Multimed Tools Appl (2019) 78:5571–5589
https://doi.org/10.1007/s11042-017-5292-7

Deep features-based speech emotion recognition for smart affective services

Abdul Malik Badshah[1] · Nasir Rahim[1] · Noor Ullah[1] ·
Jamil Ahmad[1] · Khan Muhammad[1] · Mi Young Lee[1] ·
Soonil Kwon[1] · Sung Wook Baik[1]

Received: 27 June 2017 / Revised: 29 September 2017 / Accepted: 5 October 2017 /
Published online: 31 October 2017

Abstract Emotion recognition from speech signals is an interesting research with several applications like smart healthcare, autonomous voice response systems, assessing situational seriousness by caller affective state analysis in emergency centers, and other smart affective services. In this paper, we present a study of speech emotion recognition based on the features extracted from spectrograms using a deep convolutional neural network (CNN) with rectangular kernels. Typically, CNNs have square shaped kernels and pooling operators at various layers, which are suited for 2D image data. However, in case of spectrograms, the information is encoded in a slightly different manner. Time is represented along the x-axis and y-axis shows frequency of the speech signal, whereas, the amplitude is indicated by the intensity value in the spectrogram at a particular position. To analyze speech through spectrograms, we propose rectangular kernels of varying shapes and sizes, along with max pooling in rectangular neighborhoods, to extract discriminative features. The proposed scheme effectively learns discriminative features from speech spectrograms and performs better than many state-of-the-art techniques when evaluated its performance on Emo-DB and Korean speech dataset.

Keywords Speech emotion recognition · Convolutional neural network · Spectrogram · Rectangular kernels

1 Introduction

Speech signal is the most natural, intuitive, and fastest means of interaction among humans. However, using speech signals to interact naturally with machines require a lot of efforts. Significant progress has been made in the recent years in speech recognition and speaker

✉ Sung Wook Baik
sbaik@sejong.ac.kr

[1] Digital Contents Research Institute, Sejong University, Seoul, Republic of Korea

recognition. Speech carries much more information than spoken words and speaker information. Speech emotions recognition (SER) has been an active area of research where it aims to enable speech analysis systems to recognize the affective state of the speaker [41]. In this context, researchers are striving to make machines understand our emotional states. To extract emotional state of a speaker from his/her speech, SER using discriminative features is a viable solution. However, finding effective and noticeable features for SER is a challenging task [14]. In the recent past, emotion detection from speech signals has gained much attention and is an important research area to make human machine interaction more natural.

With the advancements in technology, there is growing interest in developing affective interaction modes, giving rise to smart affective services [51]. Hence, applications of SER are increasing at a rapid pace. Bjorn Schuller et al. [43] have suggested usage of SER within automotive environments for an in-car board system where strategies can be initiated by determining the mental state of the driver to the system. SER can serve as an essential component in developing smart affective services for healthcare, surveillance, human-machine-interaction, audio forensics [27], and affective computing. For instance, it can be used as a diagnostic tool for therapist [17] and to assess situational seriousness in emergency centers by analyzing the affective state of the callers through their speech [2, 6, 50]. It can also help in automatic translation systems where identifying emotional state of the speaker plays a vital role in communication between parties. It can collect the supporting evaluation data from the emotions of pilgrims from their normal speaking within Hajj services helping in improving data visualization affecting serious decision making [20]. SER can act as an important tool in helping automatically understand people's physical interactions in different dense crowds which is difficult to do with manual methods [25, 26, 28]. Further SER can help in assessing voices of people in dense crowds for violent and aggressive behaviors prediction in surveillance streams [1, 10, 36]. It can also be used in supporting data verification scheme for input from auto sensor device for proper information gathering [5].

