, Deep Neural Network, Recurrent Neural Network, Gated Recurrent Unit

Introduction

One can express his or her perspective on various matters through emotions. Emotions play an important role in the field of Human Computer Interaction as they can provide improved services based on the emotions of users (Anagnostopoulos and Iliou, 2010). The human emotions are expressed mostly by speech and face. Speech or voice is the most important medium of communication between humans. Speech emotion recognition is a challenging task as the emotions are recognized based only on the voice of the speaker. The extraction of the features which are discriminative and effective to recognize the emotions is a very big issue while considering speech emotion recognition (El Ayadi *et al*., 2011; Schuller *et al*., 2011). Lot of features are used for recognizing the emotions from speech and they can be classified as (1) logical (2) acoustic (3) hybrid (4) context information (Luengo *et al*., 2010). The performance of each feature varies with respect to situations. Mapping from perceptual input to the output is very complicated in many machine learning problems. Deep Neural Networks can handle large amount of data and as a result DNNs are dominating the traditional machine learning algorithms. The feature hierarchies are learned by deep learning by forming the higher level features from the lower level features (Bengio, 2009). Complex tasks like voice search, automatic text generation, speech emotion recognition can be made possible by deep learning. For such applications the deep learning networks are accurate and faster than the traditional learning networks and they produce the best results by automatically learning the features. Thus deep learning can be considered as a powerful framework for automatic learning.

The organization of this paper is as follows. The section 'Related Works' will describe about the previous works on speech emotion recognition. A detailed description about feature extraction is given in the section 'Featuture Extraction'. Following that is the 'Implementation' section in which the implementation of the DNN model and the GRU model is explained. The section 'Results and Discussion' discusses about the results obtained. Finally, the conclusion and future scope of the work is given in the section 'Conclusion'.

Related Works

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 has a good scope for research since it can be applied in many areas of human machine interaction. Some of the researches conducted in this area are briefed here.

Eyben *et al*. (2009) extracted the low level features from sound signals and an utterence level calculation is made for recognizing the emotions. Stuhlsatz *et al*. (2011) 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. They compared the performance of a deep neural network with SVM on speech emotion recognition. The extracted acoustic features are given as input to the classifiers. The experimental results showed a good performance of DNN over the SVM. Cibau *et al*. (2013) and Kim *et al*. (2013) proposed deep learning models which learn the feature representation with deep network architectures. A comparison of DNN-HMM classifier and GMM-HMM classifier on speech emotion recognition is done by Li *et al*. (2013) in their work. In DNN-HMM, the discriminative speech features are extracted using the DNN and the HMM uses these features for classification of emotions. The DNN-HMM performed well compared to the GMM-HMM.

Mao *et al*. (2014) used the narrowband spectrogram as input to the CNN to learn the discriminative speech features. Then these features are given to an SVM classifier to classify the speech signals according to the emotions. A Deep Neural Network (DNN) model is explained by Han *et al*. (2014) in which the utterence level characteristics is learned from the frame level features. A DNN with narrowband spectrogram is used by Fayek *et al*. (2015) for emotion classification. The authors used narrow band spectrograms as input and the DNN outperformed the traditional machine learning methods. Bhargava and Rose (2015) proposed an advanced bottleneck deep neural networks (DNNs) which take the windowed speech waveforms as input. A model using auto-encoder is explained by Ghosh *et al*. (2015) to learn the frame-level features for computing the utterance level statistics. RNN with Bidirectional Long-Short Term Memory (BLSTM) model is used by Lee and Tashev (2015) 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.

A Recurrent Neural Network (RNN) model with Long-Short Term Memory (LSTM) is used by Keren and Schuller (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. A sequence of two CNN layers applied at two different time resolutions followed by a LSTM RNN is proposed by Trigeorgis *et al*. (2016) for speech emotion recognition. Wang and Tashev (2017) proposed 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 (ELM).

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 two different deep learning models and the results are compared. The first model is a simple DNN model which uses a feed forward network and the second model uses Gated Recurrent Units (GRUs). GRU is a new version of Recurrent Neural Network (RNN) which eliminates the problem of vanishing gradients. GRU networks have not yet been used for automatic speech emotion recognition. Our work compares the performance of DNN model and GRU model for speech emotion recognition.

We have used 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.

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'.

The energy of the speech signal is increased by passing the signal through a filter. This is the pre-emphasis (Basu *et al*., 2017) 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 1sec speech we will get 100 frames and so speech vectors. The frames are then converted from time domain to frequency domain and the frequencies which are present in the frames are

DOI: 10.3844/jcssp.2018.1577.1587

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 MFCC calculation and spectrogram generation are shown in Fig. 2 and 3.

