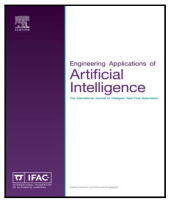
Engineering Applications of Artificial Intelligence 94 (2020) 103775

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Emotion recognition using speech and neural structured learning to facilitate edge intelligence

Md. Zia Uddin *, Erik G. Nilsson

SINTEF Digital, Forskningsveien 1, Oslo, Norway

ARTICLE INFO

Keywords:
Audio
Emotion
MFCC
LDA
NSL

ABSTRACT

Emotions are quite important in our daily communications and recent years have witnessed a lot of research works to develop reliable emotion recognition systems based on various types data sources such as audio and video. Since there is no apparently visual information of human faces, emotion analysis based on only audio data is a very challenging task. In this work, a novel emotion recognition is proposed based on robust features and machine learning from audio speech. For a person independent emotion recognition system, audio data is used as input to the system from which, Mel Frequency Cepstrum Coefficients (MFCC) are calculated as features. The MFCC features are then followed by discriminant analysis to minimize the inner-class scatterings while maximizing the inter-class scatterings. The robust discriminant features are then applied with an efficient and fast deep learning approach Neural Structured Learning (NSL) for emotion training and recognition. The proposed approach of combining MFCC, discriminant analysis and NSL generated superior recognition rates compared to other traditional approaches such as MFCC-DBN, MFCC-CNN, and MFCC-RNN during the experiments on an emotion dataset of audio speeches. The system can be adopted in smart environments such as homes or clinics to provide affective healthcare. Since NSL is fast and easy to implement, it can be tried on edge devices with limited datasets collected from edge sensors. Hence, we can push the decision-making step towards where data resides rather than conventionally processing of data and making decisions from far away of the data sources. The proposed approach can be applied in different practical applications such as understanding peoples' emotions in their daily life and stress from the voice of the pilots or air traffic controllers in air traffic management systems.

1. Introduction

Human computer interaction (HCI) is getting considerable attentions from lots of researchers due to its practical applications in ubiquitous systems (Hassan et al., 2019; Yang et al., 2020; Pace et al., 2019; Gravina and Fortino, 2016; Zhang et al., 2018). For instance, adopting HCI systems in a ubiquitous healthcare system can improve it by perceiving people's accurate emotions and proactively act to help them improving their lifestyle. Alongside other data sources, research on emotion recognition from audio is increasing day by day for healthcare in a smartly controlled environment. Speech is a natural way of humans to communicate each other in daily life. In affective computing research, speech has a vital role in promoting harmonious HCI systems and emotion recognition from speech is the first step. However, due to the lack of an exact definition of emotion, robust emotion recognition from audio speech seems to be quite complex. Hence, it demands a lot of research to solve the challenging problems beneath the audio-based emotion recognition (Sonmez and Varol, 2019).

Sound is a wave of pressure arises out of the vibration of molecules in a substance. The sound waves from a single source usually spread in all directions, eventually creates physical pressure on our ears. These waves are interpreted as electrical signals once transmitted to neurons (Tarunika et al., 2018). When a sound wave is generated from its source, it also vibrates the particles such as solid, liquid, and gas in its environment due to the its energy. Sound waves always need a material form (i.e., solid, liquid, or gas) in an environment to travel. Sound waves can be divided in to three categories based on their frequencies: below 20, between 20 and 20 000 Hz, and more than 20 000 Hz. Among which, sound waves in the middle category are human audible sound waves. These waves can be generated in various ways (e.g., musical instruments and vocal cords). Sound waves less than 20 Hz are called infrasonic sound waves. For example, earthquake waves. Sound waves of more than 20 000 Hz are called ultrasonic sound waves. Sound waves are used in different ways in science and technology such as ultrasound devices are used for imaging internal organs. High frequency ultrasonic waves are also used to break up stones in the kidneys.

Speech signal carries feelings and intentions of the speaker (Zhao et al., 2018). Speech signal analysis can be done in both time and

* Corresponding author.
E-mail addresses: zia.uddin@sintef.no (M.Z. Uddin), erik.g.nilsson@sintef.no (E.G. Nilsson).

https://doi.org/10.1016/j.engappai.2020.103775
Received 13 January 2020; Received in revised form 6 June 2020; Accepted 17 June 2020
Available online 24 June 2020

frequency domains to obtain features to model underlying events (e.g., speaker, meaning of the speech, and emotion recognition) in the signals. Hence, original speech signal and corresponding spectrum diagram can be explored for robust emotion recognition for both the domains. In Trigeorgis et al. (2016), the speech signal in time domain was used as input and combined with machine learning model for emotion recognition. In Sivanagaraja et al. (2017), the authors simultaneously applied original speech, multiscale, and multi-frequency signals to predict different emotions. In audio speech signal, the waveform characteristics vary irregularly. Hence, typical digital signal processing techniques are typically not directly utilized as audio signals for a speech are also usually continuous. However, they can be regarded as short-term stationary and then can be analysed in the frequency domain. While studying the affective identification of speech information, the typical approach is to first use the raw audio signal processing and then followed by learning the extracted features with some machine learning models, for comprehensive pattern recognition or event prediction. Spectrogram analysis of the speech signal is also very common for speech pattern recognition. In that case, the speech signal is windowed to small chunks and then divided into narrowband and broadband spectrum (Loweimi, 2016). Emotion recognition from speech signal based on spectrum may contribute much in feature engineering process. For instance, the authors in Fujimoto (2017), Sivanagaraja et al. (2017) and Le and Provost (2013) studied spectrogram with deep learning to extract features from the spectrogram of audio speech for emotion recognition.

Deep learning algorithms have been getting huge attention by pattern recognition and artificial intelligence researchers these days (Fujimoto, 2017; Sivanagaraja et al., 2017; Le and Provost, 2013; Hinton et al., 2006; Fischer, 2014; Asl et al., 2008; Uddin, 2016; Uddin et al., 2017; Li et al., 2008; Wang et al., 2012; Yang et al., 2012; Uddin et al., 2020). Deep neural network is typically better than the conventional neural networks. However, they often result in overfitting problem and take much time during training. Deep Belief Network (DBN) was a pioneer deep learning approach which utilizes Restricted Boltzmann Machines (RBMs) for training (Hinton et al., 2006). Use of RBM makes DBN faster than typical neural network (Fischer, 2014). Later, Convolutional Neural Networks (CNN) became very popular because of its improved discriminative power compared to DBN. A typical CNN algorithm consists of convolution, pooling, tangent squashing, rectifier, and normalization. CNN consists of feature extractions and some convolutional stacks to create a progressive hierarchy of useful features, especially effective for image processing tasks (Uddin et al., 2017). Basically, it follows a hierarchical neural network structure where convolutional layers are followed by subsampling layers. Finally, they are followed by fully connected layers that are identical to typical multilayer perceptron-based neural network. CNN-based deep learning approaches are very much used in visual scenery-based applications e.g., object detection in a large image achieve. Though CNN is used for many applications such as computer vision, most of the analysis of temporal events in time-sequential applications are adopt as Long Short-Term Memory (LSTM)-based Recurrent Neural Networks (RNNs) (Uddin et al., 2020). Hence, RNNs have become very popular for time-sequential event analysis. Besides, it can provide better discriminative power over DBN and CNN so far, for sequence-based pattern analysis. Amongst the available deep learning tools, TensorFlow is one of the famous tools for deep learning-based event modelling tasks such as classification, prediction, and perception. Very recently, Google has introduced Neural Structured Learning (NSL) (Anon, 2020), an open-source framework to learn neural networks. NSL can be utilized to construct robust models for in a wide range of research fields such as vision, natural language processing, and prediction in general. It can be applied for training deep neural networks by leveraging structured signals with feature inputs. An NSL algorithm basically implements neural graph learning to train neural networks with the help of graphs and structured data (Bui et al., 2018). The graphs can be obtained from

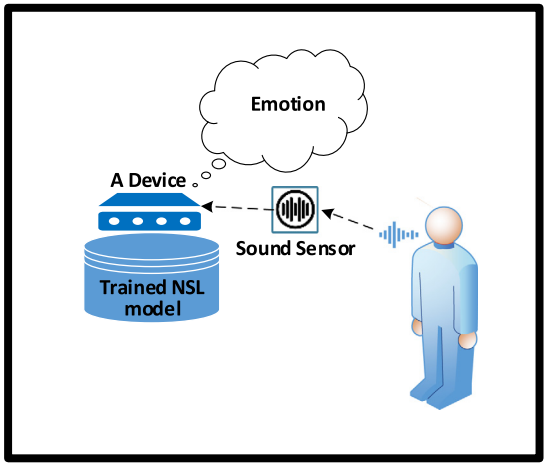


Fig. 1. A schematic picture of audio-based emotion recognition for human machine interaction in a room.

different sources e.g., knowledge graphs, medical records, genomic data, and multimodal relations. Besides, NSL generalizes to adversarial learning (Aghdam et al., 2017). Structured data usually contains good relational information among the data samples. During the training process of a deep learning model, leveraging the structured signals provide supports to obtain better model accuracy. The structured signals are primarily applied to regularize the training of a neural network by driving the model towards learning as much as accurate predictions along with maintaining the input structural similarity. Thus, training with structured signals can lead to more robust deep learning models. Hence, NSL is adopted in this work from audio speech for better emotion recognition than the traditional approaches. Besides, the approach seems to be fast and accurate enough to be applied on edge devices (e.g., Raspberry Pi®) for smart audio-based emotion recognition system with a limited number of training samples. Fig. 1 shows a schematic setup of an audio-based emotion recognition in a smart room.