SER has been an active area of research where a variety of approaches have been presented. Significant work has been carried out in affective features extraction from speech signals and efficient classification. Major issue faced by SER systems is the detection of affect-oriented discriminative features to represent emotional speech signals. Recently, deep neural networks (DNNs) have been tested to extract high level features from raw speech signals which exhibited interesting and highly acceptable results. Effectively modeling speech signals for classifiers to efficiently classify emotions is a critical design issue for SER [14]. In this paper, we present a convolutional neural network architecture with rectangular kernels and modified pooling strategy to detect emotional speech. Furthermore, we also evaluate the proposed method to recognize emotions in emergency calls which are plagued with background noise and poor voice quality, making SER even more challenging. Through extensive experiments, we show that the proposed scheme significantly outperforms existing hand-crafted features based SER approaches on the two challenging datasets.

The rest of the paper is organized as follows: Section 2 presents an overview of the state-of-the-art SER approaches with their strengths and weaknesses. The proposed method is explained in Section 3, and experimental results and discussions are provided in Section 4. Section 5 concludes the paper with some future research directions.

2 Related work

Typical SER is composed of two main portions 1) a processing unit that extracts the most suitable features from speech signals and 2) a classifier to recognize the hidden emotions in speech using its features vectors. This section provides a quick overview of existing feature extraction methods and classification strategies.

Common challenges being faced by SER systems include the selection of the speech features which allow clear discrimination among distinct emotions. However, acoustic variability due to the variation of different speakers, speaking styles, speaking rates, and different sentences directly affect extracted features such as pitch and energy contour [3, 7]. To tackle this issue, one possible approach is to divide speech signals into multiple small chunks called frames and construct a feature vector for each frame. For example, building prosodic feature vector for each frame such as pitch and energy [40, 48]. Furthermore, global features can be extracted from the whole speech utterance which offer lower dimensionality details as compared to local features extracted from each frame, thereby reducing computations. Moreover, it is possible that a particular utterance has more than one emotion; each emotion corresponds to different frame of the spoken utterance. In addition to this, detecting boundaries of such frames is difficult because expression of certain emotion varies from speaker to speaker, cultural differences, and variations in environmental conditions. In literature, most experiments were conducted in monolingual emotion classification environment, where the cultural differences among speakers were ignored.

Recently, unsupervised feature learning techniques have shown improved results for automatic speech recognition system [52] and image understanding [8, 19]. Stuhlsatz et al. [45] proposed a method which yielded improved results in both weighted and un-weighted recall by using generatively pre-trained artificial neural network to construct low dimensional discriminative features vector in a multi-emotion corpora. Schmidt and Kim [42] used deep belief network for emotional music recognition. Their method aimed to learn high level features from magnitude spectra directly as compared to hand-crafted features. The authors in [12, 15, 16, 19, 21–23, 49] replaced a collection of Gaussian mixtures by single context dependent DNN using variety of large scale speech task. Wollmer et al. [48] proposed a method to analyze back and forth speech utterance-levels and used this analysis to predict emotion for a given utterance.

Different types of classifiers have been used for SER including hidden Markov model (HMM) [30, 39], Gaussian mixtures model [53], support vector machine (SVM) [33], artificial neural network [18], K-nearest neighbor [37] and many others [38]. Among these, SVM and HMM are the most widely used learning algorithm for speech related applications [30, 35, 47, 49]. However, experiments show that each classifier is domain dependent in terms of accuracy and the quality of data. Apart from single classifiers, an aggregated system of multiple classifiers has also been studied for SER for improving accuracy [32].