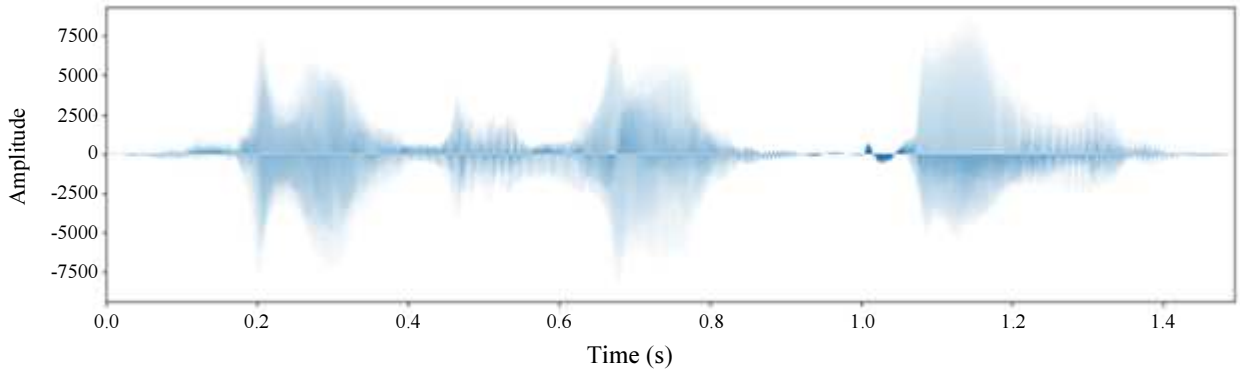

Fig. 1: A sample speech signal with emotion 'angry'

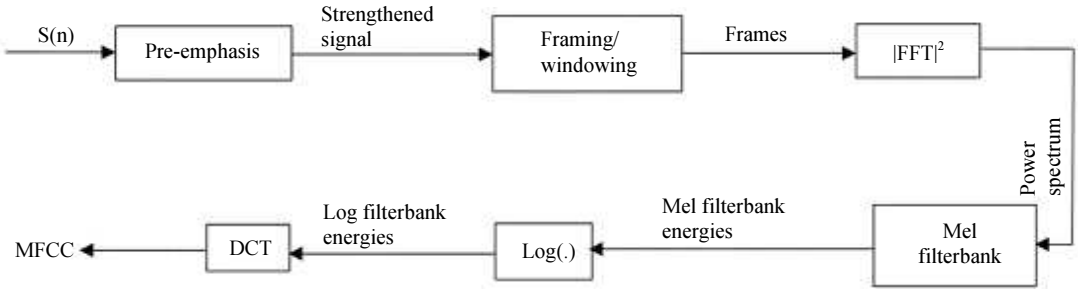

Fig. 2: Different stages in MFCC feature extraction

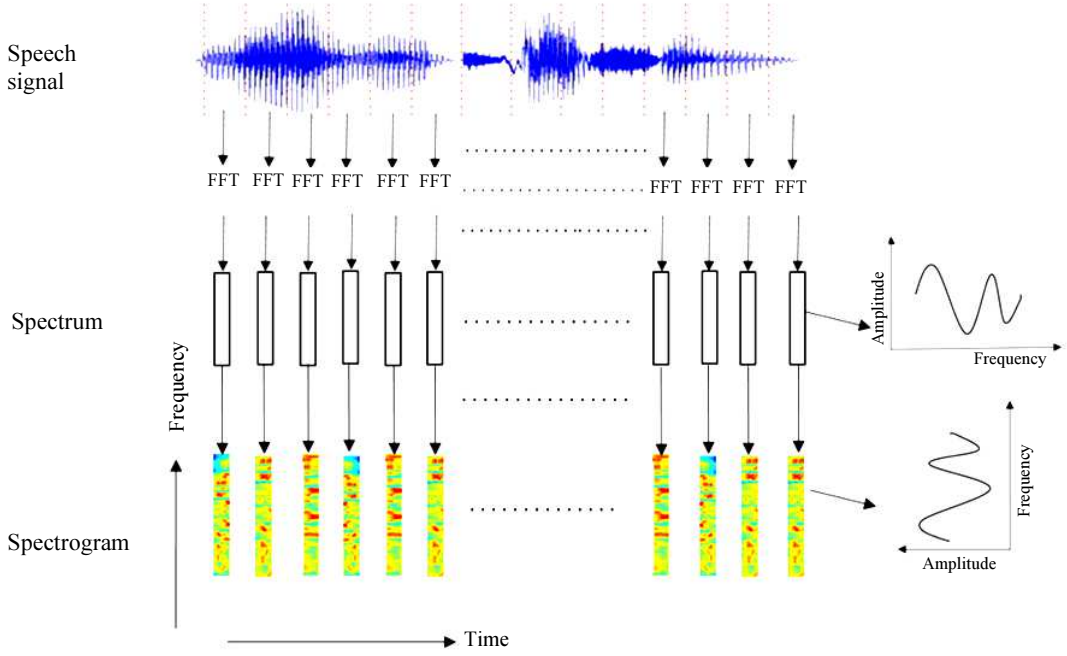

Fig. 3: Spectrogram generation

A set of mel filters 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:

$$m = 2585 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

$$f = 700\left(10^{m/2595} - 1\right) \quad (2)$$

The MFCCs are extracted from each speech signal and then the delta MFCCs are calculated. Our first GRU model uses 40 features per frame: 20 MFCCs and 20 delta MFCCs. The second GRU model uses filter bank energies from 26 filters as features.

Implementation

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*., 2016), Theano (Bergstra *et al*., 2010) and Torch (Collobert *et al*., 2002). 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*., 2016). 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*., 2018). The computations with Tensorflow are expressed as a dataflow 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.