In this work, a novel emotion recognition method is proposed combining NSL with the Mel Frequency Cepstrum Coefficients (MFCC)-based robust features obtained from audio speech. The MFCC features are first extracted from the raw speech data and then followed by discriminant analysis to represent a feature space highlighting to minimize the inner-class discrimination while maximizing the inter-class discrimination of different emotions. For the person independent emotion recognition modelling, the robust features are followed by a combination of neural graph learning of structured data as well as adversarial learning (i.e., NSL). The proposed emotion recognition method combining MFCC-based robust features, discriminant analysis, and deep learning by NSL was compared with the traditional deep learning approaches such as DBN, CNN, or RNN where it showed the superior recognition performance over all the conventional methods. Since the approach is fast and robust on small datasets, it can be also adopted in various smartly controlled environments such as edge devices in smart homes or clinics to provide better affective healthcare. The proposed system can be adopted in various practical applications. For instance, such systems can be applied in a smart home for understanding the emotions in the daily life of people (e.g., elderly in smart old homes or clinics) or for predicting the mental stress of the pilots and air traffic controllers in the air traffic management systems.

2. Methods

In the proposed emotion recognition method, the audio sensor data of all emotions is acquired for feature extraction and then applied for training an NSL model. For recognition process, an edge device obtains the features from a small chunk of audio speech and apply

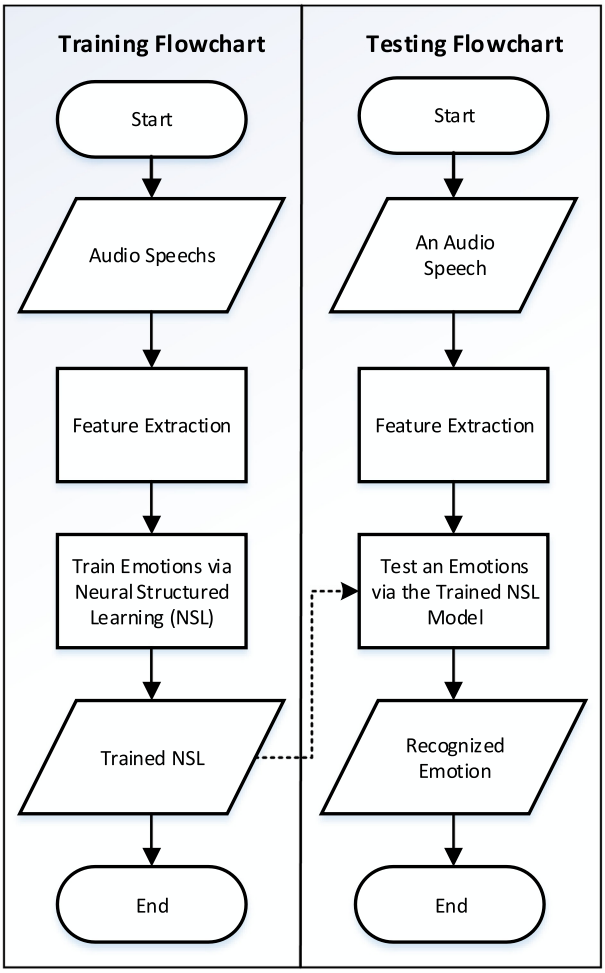


Fig. 2. Training and testing flowcharts of the proposed system.

on the trained model for emotion recognition. Fig. 2 shows the basic architecture of the proposed method from signal data collection to the recognition process via feature extraction and training.

2.1. Signal pre-processing

A sound signal is an electrical form of sound. An analog sound signal is the same copy of the sound whereas a digital sound signal is a numerical form derived from the analog sound followed by sampling and transforming them into digital ranging between 1 and 0, as shown in Fig. 3. Speech signals are basically complex signals in different frequencies where the spectrum analysis is done using a spectrogram (Uddin et al., 2018; Fleury et al., 2008; Li et al., 2010, 2011, 2012; Popescu et al., 2008; Popescu and Mahnot, 2009; Vacher et al., 2011; Zhuang et al., 2009). The Fast Fourier Transform (FFT) is most popular for spectrogram analysis in this kind of signals.

FFT is applied to a window of signals to convert a sound signal from the time domain to the frequency domain. FFT is often used to measure of the frequencies of a signal using spectrogram. Spectrogram shows the intensity of vibrations based on frequencies. FFT can adopt a fast algorithm to reduce the time complication in the measurement of Discrete Fourier Transform (DFT). This transformation is done as

$$Q_k = \sum_{n=0}^{(N-1)} Q_n e^{-i*\left(\frac{2\pi kn}{N}\right)} \qquad k = 0, 1, 2, \ldots, (N-1) \tag{1}$$

$$e^{iQ} = cosQ + isinQ \tag{2}$$

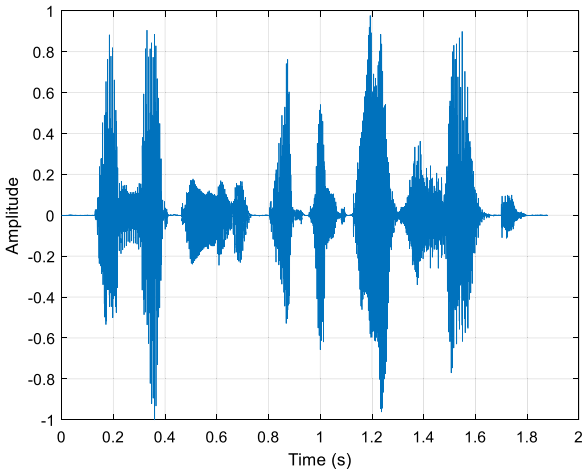


Fig. 3. A sample anger audio file from the experimental dataset.

where k represents the subsequent frequency element, N the number of samples, i the square root of (1), Q the sampled signal data, and n the index of the subsequent sample to be processed.

2.2. Feature extraction with Mel Frequency Cepstrum Coefficients (MFCC)

Mel Frequency Cepstrum Coefficients (MFCC) (Zhang et al., 2018) are coefficients of a short-time windowed signal that is obtained through FFT. MFCC provides better results than the operations in time domain. MFCC utilizes the Mel scale based on the sensitivity of the human ear. MFCC is very popular and often used for feature extraction in the frequency domain for different sound-based applications (Sonmez and Varol, 2019).

Human ears usually pick up sound frequencies until 1000 Hz in Mel scale. A triangular filter is applied in the Mel spectrum where the bandwidth varies according to the Mel scale. A normal frequency E is converted to Mel frequency F as

$$F = 2595\, log_{10}\left(\frac{E}{100} + 1\right). \tag{3}$$

Furthermore, DFT coefficients are explored according to the amplitude frequency response of the Mel filter bank. The amplitude spectrum of the signal is distributed along the Mel scale. The spectrum contains equal intervals and multiplied by the triangular filter. Then, the logarithm of the remaining energy is calculated. As the logarithm of the Mel spectrum coefficients are real numbers, the time domain values can be returned using the discrete cosine transform. The coefficients obtained via this process are called MFCC and can be represented as

$$M_n = \sum_{k=1}^{L} \left(\log S_k\right) \cos\left\lfloor n\left(k - \frac{1}{2}\right)\frac{\pi}{L}\right\rfloor \tag{4}$$

where S_k represents the Mel spectrum coefficients. Fig. 4 shows some MFCC for a sample audio speech from the audio speech clip shown in Fig. 3.