Accurate and efficient SER systems can substantially improve affective smart services. With the rapid adaptation of end-to-end procedures for classification tasks using deep learning algorithms, it becomes imperative to explore these hierarchical architectures for the task of SER on highly challenging datasets. The strength of these end-to-end learning methods lie in the automatic extraction of discriminative features for efficient classification for a variety of data. Dennis et al. [13] proposed a novel feature extraction method for classification of sound events. They extracted the visual signature from sound's time-frequency representation (i.e., spectrograms). They tested their method on a database consisting of 60 sound classes. They claimed a remarkable improvement over other methods in conditions with mismatch. Deng Li et al. [11] explored a layer-by-layer learning strategy of patches of speech spectrograms for training a multi-layer generative model. Qirong Mao et al. [34] introduced feature learning to SER by learning affected-salient features using CNN. They used public emotion speech databases with different languages. With regard to speaker variation, language variation and environmental noise, they achieved high results with learned features compared to other established feature representations. There are many methods to perform emotion recognition using CNNs, however few of them are using spectrograms to recognize emotions from speech which indeed is a new approach in SER. Among the methods using spectrograms for SER, some of them have used an extra classifier at fully connected layer which increases the computational complexity of the overall model. For example, in [34] the effect-salient feature block obtained from the final feature vector is then passed to a SVM classifier [31] to discover the emotion class of the speech utterance. The second reason which makes our work different from the existing works is that we have introduced the use of rectangular kernels which allow extraction of meaningful features from spectrograms. The rectangular kernels and pooling operations are used keeping in view the format of information presented in spectrograms. Lastly, we have used a modified AlexNet architecture which uses a relatively simple layout, compared to modern architectures and is less prone to overfitting with limited training data.

3 Proposed method

The proposed framework utilizes feature learning scheme powered by a discriminative CNN using spectrograms in order to recognize the emotional state of the speaker. The main components of the proposed framework are explained in the subsequent sections.

3.1 Spectrograms extraction from speech

A spectrogram represents the strength or loudness of a signal over time at different frequencies in a particular waveform. With the energy strength at a particular region, we can also see the variation in the energy over time. In general, spectrograms are used to see the frequencies in continuous signals. It is a graph with two geometric dimensions in which time is shown on the horizontal axis, while the vertical axis represents frequency, and the intensity or color of each point in the image corresponds to amplitude of particular frequency at particular time.

Short term Fourier transform (STFT) is usually applied to an electronically recorded sound, to generate spectrograms from the time signal. Using fast Fourier transform (FFT) for generating the spectrogram is a digital process. To discover frequencies at each point in the speech signal, a small sliding window is moved over the signal and FFT is computed for the signal within each window. For a given spectrogram S, the strength of a given frequency component f at a given time t in the speech signal is represented by the darkness or color of the corresponding point $S(t,f)$.We extracted spectrograms as shown in Fig. 1 by using STFT for each audio file in the dataset. Figure 1 contains sample spectrograms for each of the seven emotions in Emo-DB dataset.

In the present work, we extracted spectrograms from each individual file and then we split the spectrogram into multiple smaller spectrograms with an overlap of 50%. This overlap served two purposes. First, it allowed us to simulate a continuity in the processing pipeline. Secondly, it caused an increase in the number of spectrograms which enables us to effectively train or fine-tune a powerful deep CNN. The resulting spectrogram images had dimensions 16 × 256, which were resized to 256 × 256 for input to the CNN.

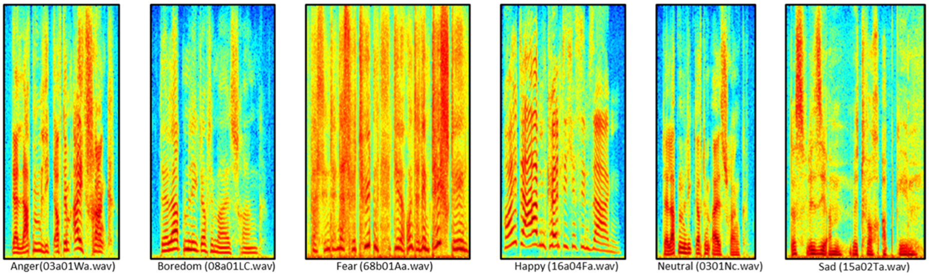


Fig. 1 Spectrograms for various emotions

3.2 Convolutional neural network

Convolutional neural networks are the current state-of-the-art models responsible for the major breakthroughs in image classifications tasks [29]. CNNs use a sequence of filters on the raw pixel data of an image in order to extract and learn high-level features. The model then uses these features to perform classification. CNN architecture consists of three major components; convolutional layers, which apply definite number of convolution filters on the image. Pooling layers; which decrease processing time by reducing dimensionality of the feature maps and fully connected layers; which extract global features from the local feature maps, and performs classification on the extracted features. These layers are usually arranged in the form of a hierarchy where we can use any number of convolutional layers followed by pooling layers and at the end fully-connected layers.