Model 1: DNN Model

A fully connected deep neural network model is built for recognizing the emotions from speech. Deep feed forward neural networks can be used to define a mapping $y = f(x, \theta)$ between the input x and output y and find the best approximation of the function by learning the value of θ. Since there is no backward connection, the model is called a feed forward model. Figure 4 shows the data flow graph of the DNN model created. The model contains an input layer, two hidden layers and an output layer. All the layers are fully connected. We pass the training samples through the input layer to the network. The error is calculated by comparing the actual output and the obtained output. The weights of the neurons are updated according to the value of the error such that the error get decreased.

The co-dependency among the neurons are avoided by using a dropout layer in between the second hidden layer and the output layer. The parameter learning rate is used to determine the amount by which the weights are getting changed. The learning rate of the model is set to 0.001 and the optimizer used for the model is Adam. The training is done for 100 epochs with a batch size of 1000. The output layer uses softmax activation function and all the other layers use ReLU activation function. The probability associated with the output is determined using the softmax function. This function is implemented in the last layer of the network. The Rectified Linear Units (ReLU) are the default activation function for feed forward neural networks.

Model 2: GRU Model

As a second experiment we have constructed a Gated Recurrent Unit (GRU) model for speech emotion recognition. GRU is the improved version of Recurrent Neural Network (RNN). RNNs are those with loops in them which persist the information. The GRU network is suitable for speech emotion recognition as they can model long range dependencies and as speech emotions are with temporal dependency. 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.

The GRU networks use fewer parameters and so it is faster. It can control the information flow from the previous activation and the whole memory is exposed to the network. 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:

$$r_t = \sigma\left(W_r \bullet \left[h_{t-1}, x_t\right]\right) \quad (3)$$

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:

$$z_t = \sigma\left(W_z \bullet \left[h_{t-1}, x_t\right]\right) \quad (4)$$

DOI: 10.3844/jcssp.2018.1577.1587

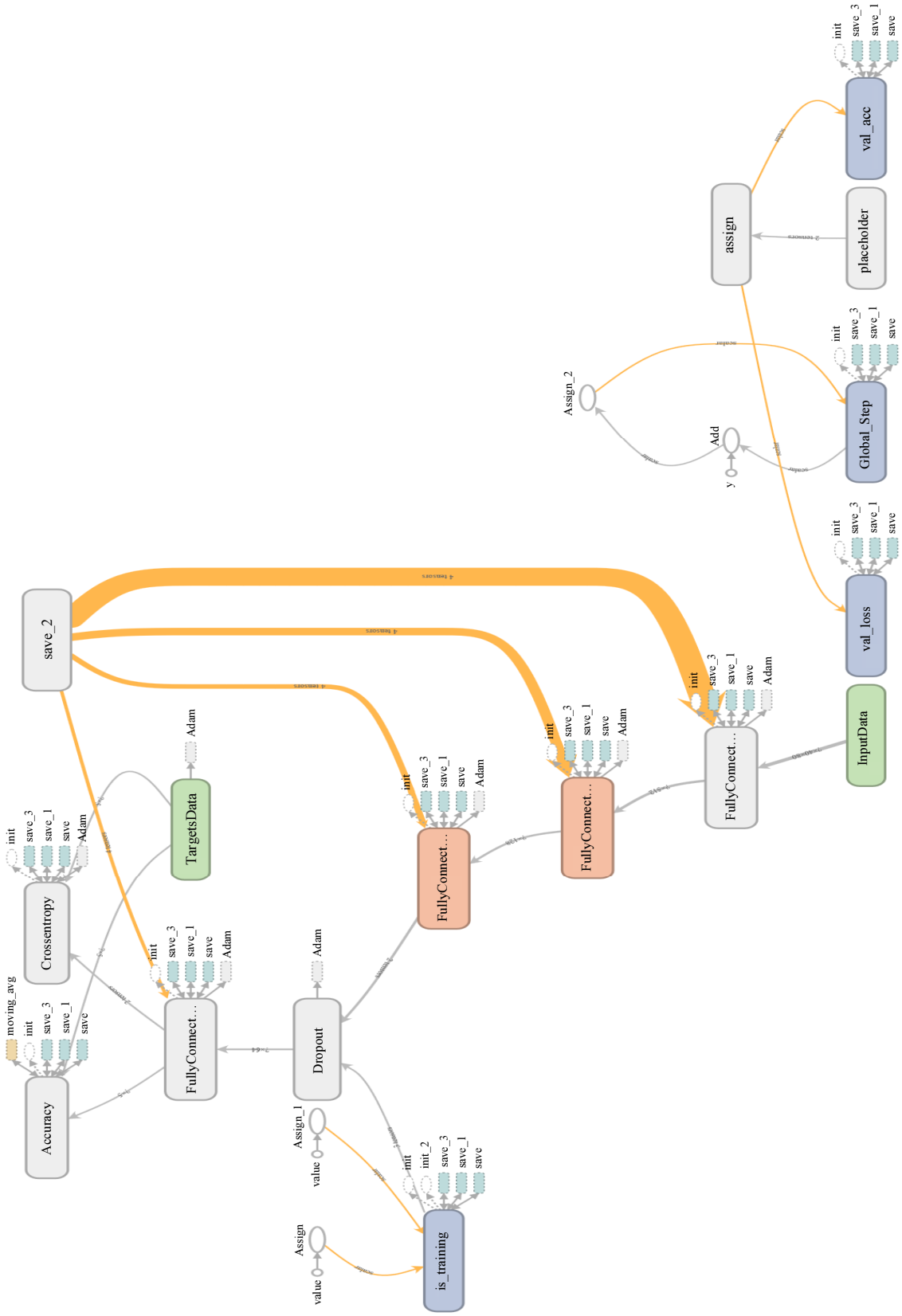

Fig. 4: Data flow graph of CNN model

DOI: 10.3844/jcssp.2018.1577.1587

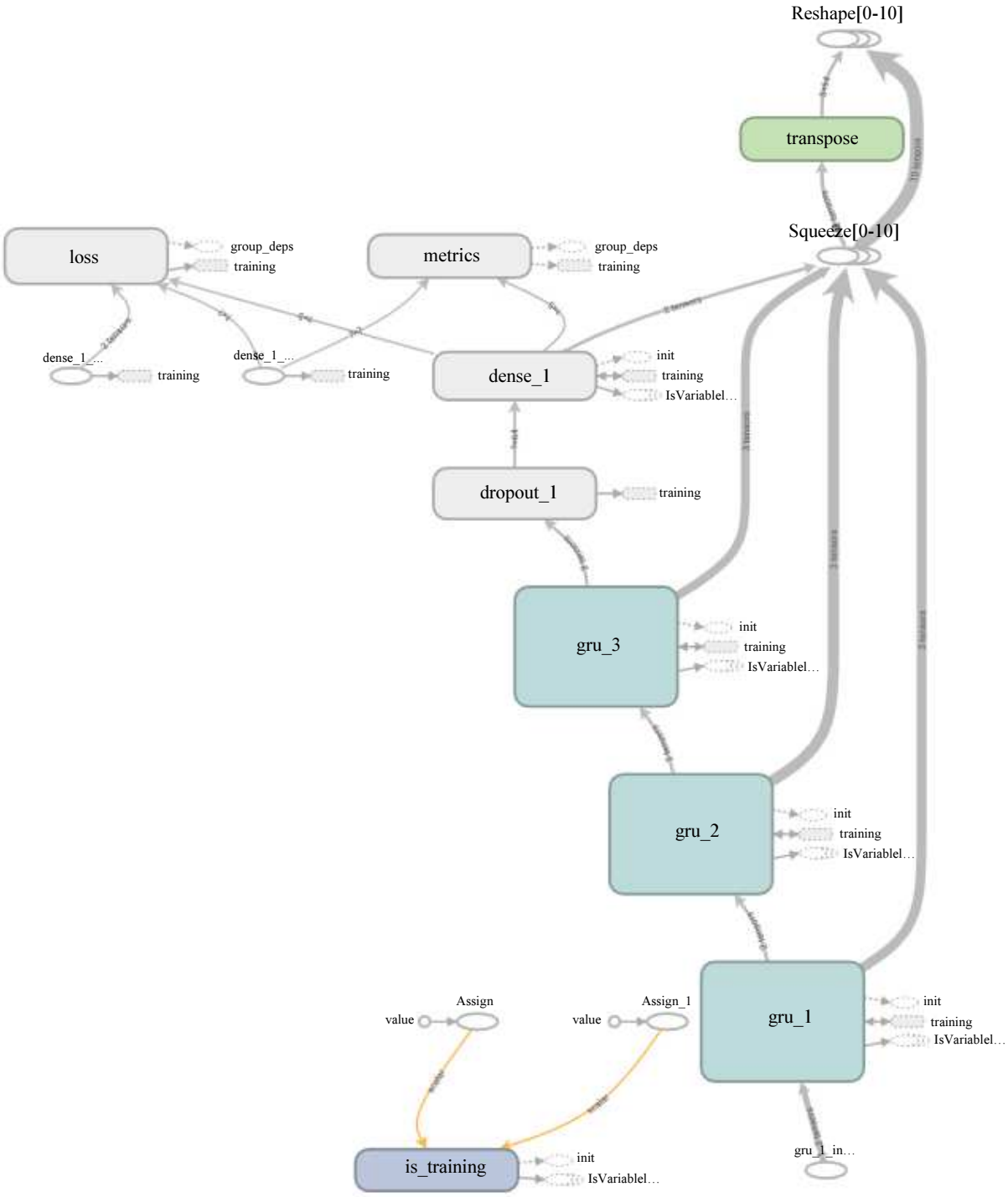

Fig. 5: Data flow graph of GRU model

The GRU activation ht at time t can be computed as:

$$h_t = \sigma\left(1 - z_t\right) \times h_{t-1} + z_t \times \tilde{h}_t \tag{5}$$

where, $h_{t\text{-}1}$ is the previous activation and $\tilde{h}_t$ is the candidate activation which is computed as:

$$\tilde{h}_t = \tanh\left(W \bullet \left[r_t \times h_{t-1}, x_t\right]\right) \tag{6}$$

GRU networks eliminate the problem of vanishing gradients and the computations are very simple for this model. The data flow graph of the GRU model is shown in Fig. 5.

DOI: 10.3844/jcssp.2018.1577.1587

Results and Discussion

Our data set contains utterances of five emotions. The emotions we used are angry, happy, sad, fear and neutral. The entire dataset had been divided into training set which contain 70% of the dataset and validation set which contain 30% of the dataset. The number of epochs is set to 100 and the batch size is kept 1000. The learning rate for both the models is set to 0.001. As the number of files in the dataset increases the model tries to learn many intrinsic features from inputs and this essentially helps to increase the accuracy of the model. The training accuracy and training loss of DNN model are shown in Fig. 6 and 7.