2.3. Discriminant Analysis (DA) on MFCC features

For visualization of the MFCC features, linear discriminant analysis (LDA) was adopted. LDA basically based on an eigenvalue resolution problem that tries to minimize the inner-class scatterings while maximizing the inter-class scatterings. The formulas for the within, C_W and between, C_B scatter matrix can be represented as follows:

$$C_W = \sum_{i=1}^{u} \sum_{m_k \in C_i} (m_k - \tilde{m}_i)(m_k - \tilde{m}_i)^T, \tag{5}$$

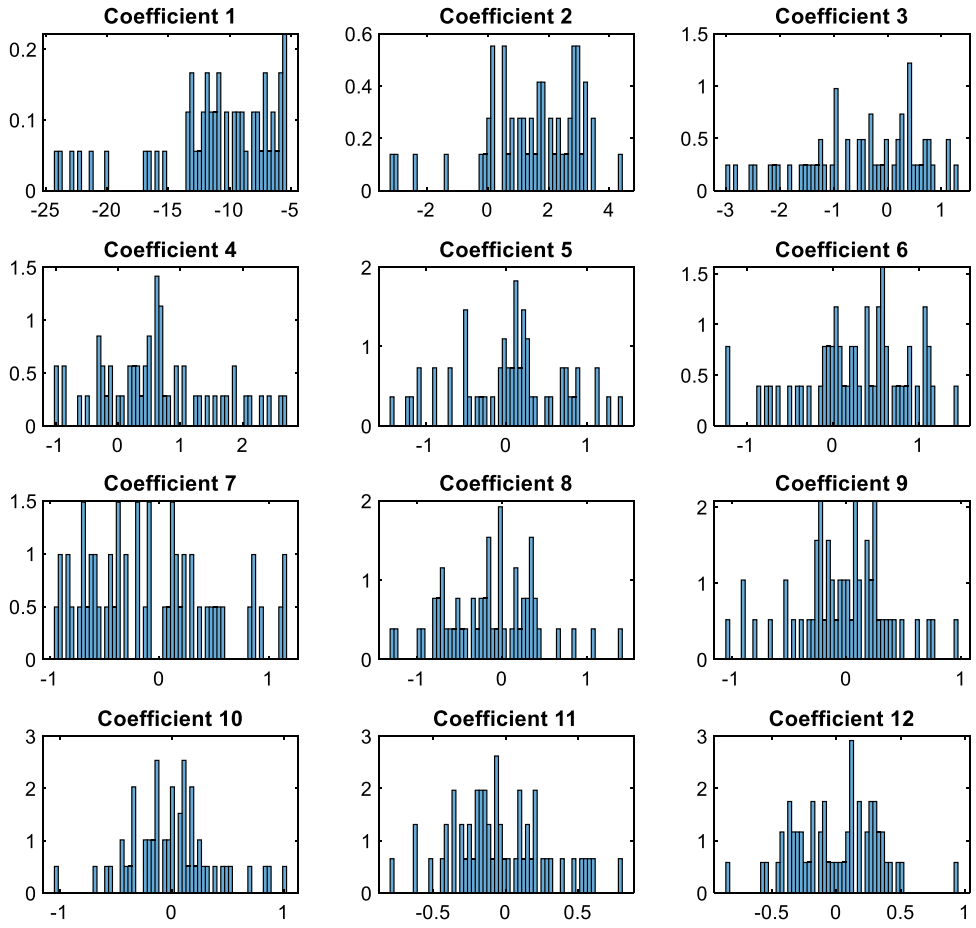


Fig. 4. MFCC co-efficients of a sample audio speech.

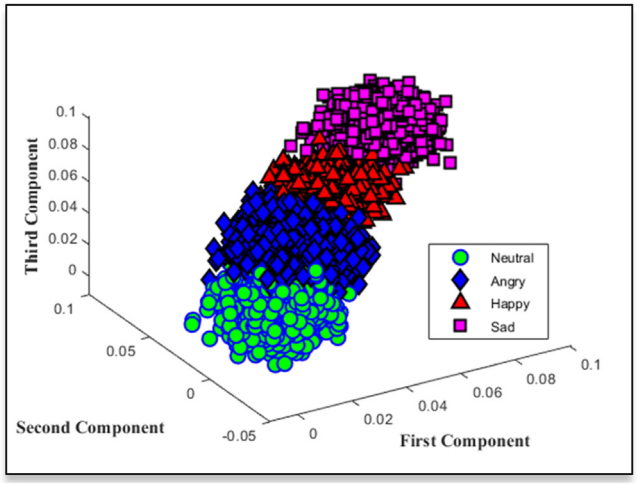


Fig. 5. 3-D plot after LDA on the audio speeches of four emotions.

$$C_B = \sum_{i=1}^{u} N_i(\tilde{m}_i - m_j)(\tilde{m}_i - m_j)^T \tag{6}$$

where u represents the number of classes, N_i the number samples in class C_i, m_k the MFCC feature vectors from all feature vectors M, m_i the mean of class i, and m_j the mean of all feature vectors.

The optimal discrimination matrix can be obtained from the maximization of the ratio between the determinant of the between-class C_B and within-class C_W scatter matrix of MFCC features as

$$J(W) = \frac{|W^T C_B W|}{|W^T C_B W|} \tag{7}$$

where W is the set of discriminant vectors of C_B and C_W corresponding to the $(u-1)$ largest eigenvalues. Thus, the discriminant ratio can be obtained by solving the following eigenvalue problem as

$$C_B W = \Lambda C_W W \tag{8}$$

where Λ represents the eigenvalue matrix. Fig. 5 shows a 3-D plot of the MFCC features in LDA features space, shows good separation among the samples of different classes.

2.4. Emotion modelling

Neural Structured Learning (NSL) is a deep learning approach that focuses on training the neural networks by leveraging structured signals along with the input features. As introduced in Bui et al. (2018), the structured signals are used to regularize the training of a neural network that forces to learn accurate predictions with the help of minimizing supervised loss. At the same time, it tries to maintain the input structural similarity with the help of minimizing the neighbour loss. The approach is very generic and can be utilized on any arbitrary neural architectures such as typical Feed-forward neural networks, CNN and RNN. Fig. 6 shows the basic structure of an NSL with features as input combined with structured signals.

The generalized neighbour loss equation can be represented as

$$loss = \sum_{k=0}^{W} L(y_i, \hat{y}_i) + \propto \sum_{k=0}^{W} L\left(y_i, x_i, N\left(x_i\right)\right). \tag{9}$$

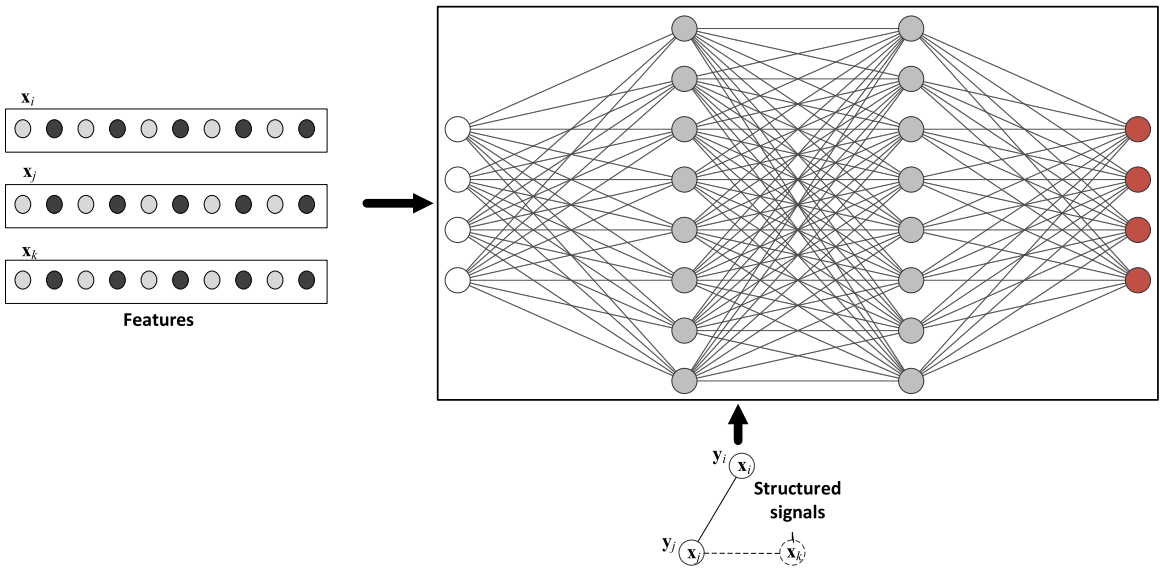


Fig. 6. A neural network with features as input combined with structured signals.