Typically, a CNN is a hierarchical neural network which consists of a stack of convolutional layers followed by pooling layer that performs feature extraction by transforming an image (i.e. spectrogram) to a higher level abstraction in a layer by layer manner. The initial layers consist of simple features like raw image pixels and edges, the higher layers contain local discriminative features, while the last dense (fully connected) layer derives a global representation from the local convolutional features which is then fed to a Softmax classifier to generate probabilities for each class. A convolutional layer applies convolution filters on the small portion of the input image and produces single value in the output feature map by performing dot product and summation operations on these small regions. Each convolutional kernel generates a feature map where the activation values correspond to the presence of particular features. Several feature maps are generated within each convolution layer. Between successive convolutional layers a pooling layer is applied which controls overfitting and reduces computations in the network. The most commonly used pooling algorithm is max pooling, which keeps the maximum value and discards all other values in a local neighborhood. Fully connected layers use more wide filters to process more complex portions in the input layer. Every node in the fully connected layer is connected to every node in preceding layer. The selection of appropriate kernels shapes and sizes, and pooling neighborhoods is key to the success of these models.

3.3 Proposed model architecture

The proposed CNN framework is shown in Fig. 2, which has an input layer, five convolutional layers, three pooling layers, three fully connected layers, and a Softmax layer. The spectrograms generated from emotional speech signals (16 × 256, resized to 256 × 256) are input to the CNN. Convolutional kernels are applied to the input in the initial layers to extract feature maps from these spectrograms. The first convolutional layer C1 has 96 kernels of size 15 × 3 which

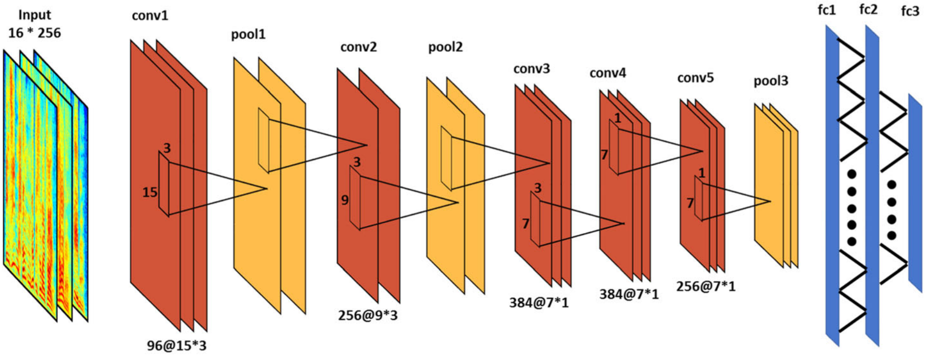


Fig. 2 Proposed CNN architecture with rectangular kernels

are applied with stride setting of (3 × 1) pixels. Its activation functions are rectified linear units (ReLU) which is then followed by max pooling layer of size (3 × 1) having stride 2. Similarly, the second convolutional Layer C2 has 256 kernels of size 9 × 3 with an input of stride 1. Layer C2 is again followed by a max pooling layer of size 3 × 1 and stride 1. In the same way, Layer C3 has 384 (7 × 3) kernels and C4 has 384 (7 × 1) kernels, respectively. The last convolutional layer C5 contains 256 kernels of size 7 × 1 followed by a max pooling layer having size of (3 × 1). Layer C5 is followed by three fully connected (FC) layers having 4096, 4096 and 7 neurons, respectively. The first two fully connected layers are followed by dropout regularization of ratio 75% to avoid overfitting [44]. The output of the last FC layer is fed to the Softmax layer which computes output probabilities for all the seven emotion classes.