The training is done in 600 steps and 100 epochs. Out of 500 input samples 350 were selected randomly as training set and the remaining 150 are selected as validation set. The accuracy of the model is found to be 89.96% over a test dataset containing 95 samples.

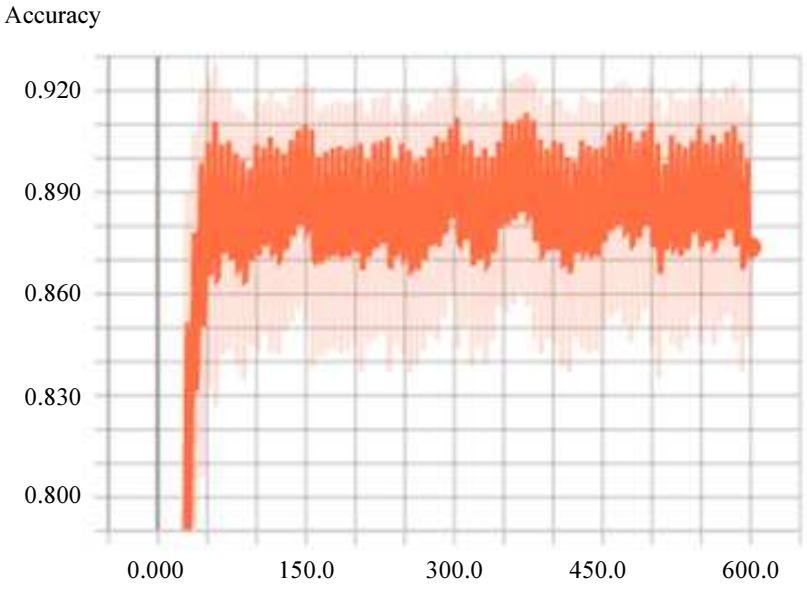

Fig. 6: Training accuracy of DNN model

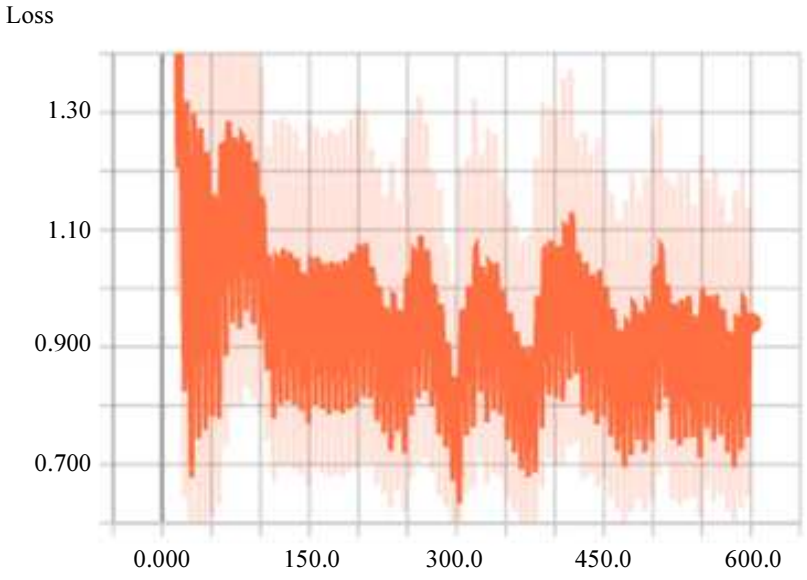

Fig. 7: Training loss of DNN model

DOI: 10.3844/jcssp.2018.1577.1587

The confusion matrix and the numerical confusion matrix for the test data set containing 95 samples with the DNN model are given in Table 1 and 2. The percentage accuracy for each emotion is given in the numerical confusion matrix. For the emotion 'fear' an accuracy of 78.95% is achieved on the test data set. For the emotion 'happy', the accuracy obtained is 84.21% and for 'angry', the accuracy obtained is 89.47%. The emotion 'neutral' has got an accuracy of 78.95% and the emotion 'sad' has got an accuracy of 84.21%.

The training accuracy and training loss of the GRU model are given in Fig. 8 and 9.

The dataset contains 1369 utterances of the above mentioned emotions and it is divided into training set and validation set in the ratio 70:30. The model is trained for 100 epochs and the accuracy of the model is found to be 95.82% over a test dataset of 100 samples. We found that the GRU model works much better than the DNN model. The validation accuracy and validation loss of the GRU model are also given in Fig. 10 and 11.