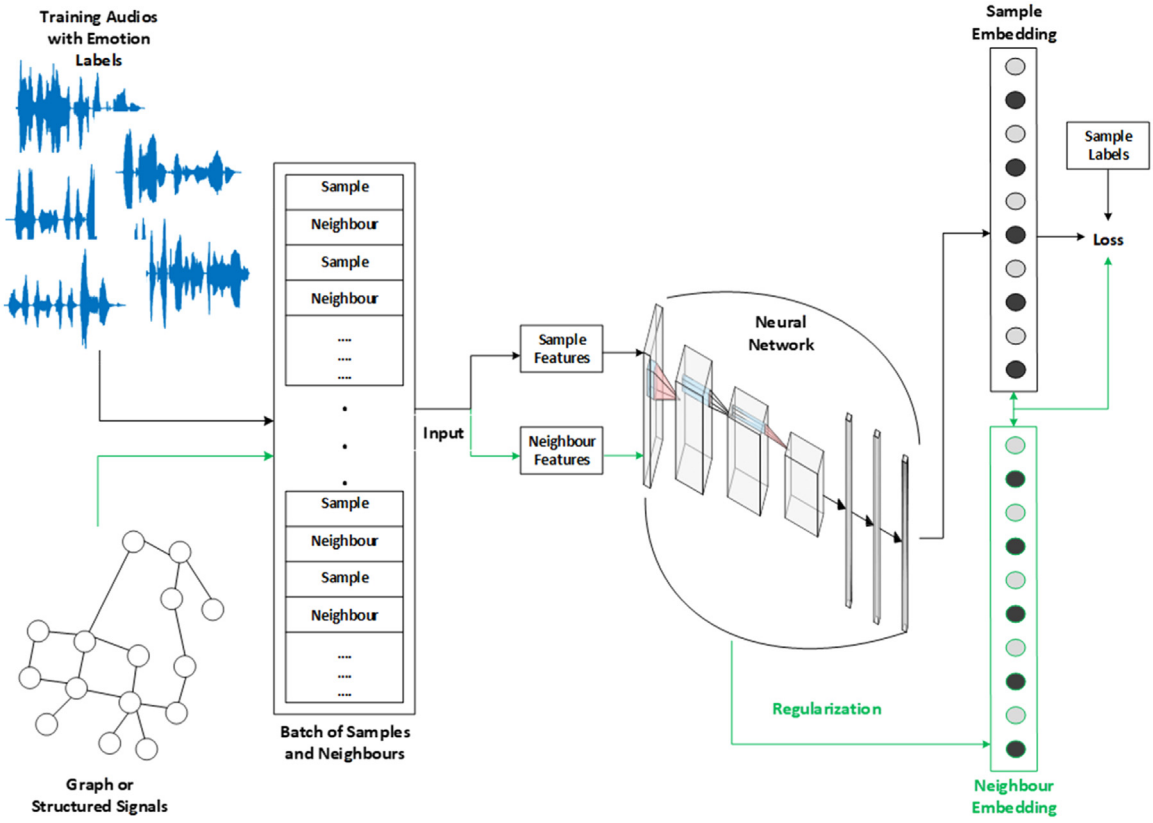


Fig. 7. A neural structured learning model designed from combining audio signals, corresponding graph of the signals, and CNN. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

According to the above equation, NSL basically generalizes the network using two different ways. First one is by using neural graph learning (NGL) where the data points are represented by a graph. Graph-based method provides a versatile, scalable, and effective solution to solve a wide range of problems. It constructs a graph over labelled as well as unlabelled data. Besides, Graph is a natural way to describe the relationships of data elements where connections in the graph connect semantically similar data. If there is connection in the data and its neighbours, the edge weights of the nodes in the graph reflect strength of the similarities. Thus, NGL refines the node labels in the graph by collecting and combining the information from neighbours and propagate the labels to the neighbours. NGL methods quickly converge and hence, can be applied in small dataset but also, can be scaled for large datasets consisting of a large label space (Bui et al., 2018). NGL architectures are basically inspired by the objective function of label propagation which enforces similarity between nodes in the graphs. This results in graph-augmented training for a wide range of neural networks and can be used in both inductive and transductive

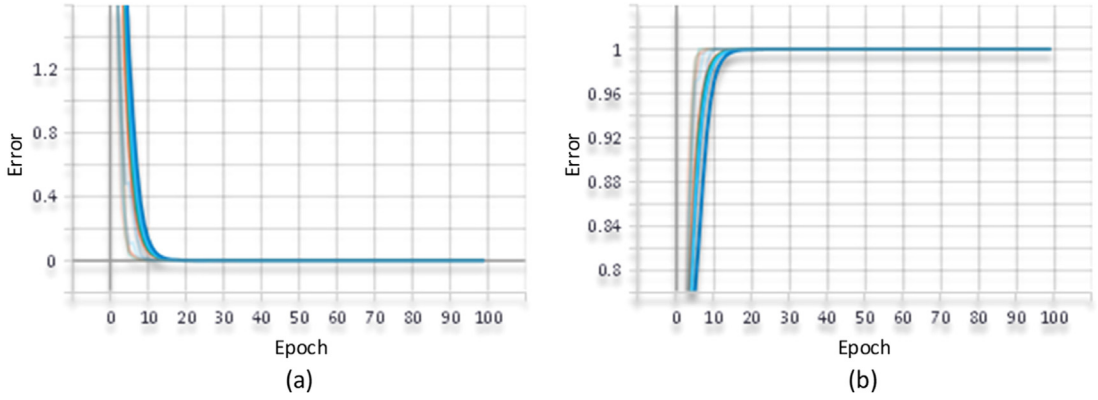


Fig. 8. (a) Loss and (b) accuracy of the ANN model for 100 epochs.

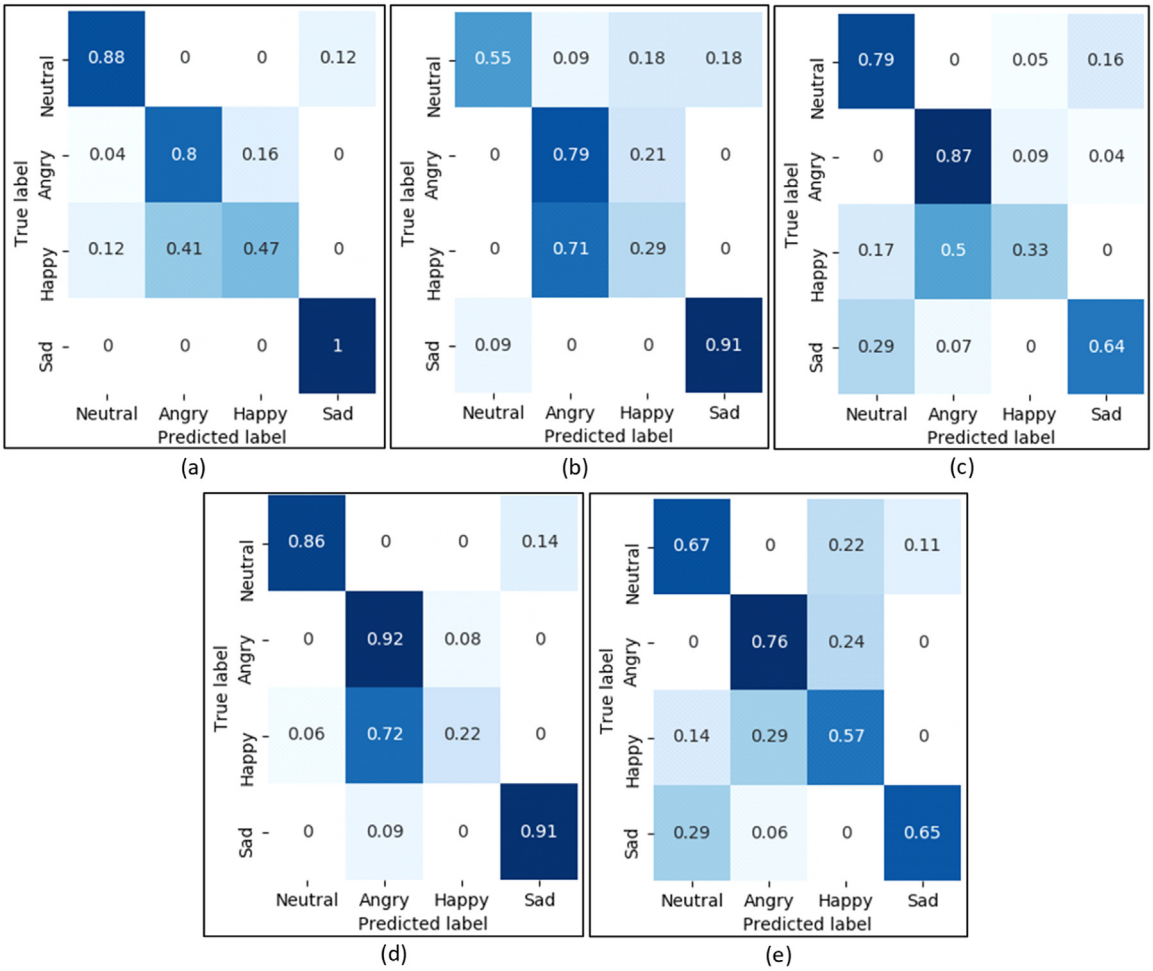


Fig. 9. Confusion matrices of five folds from (a) to (e) using ANN.