This network was designed keeping in view the format of information encoded in the spectrograms. Each input spectrogram corresponds to a shorter sample of the input speech where the frequencies and amplitudes are encoded. In the lower layers, the kernels have greater heights and relatively lower widths so that they can capture local features effectively from the neighborhood. In the subsequent layers, both the height and width are reduced but the shape of the kernels still remain rectangular. It helps to construct effective local receptive fields for spectrograms. The salient features of this architecture are the rectangular shaped kernels, strides, and pooling neighborhoods, which make it possible for the CNN to effectively capture discriminative features from spectrograms.

3.4 Model training

The proposed CNN architecture was implemented in Caffe [24], using NVidia DIGITS 5.0 as frontend [54] for training and validating models. MATLAB was used to generate spectrograms from each emotion in the dataset. The spectrograms of size 16 × 256 were generated with a 50% overlap. Around 1500 spectrograms were generated for each emotion in the dataset. Overall, more than 10,000 spectrograms were generated for all the audio files in the dataset. These spectrograms were divided for training and testing in such a way that 75% of the data was used for training and 25% data was used to validate the performance of the model. A 5-fold cross validation was used for all experiments.

The training process was run for 30 epochs with a batch size of 128. We set the initial learning rate to 0.01 with a decay of 1 after every 10 epochs. A single NVidia GeForce GTX Titan X GPU with 12 GB on-board memory, was used to train and later fine-tune the models. The best accuracy was achieved after 29 epochs. A loss of 0.8066 was obtained on the training set, whereas on the test set a loss of 1.0695 was observed.

A single audio recording usually consists of many spectrograms (as shown in Fig. 3). It is worth mentioning that if more that 30% spectrograms for a single file were correctly classified, then the chance of correctly predicting the entire file will be significantly high. Both the result per image and per file are shown in the result section in detail.

3.5 Emotion prediction using majority voting

Individual spectrograms generated for audio stream are input to the trained CNN for prediction. Such a tiny fragment of speech may not be sufficient to correctly predict any emotions. Hence, prediction results from multiple spectrograms are combined using majority voting scheme for reliable prediction performance. For the speech stream, a large spectrogram is generated which is later segmented into multiple short spectrograms as shown in Fig. 4. Predictions are obtained by using the trained model for each generated spectrogram. Probabilities of seven different emotions are obtained from the Softmax layer of the model for the Emo-DB dataset. Similarly, for the Korean speech dataset, the same architecture was used to predict emotional or normal speech. The overall prediction scores for these emotions are then obtained by using a majority voting scheme where the most frequent label is assigned to the speech stream or a collection of streams if the recording is lengthy and may contain multiple emotions. In the current scenario based on the collected evidence from multiple spectrograms if roughly 25–30% predictions of individual spectrograms for a single audio file are made correctly, then there exists a good chance that the particular emotion will be predicted accurately.

4 Experimental results & analysis

We performed SER on Emo-DB dataset using spectrogram images generated from speech signals. To evaluate performance of the proposed framework, we conducted several experiments and compared the performance of fresh trained CNN with fine-tuned CNN on the extracted spectrograms. The following sections provide details about the experimental setup, experiments conducted, and the analysis of results.

4.1 Datasets

The proposed SER method was trained and evaluated on Berlin Emotional Database (Emo-DB) [9], which contains emotional speech utterances by 10 German actors, and Korean emotional speech dataset obtained from Emergency call centers. Each emotional utterance in Emo-DB is annotated using one of the seven emotions (i.e., anger, boredom, disgust, fear, happy, neutral, and sad). The Korean dataset consists of real calls made by regular people in