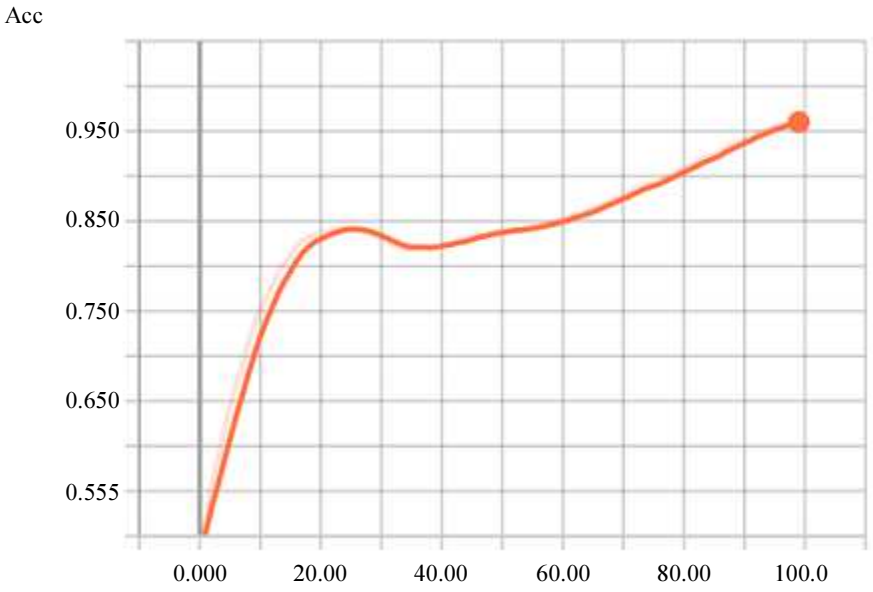

Fig. 8: Training Accuracy of GRU model

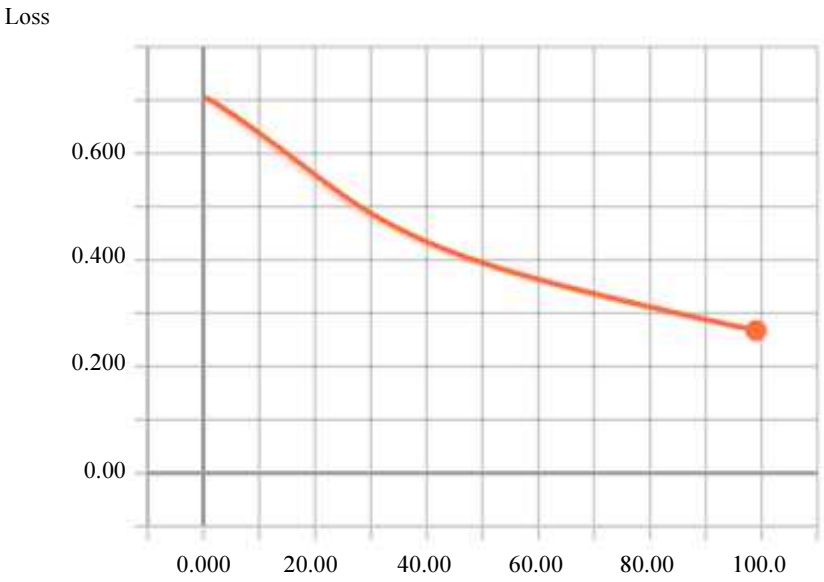

Fig. 9: Training loss of GRU model

DOI: 10.3844/jcssp.2018.1577.1587

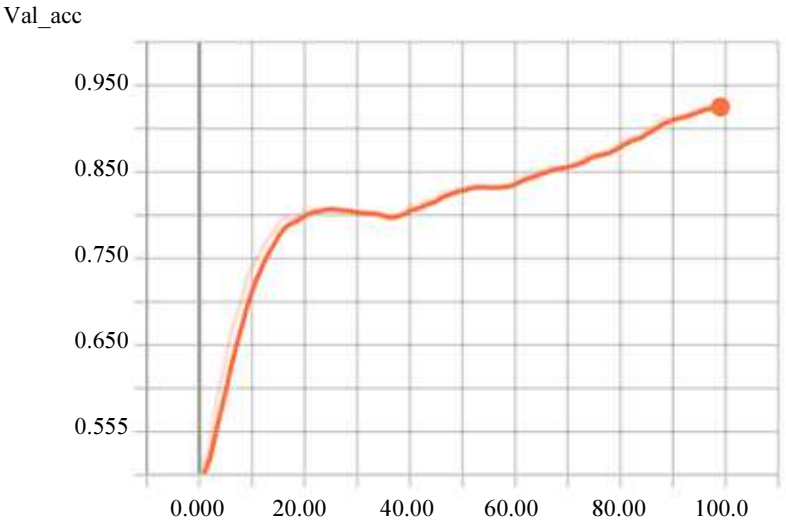

Fig. 10: Validation Accuracy of GRU model

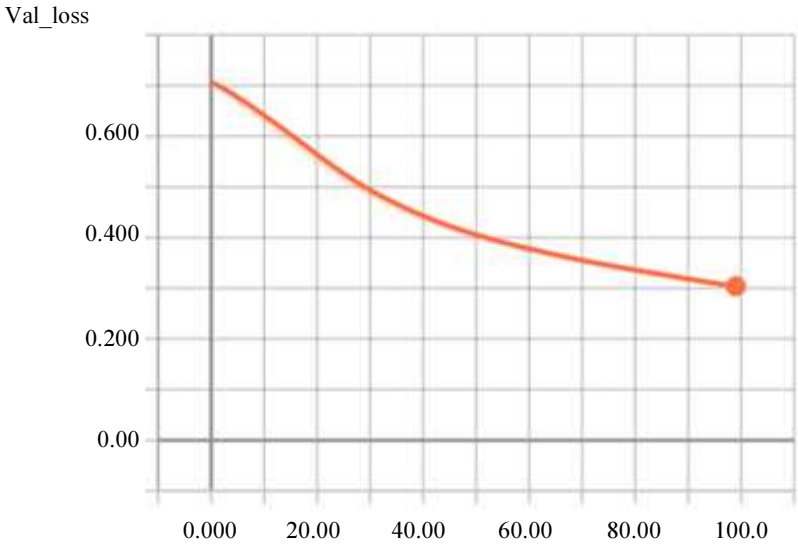

Fig. 11: Validation loss of GRU model

Table 1: Confusion matrix of DNN Model

	Fear	Happy	Angry	Neutral	Sad
Fear	15	2	2	0	0
Happy	1	16	1	1	0
Angry	0	0	17	1	1
Neutral	1	1	0	15	2
Sad	1	1	0	0	16