learning. The NGL framework can also handle multiple forms of graphs such as natural or developed from the knowledge bases in data. Thus, NGL framework can be adopted for organizing the relations in data based on the neighbourhoods.

The second one is by using adversarial learning (Aghdam et al., 2017) if neighbours are induced by the adversarial perturbation.

Traditional machine learning models including state-of-the-art neural networks are usually vulnerable to adversarial examples. More clearly, these models basically misclassify examples that are a little different from correctly classified examples drawn from the given data. In many cases, a wide range of different models with distinguished architectures trained on different subsets of the training data misclassify even the same adversarial example. Hence, this suggests that adversarial examples indicate blind spots in our typical adversarial training methods. The cause of these adversarial examples may be due to extreme nonlinearity of deep neural networks. Perhaps, due to insufficient model averaging or regularization of the purely supervised learning problems. Thus, in this work, linear behaviour of the models in high-dimensional spaces seems to be sufficient to cause adversarial examples. This view makes it possible to design a fast method of generating adversarial examples that makes adversarial training more

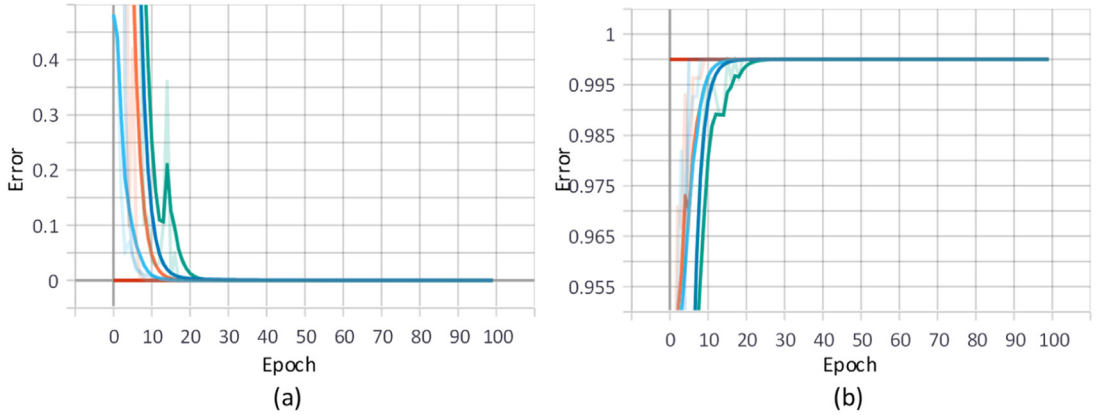


Fig. 10. (a) Loss and (b) accuracy of the CNN model for 100 epochs.

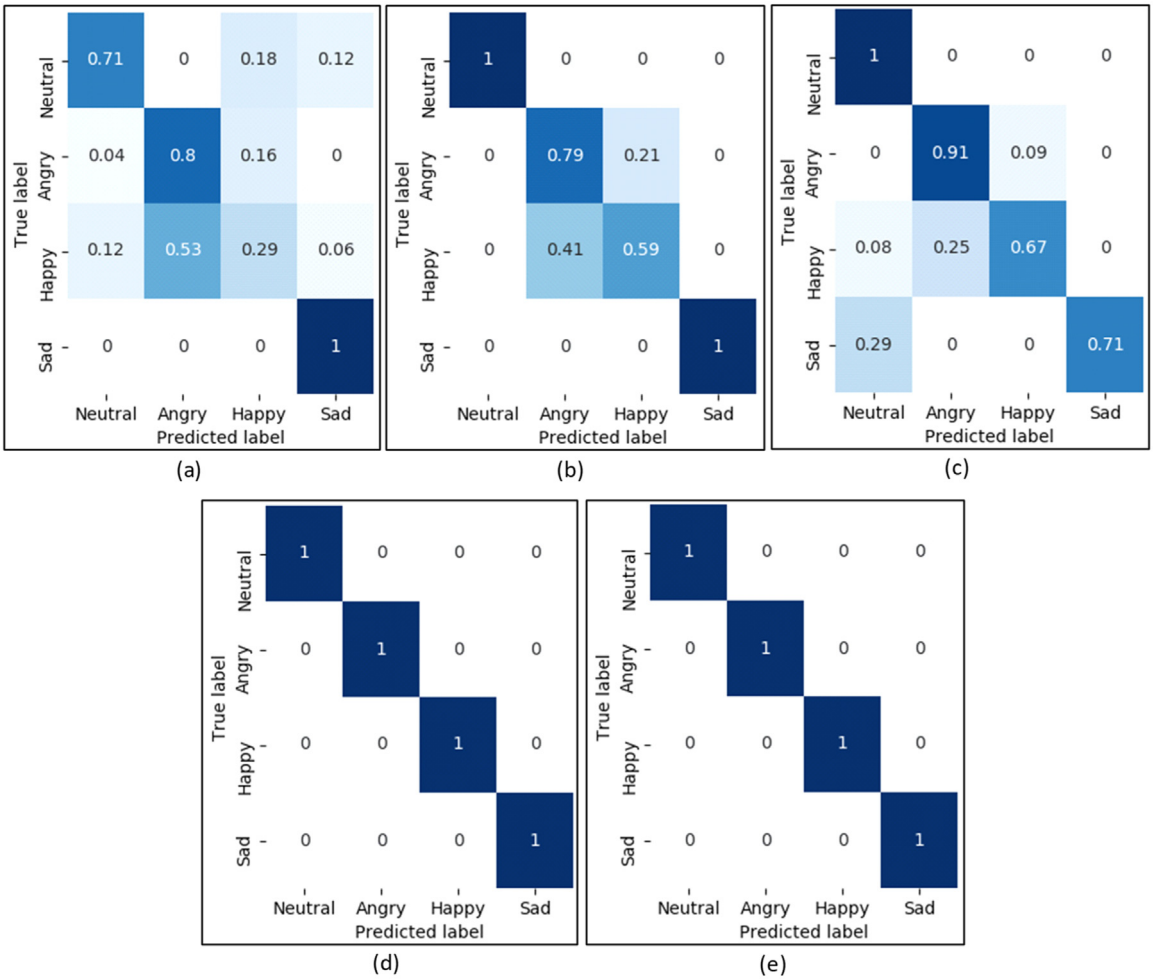


Fig. 11. Confusion matrices of five folds from (a) to (e) using CNN.

practical than before. Besides, general regularization strategies in common deep learning models such as dropout, pretraining, and model averaging do not usually reduce the models' vulnerability to adversarial examples.

The overall workflow for NSL is depicted in Fig. 7 where the black arrows indicate the conventional process of training while the green arrows show the workflow in NSL to take the advantage of structured signals. In NSL, the training samples are augmented to include structured signals and if structured signals are not explicitly provided, they are either constructed or induced by adversarial learning. Later, the augmented training samples including original and neighbouring samples are fed into the neural network for calculating their embeddings. The distance between the embedding of a sample and its neighbour is obtained and used as the neighbour loss. This process is treated as a regularization term and later added to the final loss. During the explicit neighbour-based regularization, any layer of the neural network may however be used to compute the neighbour loss. On the contrary, for induced neighbour-based regularization (i.e., adversarial), neighbour loss is calculated according to the distance between the ground truth and output prediction of the induced adversarial neighbour.