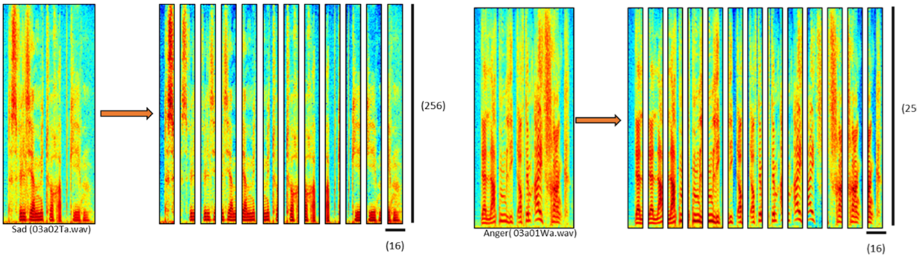
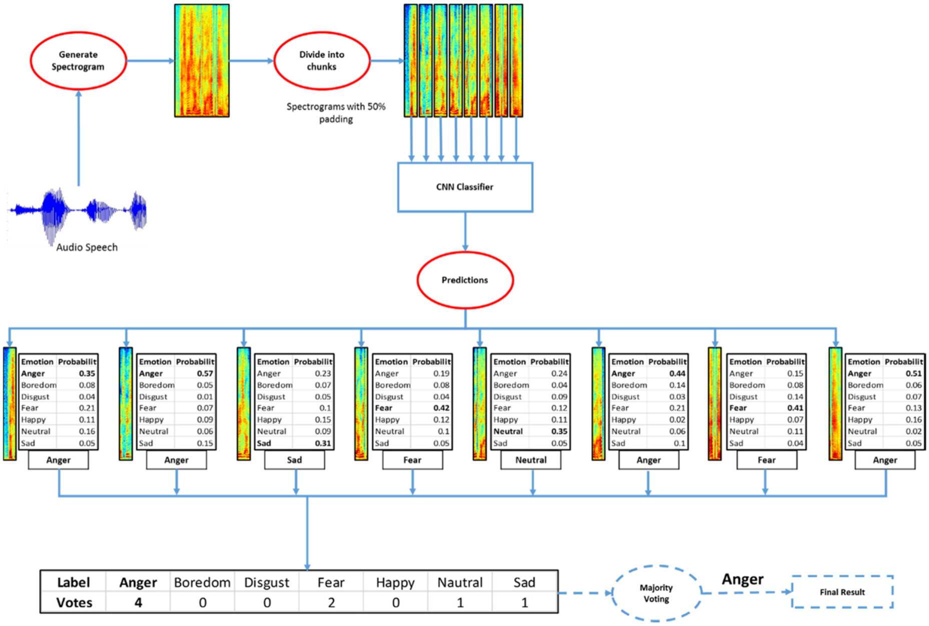


Fig. 3 Spectrogram segmentation with 50% overlap



Emotion	Probabilit
Anger	**0.35**
Boredom	0.08
Disgust	0.04
Fear	0.21
Happy	0.11
Neutral	0.16
Sad	0.05

Anger

Emotion	Probabilit
Anger	**0.57**
Boredom	0.05
Disgust	0.01
Fear	0.07
Happy	0.09
Neutral	0.06
Sad	0.15

Anger

Emotion	Probabilit
Anger	0.23
Boredom	0.07
Disgust	0.05
Fear	0.1
Happy	0.15
Neutral	0.09
Sad	**0.31**

Sad

Emotion	Probabilit
Anger	0.19
Boredom	0.08
Disgust	0.04
Fear	**0.42**
Happy	0.12
Neutral	0.1
Sad	0.05

Fear

Emotion	Probabilit
Anger	0.24
Boredom	0.04
Disgust	0.09
Fear	0.12
Happy	0.11
Neutral	**0.35**
Sad	0.05

Neutral

Emotion	Probabilit
Anger	**0.44**
Boredom	0.14
Disgust	0.03
Fear	0.21
Happy	0.02
Neutral	0.06
Sad	0.1

Anger

Emotion	Probabilit
Anger	0.15
Boredom	0.08
Disgust	0.14
Fear	**0.41**
Happy	0.07
Neutral	0.11
Sad	0.04