Table 2: Numerical Confusion matrix of DNN Model

	Fear	Happy	Angry	Neutral	Sad
Fear	**78.95**	0.52	10.52	0	0
Happy	5.26	**84.21**	5.26	5.26	0
Angry	0	0	**89.47**	5.26	5.26
Neutral	5.26	5.26	0	**78.95**	10.52
Sad	5.26	5.26	5.26	0	**84.21**

Table 3: Confusion matrix of GRU Model

	Fear	Happy	Angry	Neutral	Sad
Fear	19	0	1	0	0
Happy	0	18	1	1	0
Angry	0	0	20	0	0
Neutral	0	0	0	20	0
Sad	0	1	0	0	19

Table 4: Numerical Confusion matrix of GRU Model

	Fear	Happy	Angry	Neutral	Sad
Fear	**95.00**	0	5.00	0.00	0
Happy	0	**90.00**	5.00	5.00	0
Angry	0	0	**100.00**	0	0
Neutral	0	0	0	**100.00**	0
Sad	0	5.00	0	0	**95.00**

Table 5: Comparison with other recent studies

Method	Emotions	Accuracy
DCNN (Zheng *et al*., 2015)	5 classes	40.02%
Acoustic and lexical feature model (Jin *et al*., 2015)	4 classes	69.20%
BLSTM (Lee and Tashev, 2015)	4 classes	63.89%
Our Model (DNN)	**5 classes**	**89.96%**
Our Model (GRU)	**5 classes**	**95.82%**

The confusion matrix and the numerical confusion matrix for the test data set containing 95 samples with the GRU model are given in Table 3 and 4. The percentage accuracy for each emotion is given in the numerical confusion matrix. For the emotion 'fear' an accuracy of 95% is achieved on the test data set. For the emotion 'happy', the accuracy obtained is 90%. For 'angry' and 'neutral', the accuracy obtained is 100%. The emotion 'sad' has got an accuracy of 95%.

GRU models have not yet explored for Automatic Speech Emotion Recognition. Since GRU models are simple in calculation, the training can be done faster. When compared with other studies, we found that the results are better when we use GRU model for recognizing emotions from speech. Table 5 shows the accuracy of recently developed models and the accuracy of our model. The DCNN model developed by Zheng *et al*. (2015) has an overall accuracy of 40.02% for five classes of emotions. The Acoustic and Lexical Feature Model proposed by Jin *et al*. (2015) has got an accuracy of 69.2%. The BLSTM model explained by (Lee and Tashev, 2015) reported an accuracy of 63.89%. Our DNN model and GRU model got good accuracy compared to the above models. The accuracy that we got with our DNN model is 89.96% and with GRU model is 95.82%.

Conclusion

Automatic recognition of speech 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. Our first model uses a feed forward DNN and the second model uses a GRU network. From our studies we conclude that the GRU model is very much suitable for automatic emotion recognition from speech and it gives better results than the DNN model. The purpose of this study is to explore GRU for speech emotion recognition and to prove that the performance of GRU model on speech emotion recognition is very good compared to the normal DNN model. Our DNN model has got an accuracy of 89:96% whereas our GRU model has got an accuracy of 95:82%.

In future, more studies on GRU models could be conducted to improve the efficiency of speech emotion recognition. Deep learning models work well when there is a large number of data for training and testing. Since the availability of database IA a major difficulty in automatic speech emotion recognition, data augmentation can be applied to increase the size of the database.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The corresponding author confirms that the coauthor has read and approved the manuscript and there are no ethical issues involved.

References

Abadi, M., A. Agarwal, P. Barham, E. Brevdo and Z. Chen *et al*., 2016. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, (OSDI'16), Savannah, GA, USA, pp: 265-283.

Anagnostopoulos, C.N. and T. Iliou, 2010. Towards Emotion Recognition from Speech: Definition, Problems and the Materials of Research. In: Semantics in Adaptive and Personalized Services, Wallace, M., I.E. Anagnostopoulos, P. Mylonas and M. Bielikova (Eds.), Springer, pp: 127-143.

Basu, S., J. Chakraborty, A. Bag and M. Aftabuddin, 2017. A review on emotion recognition using speech. Proceedings of the International Conference on Inventive Communication and Computational Technologies, Mar. 10-11, IEEE Xplore Press, Coimbatore, India, pp: 109-114. DOI: 10.1109/ICICCT.2017.7975169

Bengio, Y., 2009. Learning deep architectures for Ai. Foundat. Trends® Mach. Learn., 2: 1-127. DOI: 10.1561/2200000006

Bergstra, J., O. Breuleux, F. Bastien, P. Lamblin and R. Pascanu *et al*., 2010. Theano: A CPU and GPU math compiler in python. Proceedings of the 9th Python in Science Conference, (PSC' 10), Austin, TX, pp: 3-10.