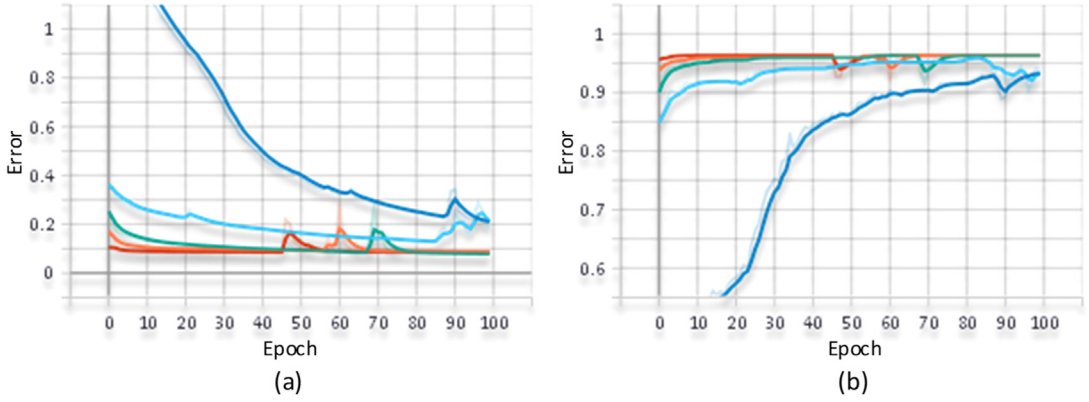


Fig. 12. (a) Loss and (b) accuracy of the LSTM model for 100 epochs.

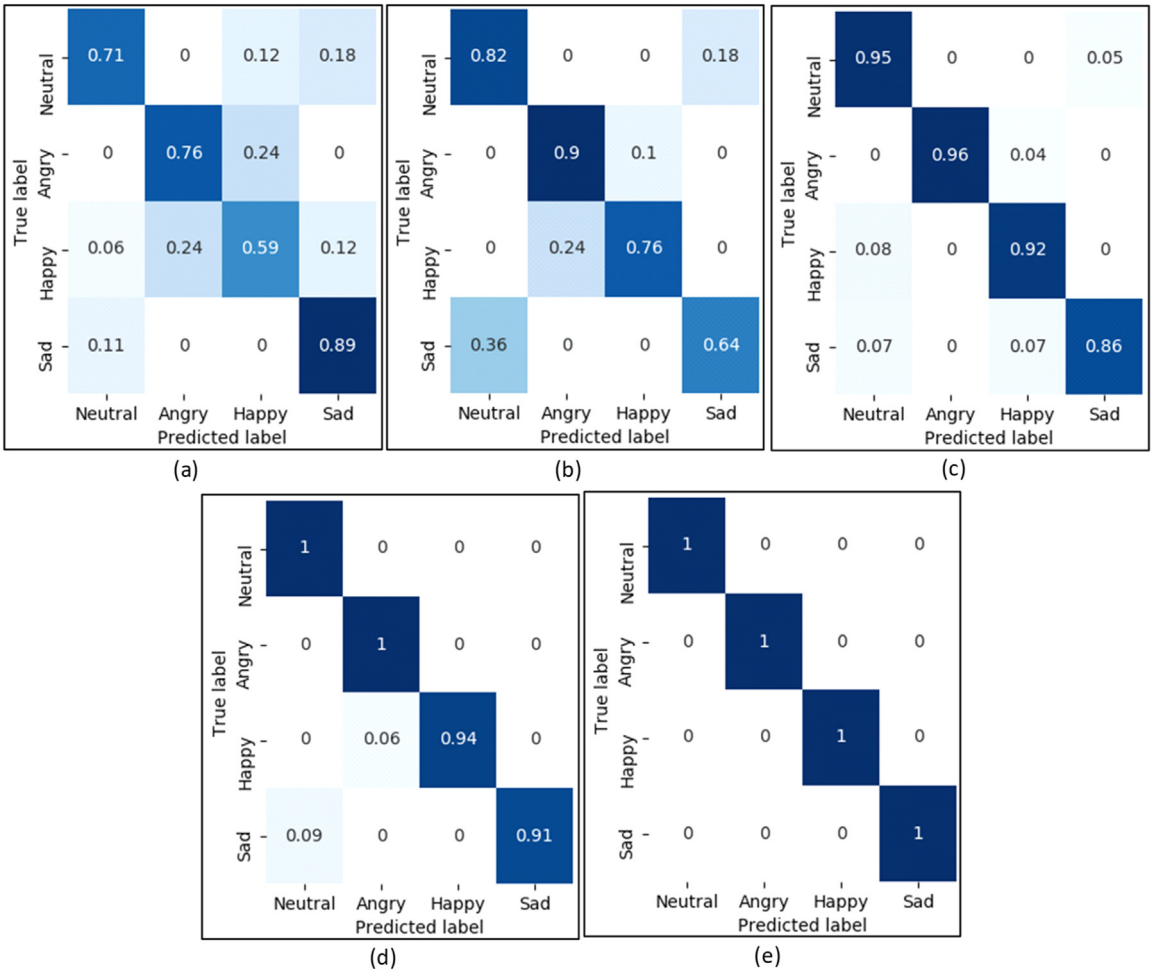


Fig. 13. Confusion matrices of five folds from (a) to (e) using LSTM.

3. Experiments and results

An audio speech database was collected for this work containing four expressions: namely neutral, angry, happy, and sad (Burkhardt et al., 2005). There were 339 audio clips where the duration of each clip was around 2 s. For the experiments, five-fold cross validation was applied. The summary of the results is shown in Table 1. To achieve a good audio quality, the recordings was done in an anechoic chamber of the Technical University Berlin, Technical Acoustics Department. Each audio consists of a sampling frequency of 48 kHz, which later was made to 16 kHz by down sampling. The actors stood before the microphone and spoke in the direction of the microphone from about the distance of 30 cm. Three phoneticians supervised the recording of each session of the users where two of them were giving instructions and one was monitoring the functions of the recording equipment. For some emotions, there were several variants of the same emotions. The actors were also instructed about not to unnecessarily shout to express anger or to avoid whispering while expressing their anxiety during the audio recording. This was done due to maintain and analyse the voice quality. In addition, the recording levels were adjusted between very loud speech and very quite speech. Finally, twenty subjects evaluated the speech data with satisfactory results, ensures the emotional quality

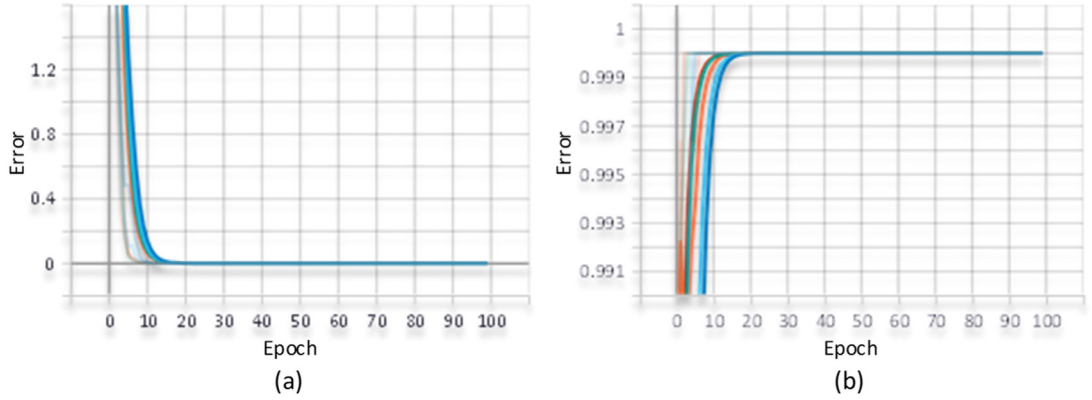


Fig. 14. (a) Loss and (b) accuracy of the NSL model for 100 epochs.