Fear

Emotion	Probabilit
Anger	**0.51**
Boredom	0.06
Disgust	0.07
Fear	0.13
Happy	0.16
Neutral	0.02
Sad	0.05

Anger

Label	Anger	Boredom	Disgust	Fear	Happy	Nautral	Sad
Votes	4	0	0	2	0	1	1

Fig. 4 Schematics of the proposed framework

case of emergencies using their phones. The recordings in these datasets contain background noise and the people involved are non-actors which make it more challenging than the Emo-DB dataset. Audio calls in this dataset were labeled as either emotional or normal. To obtain results, five folds cross validation was performed.

4.2 Experiments

In our experiments we first extracted spectrograms from each utterance in the database. Each spectrogram generated from this time-domain signals was then split into multiple smaller spectrograms of size 16 × 256. For input to the CNN, the spectrogram images were resized to 256 × 256. We performed two different sets of experiments. In the first experiment, the prediction performance was assessed by training a fresh CNN model on the training dataset. In the second experiment, we explored transfer learning approach to determine the suitability of using spectrograms for the task of emotions prediction.

4.2.1 CNN performance with square shaped kernels

In this experiment, we used the AlexNet architecture described in [29]. It was trained on the spectrograms generated from Emo-DB dataset using a 75/25 split approach. Tables 1 and 2 summarize performance of the trained model on the test dataset. Table 1 shows the accuracy of the model on individual chunks of generated spectrograms. Table 2 shows the overall prediction accuracy for each emotion. It clearly indicates that for 6 emotions, namely anger, boredom, disgust, fear, neutral and sad, the trained model was able to perform prediction with accuracy above 60%. However, the model was not able to recognize happy emotion with the same accuracy. It was being confused with anger emotion and therefore achieved only

Table 1 Confusion matrix for emotion prediction on individual spectrograms using AlexNet model

Emotion Class	Angry	Boredom	Disgust	Fear	Happy	Neutral	Sad
Angry	**85.09%**	0.45%	2.12%	4.08%	6.48%	1.51%	0.28%
Boredom	4.47%	**43.76%**	4.38%	4.30%	0.50%	26.80%	15.80%
Disgust	9.89%	3.53%	**63.96%**	6.12%	3.42%	5.42%	7.66%
Fear	20.65%	2.71%	11.10%	**47.35%**	8.77%	5.68%	3.74%
Happy	47.63%	2.16%	5.71%	6.14%	**32.76%**	4.20%	1.40%
Neutral	8.41%	11.85%	4.05%	4.96%	1.42%	**62.72%**	6.59%
Sad	0.29%	10.58%	2.50%	1.14%	0.00%	8.22%	**77.27%**

30.88% accuracy. Among the six emotions with accuracy more than 60%, disgust and neutral achieved accuracy above 85%, whereas angry and sad achieved the highest accuracy, which is above 95% as shown in Table 2.

4.2.2 CNN performance with rectangular shaped kernels

Typically, the convolution layers in CNNs consist of many square shaped kernels which are suitable for the type of content, computer vision systems usually deal with. However, in this work, the type of images we are dealing with, are a bit different than the regular images. Information is encoded in a different manner in spectrograms and the previous experiments revealed that the square shaped kernels cannot effectively extract discriminative features from the spectrograms. In order to allow CNN to perform the features extraction effectively, we introduced rectangular kernels, rectangular strides, and max pooling in rectangular regions into the CNN architecture as shown in Fig. 2 which are more suitable for analyzing spectrograms than the square shaped kernels. In this experiment, we trained the CNN model (Fig. 2) using the training dataset and achieved improved classification performance for all emotions. Table 3 shows the confusion matrix obtained from the model for recognizing emotions using individual spectrograms while Table 4 shows the overall prediction result for each emotion. It can be clearly seen that a significant amount of improvement is achieved using the modified CNN model for all emotions. However prediction for emotions like angry, fear and sad got decreased slightly. There still exists some degree of confusion between happy and angry emotions but it does not affect the overall recognition performance, as majority of the spectrograms were classified accurately.