Bhargava, M. and R. Rose, 2015. Architectures for deep neural network based acoustic models defined over windowed speech waveforms. Proceedings of the 16th Annual Conference of the International Speech Communication Association, Sept. 6-10, Dresden, Germany, pp: 6-10.

Cibau, N.E., E.M. Albornoz and H.L. Rufiner, 2013. Speech emotion recognition using a deep autoencoder. Proceedings of the 15th Reunion de Trabajo en Procesamiento de la Informacion y Control (PIC' 13), San Carlos de Bariloche.

Collobert, R., S. Bengio and J. Mariethoz, 2002. Torch: A modular machine learning software library. Technical Report, Idiap.

El Ayadi, M., M.S. Kamel and F. Karray, 2011. Survey on speech emotion recognition: Features, classification schemes and databases. Patt. Recognit., 44: 572-587. DOI: 10.1016/j.patcog.2010.09.020

Eyben, F., M. Wollmer and B. Schuller, 2009. OpenEAR-introducing the Munich open-source emotion and affect recognition toolkit. Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, Sept. 10-12, IEEE Xplore Press, Amsterdam, Netherlands, pp: 1-6. DOI: 10.1109/ACII.2009.5349350

Fayek, H.M., M. Lech and L. Cavedon, 2015. Towards real-time speech emotion recognition using deep neural networks. Proceedings of the 9th International Conference on Signal Processing and Communication Systems, Dec. 14-16, IEEE Xplore Press, Cairns, QLD, Australia, pp: 1-5. DOI: 10.1109/ICSPCS.2015.7391796

Ghosh, S., E. Laksana, L.P. Morency and S. Scherer, 2015. Learning representations of affect from speech. arXiv preprint arXiv:1511.04747.

Han, K., D. Yu and I. Tashev, 2014. Speech emotion recognition using deep neural network and extreme learning machine. Proceedings of the 15h Annual Conference of the International Speech Communication Association, Sept. 14-18, Singapore.

Jin, Q., C. Li, S. Chen and H. Wu, 2015. Speech emotion recognition with acoustic and lexical features. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 19-24, IEEE Xplore Press, Brisbane, QLD, Australia, pp: 4749-4753. DOI: 10.1109/ICASSP.2015.7178872

Keren, G. and B. Schuller, 2016. Convolutional RNN: An enhanced model for extracting features from sequential data. Proceedings of the International Joint Conference on Neural Networks, Jul. 24-29, IEEE Xplore Press, Vancouver, BC, Canada, pp: 3412-3419. DOI: 10.1109/IJCNN.2016.7727636

Kim, Y., H. Lee and E.M. Provost, 2013. Deep learning for robust feature generation in audiovisual emotion recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 26-31, IEEE Xplore Press, Vancouver, BC, Canada, pp: 3687-3691. DOI: 10.1109/ICASSP.2013.6638346

Lee, J. and I. Tashev, 2015. High-level feature representation using recurrent neural network for speech emotion recognition.

Li, L., Y. Zhao, D. Jiang, Y. Zhang and F. Wang *et al*., 2013. Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) based speech emotion recognition. Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, Sept. 2-5, IEEE Xplore Press, Geneva, Switzerland, pp: 312-317. DOI: 10.1109/ACII.2013.58

Luengo, I., E. Navas and I. Hernaez, 2010. Feature analysis and evaluation for automatic emotion identification in speech. IEEE Trans. Multimedia, 12: 490-501. DOI: 10.1109/TMM.2010.2051872

Mao, Q., M. Dong, Z. Huang and Y. Zhan, 2014. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimedia, 16: 2203-2213. DOI: 10.1109/TMM.2014.2360798

Schuller, B., A. Batliner, S. Steidl and D. Seppi, 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Commun., 53: 1062-1087. DOI: 10.1016/j.specom.2011.01.011

Stuhlsatz, A., C. Meyer, F. Eyben, T. Zielke and G. Meier *et al*., 2011. Deep neural networks for acoustic emotion recognition: raising the benchmarks. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 22-27, IEEE Xplore Press, Prague, Czech Republic, pp: 5688-5691. DOI: 10.1109/ICASSP.2011.5947651

Trigeorgis, G., F. Ringeval, R. Brueckner, E. Marchi and M.A. Nicolaou *et al*., 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 20-25, IEEE Xplore Press, Shanghai, China, pp: 5200-5204. DOI: 10.1109/ICASSP.2016.7472669

Wang, Z.Q. and I. Tashev, 2017. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 5-9, IEEE Xplore Press, pp: 5150-5154. DOI: 10.1109/ICASSP.2017.7953138

Wongsuphasawat, K., D. Smilkov, J. Wexler, J. Wilson and D. Mané *et al*., 2018. Visualizing dataflow graphs of deep learning models in tensorflow. IEEE Trans. Visualizat. Comput. Graph., 24: 1-12. DOI: 10.1109/TVCG.2017.2744878

Zheng, W., J. Yu and Y. Zou, 2015. An experimental study of speech emotion recognition based on deep convolutional neural networks. Proceedings of the International Conference on Affective Computing and Intelligent Interaction, Sept. 21-24, IEEE Xplore Press, Xi'an, China, pp: 827-831. DOI: 10.1109/ACII.2015.7344669