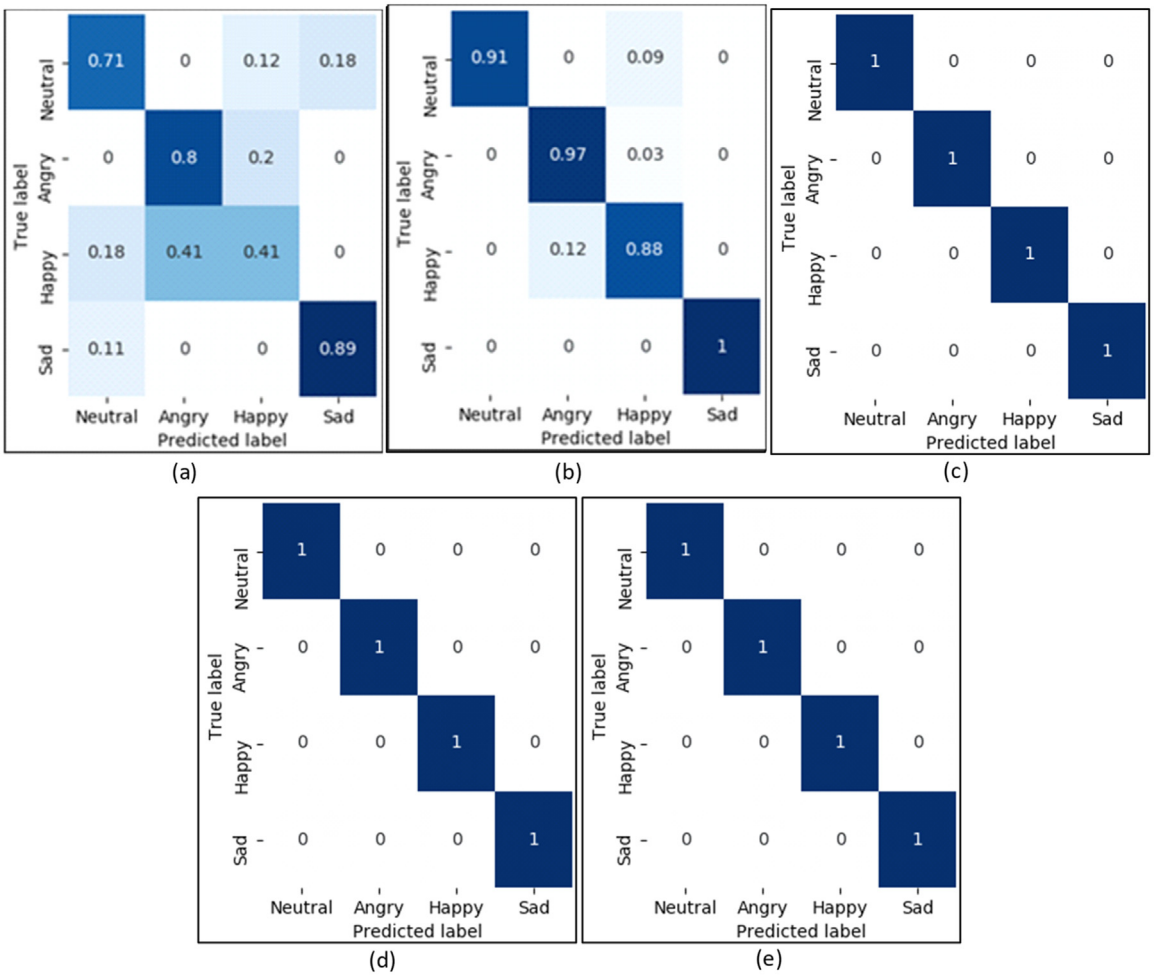


Fig. 15. Confusion matrices of five folds from (a) to (e) using NSL.

and naturalness of the utterances of the audio data. The experiments were done on a mobile workstation with the configurations of CPU as Intel Core i7-7700HQ, memory: 32 GB, and 2 graphics processing units (GPUs): Intel HD 630 and NVIDIA Quadro M2200. The deep learning platform TensorFlow version 2.2.0 was adopted to apply deep learning algorithms.

We first started with the typical artificial feed forward neural network (i.e., ANN)-based experiments that achieved the mean recognition rate of the five folds is 0.69, the least recognition performance. The mean recognition of the emotions from five folds are 0.75, 0.82, 0.37, and 0.82. The ANN model consisted of three hidden layers with 500, 200, and 100 neurons, respectively. Though the approach showed more than average performance for neutral, angry, and sad expression but it failed for happy emotion by achieving less than 40% mean recall rate. Figs. 8 and 9 show the ANN model characteristics based on epochs and confusion matrices for different folds, respectively.

Then, the experiments were continued to CNN-based ones, better approach than typical ANN. CNNS has had good results over last some years in a variety of fields of pattern recognition. A very important aspect of CNN is that CNNs have reduced number of parameters than ANN. This aspect has attracted a lot of researchers to solve complex tasks, which are apparently not possible using classic ANN. Also, CNN

Layer (type)	Output Shape	Param #
flatten (Flatten)	(None, 7722)	0
dense (Dense)	(None, 128)	988544
dense_1 (Dense)	(None, 64)	8256
dense_2 (Dense)	(None, 4)	260
Total params: 997,060		

Fig. 16. A sample NSL model summary.

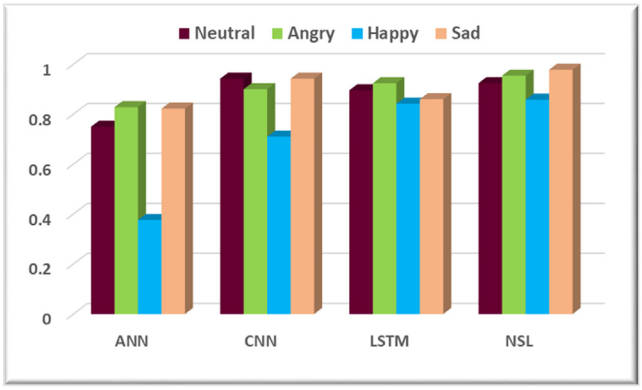


Fig. 17. Mean of the recalls of the emotions using four different approaches.

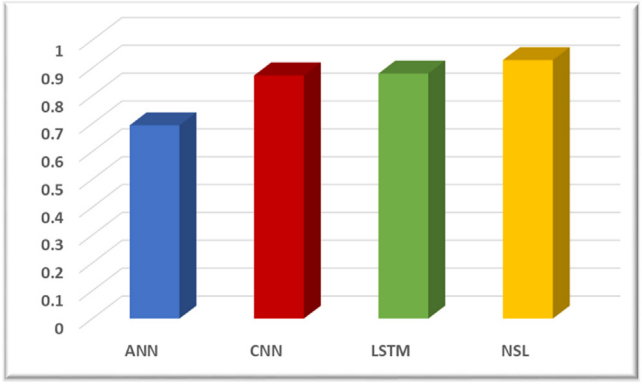


Fig. 18. Overall mean recognition rates using four different approaches.

Table 1
The mean recall rates of the emotions for all the folds using different approaches.

Emotion/Model	ANN	CNN	LSTM	NSL
Neutral	0.75	0.94	0.89	0.92
Angry	0.82	0.90	0.92	0.95
Happy	0.37	0.71	0.84	0.85
Sad	0.82	0.94	0.8	0.97
Mean	**0.69**	**0.87**	**0.88**	**0.93**

represents features spatially independent features. A CNN basically consists of multiple layers: convolutional, pooling, and fully-connected layer. Among which, the convolutional and fully-connected layers have parameters whereas pooling layers do not have any. We applied a CNN consisting of consists of three convolution layers, three pooling layers, and one fully connected layer. The CNN-based approach achieved the mean recognition rate for the five folds as 0.87, a better performance than ANN. The mean recognition of the five folds are 0.94, 0.90, 0.71, and 0.94 for all the emotions, respectively. Figs. 10 and 11 depicts the CNN model characteristics based on epochs and confusion matrices for different folds, respectively.

Furthermore, LSTM-based experiments were done. LSTM units are basically recurrent modules with the ability to learn the data in sequence. LSTM units consist of hidden states with gating functions. A typical LSTM unit is represented by three gates: input, forget, and the output gate. The input gate basically contains weight matrix, bias, and a logistic function. Forget gates usually allow the neural network to forget the memory, where applicable. At last, the output gates determine the information to transfer in between the hidden states. The output unit takes the decision of forgetting hidden states or update hidden states with the memory. We used 50 and 20 LSTM units for two layers to model the emotions in the audio speech data. The LSTM-based experiments achieved the mean recognition rate 0.88, further better recognition performance than ANN and CNN. The mean recognition rate of all the folds for the emotions are 0.94, 0.90, 0.71, and 0.94, respectively. Figs. 12 and 13 illustrate the LSTM-based RNN model characteristics based on epochs and confusion matrices for different folds, respectively.

Finally, we proceeded to the proposed NSL-based experiments for emotion modelling and recognition. NSL is a quite new learning paradigm than other existing deep learning algorithms such as LSTM and CNN. One advantage of NSL is to train the networks by leveraging structured signals in addition to the available feature inputs to the model. Two kinds of structures are incorporated in NSL: graph and adversarial perturbation. Structured signals can basically represent the similarity among labelled or unlabelled samples. Thus, leveraging structured signals during training process harnesses both kind of data, which can improve model accuracy. NSL obtained the mean recognition rate of 0.93, the highest among all the approaches. The mean recognition rate of all the folds for the emotions are 0.92, 0.95, 0.85, and 0.97, respectively. Figs. 14 and 15 represents the NSL model characteristics based on epochs and confusion matrices for different folds, respectively. Fig. 16 shows the summary of a simple and small NSL model parameters used in this work, indicating good scope of its implementation in low performance devices. To show a clear picture of the performances of different approaches, Fig. 17 shows the mean of the recalls of the emotions using four different approaches and Fig. 18 the overall mean recognition rates using four different approaches where NSL shows its leading performance over others. The training time for each fold consisting of features from 271 audio clips took on an average of 37.08 s and the testing of features from an audio clip from a testing fold took 0.006 s only, indicates the suitability of the implementation of the proposed approach in real-time.

4. Concluding remarks

A basic audio-based emotion recognition system consists of three major parts: acquisition of audio signals, feature processing that tries to obtain distinguishable robust features for each emotion so that each expression can be represented as much different from each other, and emotion recognition that recognizes emotion by applying robust features on a strong pre-trained expression model. In this work, we have proposed a novel approach for emotion recognition from audio speech signals where MFCC features are tried with discriminant analysis and a state-of-the-art deep learning approach i.e., neural structured learning based on neural graph learning and adversarial learning. The proposed emotion recognition approach was compared with traditional approaches where it showed its superiority over others. The proposed

system could be adopted to contribute in any smartly controlled environment for audio-based emotional healthcare. The system can also be tried on edge devices with a limited audio-based emotion dataset collected from sound sensors in edges. Furthermore, the approach can be extended in future with more efficient deep learning methods such as applying LSTM or CNN under NSL structure (i.e., neural graph and adversarial learning) to develop more robust emotion recognition model.

CRediT authorship contribution statement

Md. Zia Uddin: Conceptualization, Methodology, Software, Writing. **Erik G. Nilsson:** Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work has been funded by an internal research project of SINTEF, Norway.

References

Aghdam, H.H., Heravi, E.J., Puig, D., 2017. Explaining adversarial examples by local properties of convolutional neural networks. In: Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.

Anon, 2020. Neural structured learning: Training with structured signals, Tensorflow. [Online]. Available: https://https://www.tensorflow.org/neural_structured_learning/. [Accessed: 06-June-2020].

Asl, B.M., Setarehdan, S.K., Mohebbi, M., 2008. Support vector mashine-based arrhythmia classification using reduced features of heart rate variability signal. Artif. Intell. Med. 44, 51–64.

Bui, T.D., Ravi, S., Ramavajjala, V., 2018. Neural graph learning. In: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F., Weiss, B., 2005. A database of German emotional speech. In: Interspeech 2005- Eurospeech, $9^{T}h$ European Conference on Speech Communication and Technology (Lisbon). pp. 1517–1520.

Fischer, A.C., 2014. Training restricted Boltzmann machines: An introduction. Pattern Recognit. 47 (1), 25–39.

Fleury, A., Noury, N., Vacher, M., Glasson, H., Seri, J.-F., 2008. Sound and speech detection and classification in a Health Smart Home. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4644–4647.

Fujimoto, M., 2017. Factored deep convolutional neural networks for noise robust speech recognition. In: Interspeech 2017.

Gravina, R., Fortino, G., 2016. Automatic methods for the detection of accelerative cardiac defense response. IEEE Trans. Affect. Comput. 7 (3), 286–298.

Hassan, M.M., Alam, M.G.R., Uddin, M.Z., Huda, S., Almogren, A., Fortino, G., 2019. Human emotion recognition using deep belief network architecture. Inf. Fusion 51, 10–18.

Hinton, G.E., Osindero, S., The, Y., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554.

Le, D., Provost, E.M., 2013. Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding.

Li, Yun, Ho, K.C., Popescu, M., 2012. A microphone array system for automatic fall detection. IEEE Trans. Biomed. Eng. 59 (5), 1291–1301.

Li, Yun., Popescu, M., Ho, K.C., Nabelek, D.P., 2011. Improving acoustic fall recognition by adaptive signal windowing. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 7589–7592.

Li, Yun, Zeng, Zhiling, Popescu, M., Ho, K.C., 2010. Acoustic fall detection using a circular microphone array. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. pp. 2242–2245.

Li, W., Zhang, Z., Liu, Z., 2008. Expandable data-driven graphical modeling of human actions based on salient postures. IEEE Trans. Circuits Syst. Video Technol. 18 (11), 1499–1510.

Loweimi, E., 2016. Emotion recognition from the speech signal by effective combination of generative and discriminative models. In: USES.

Pace, P., Aloi, G., Gravina, R., Caliciuri, G., Fortino, G., Liotta, A., 2019. An edge-based architecture to support efficient applications for healthcare industry 4.0. IEEE Trans. Ind. Inf. 15 (1), 481–489.

Popescu, M., Li, Y., Skubic, M., Rantz, M., 2008. An acoustic fall detector system that uses sound height information to reduce the false alarm rate. In: 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4628–4631.

Popescu, M., Mahnot, A., 2009. Acoustic fall detection using one-class classifiers. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 3505–3508.

Sivanagaraja, T., Ho, M.K., Khong, A.W.H., Wang, Y., 2017. End-to-end speech emotion recognition using multi-scale convolution networks. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).

5, Sonmez, Y.U., Varol, A., 2019. New trends in speech emotion recognition. In: 2019 7th International Symposium on Digital Forensics and Security (ISDFS).

Tarunika, K., Pradeeba, R.B., Aruna, P., 2018. Applying machine learning techniques for speech emotion recognition. In: 9th International Conference on Computing Communication and Networking Technologies (ICCCNT). pp. 10–12.

Trigeorgis, George, et al., 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: IEEE International Conference on Acoustics, Speech and Signal Processing.

Uddin, M.Z., 2016. A depth video-based facial expression recognition system utilizing generalized local directional deviation-based binary pattern feature discriminant analysis. Multimedia Tools Appl. 75 (12), 6871–6886.

Uddin, M.Z., Hassan, M.M., Alsanad, A., Savaglio, C., 2020. A body sensor data fusion and deep recurrent neural network-based behavior recognition approach for robust healthcare. Inf. Fusion 55, 105–115.

Uddin, M.Z., Khaksar, W., Torresen, J., 2017. Facial expression recognition using salient features and convolutional neural network. IEEE Access 5, 26146–26161.

Uddin, M., Khaksar, W., Torresen, J., 2018. Ambient sensors for Elderly care and independent living: A survey. Sensors 18 (7), 2027.

Vacher, M., Istrate, D., Portet, F., Joubert, T., Chevalier, T., Smidtas, S., Meillon, B., Lecouteux, B., Sehili, M., Chahuara, P., Meniard, S., 2011. The sweet-home project: Audio technology in smart homes to improve well-being and reliance. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 5291–5294.

Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y., 2012. Robust 3d action recognition with random occupancy patterns. In: European Conference on Computer Vision. pp. 872–885.

Yang, J., Wang, R., Guan, X., Hassan, M.M., Almogren, A., Alsanad, A., 2020. AI-enabled emotion-aware robot: The fusion of smart clothing, edge clouds and robotics. Future Gener. Comput. Syst. 102, 701–709.

Yang, X., Zhang, C., Tian, Y., 2012. Recognizing actions using depth motion mapsbased histograms of oriented gradients. In: ACM International Conference on Multimedia. pp. 1057–1060.

Zhang, Y., Gravina, R., Lu, H., Villari, M., Fortino, G., 2018. PEA: Parallel electrocardiogram-based authentication for smart healthcare systems. J. Netw. Comput. Appl. 117, 10–16.

Zhao, H., Ye, N., Wang, R., 2018. A survey on automatic emotion recognition using audio big data and deep learning architectures. In: 2018 IEEE 4th International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing, (HPSC) and IEEE International Conference on Intelligent Data and Security (IDS).

Zhuang, X., Huang, J., Potamianos, G., Hasegawa-Johnson, M., 2009. Acoustic fall detection using Gaussian mixture models and GMM supervectors. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 69–72.