4.2.3 Prediction performance

Figure 5a-g show prediction results for spectrograms generated from individual audio files for each emotion. A single file was picked randomly from each emotion class to see the trend for

Table 2 Confusion matrix for emotion prediction of each file using AlexNet model

Emotion Class	Angry	Boredom	Disgust	Fear	Happy	Neutral	Sad
Angry	**99.8%**	0	0	0.2%	0	0	0
Boredom	1.25%	**68.75%**	0	0	0	21.25%	8.75%
Disgust	6.52%	0	**89.13%**	0	0	2.17%	2.17%
Fear	22.72%	0	3.03%	**66.66%**	4.54%	1.51%	1.51%
Happy	64.70%	0	2.94%	1.47%	**30.88%**	0	0
Neutral	2.56%	3.84%	1.28%	0	0	**89.74%**	2.56%
Sad	0	3.22%	0	0	0	0	**96.77%**

Table 3 Confusion matrix for emotion prediction on individual spectrogram using CNN with rectangular kernels

Emotion Class	Angry	Boredom	Disgust	Fear	Happy	Neutral	Sad
Angry	**83.47%**	0.67%	3.91%	3.57%	6.81%	1.45%	0.11%
Boredom	4.12%	**54.96%**	2.58%	2.74%	0.89%	23.08%	11.62%
Disgust	7.18%	2.59%	**69.61%**	6.12%	2.94%	6.36%	5.18%
Fear	17.42%	4.13%	12.90%	**48.52%**	6.71%	4.90%	5.42%
Happy	37.82%	1.51%	8.62%	7.11%	**40.19%**	3.66%	1.08%
Neutral	6.89%	15.40%	3.85%	2.13%	1.52%	**64.03%**	6.18%
Sad	0.14%	13.72%	2.22%	1.14%	0.07%	7.22%	**75.48%**

each emotion in the file. In each graph the x-axis shows the number of spectrograms for a single file whereas y-axis indicates probability of each of the seven emotions represented in different colors. Figure 5a shows the prediction performance for anger emotion. The selected file from anger emotion test set contains 21 spectrograms in total, out of which 14 were correctly classified as anger, and the remaining 4 were predicted as happy. The mean prediction for anger emotion was 0.47. Predictions for the boredom emotion are shown in Fig. 5b which consist of 26 spectrograms, out of which 16 were predicted correctly. In the remaining spectrograms, 3 spectrograms were classified as fear and 2 as sad. The mean prediction for boredom emotion was 0.43. Figure 5c presents prediction performance for disgust emotion which contains 23 spectrograms. Fourteen individual spectrograms were classified correctly and five spectrograms were misclassified as anger. The overall mean prediction for disgust emotion was 0.42. Twenty one spectrograms were generated from the randomly chosen file for fear emotion as shown in Fig. 5d, out of which, 12 were classified correctly, 5 were misclassified as disgust, 2 as anger, and 1 as happy and neutral. The overall mean prediction for fear emotion was 0.41. Happy emotion file consisted of 23 spectrograms as shown in Fig. 5e, out of which 9 spectrograms were correctly classified with more than 0.50 prediction rate for most of the files. However 6 files were confused with anger emotion. The overall mean prediction for happy was still 0.40 for the selected file. In the same way, 27 spectrograms were generated for a file labelled as neutral emotion. The prediction performance is reported in Fig. 5f, out of total 27 total spectrograms, 17 were predicted correctly, 5 were confused with boredom and two with sad emotion. The overall mean prediction for neutral emotion was 0.42. Figure 5g presents prediction performance for sad emotion. 26 spectrograms were generated from the file, out of which 19 were predicted correctly, 5 were confused with boredom emotion, while one file was predicted as neutral and one as fear emotion. Mean prediction rate for sad emotion was 0.47 which is highest among the files selected for each emotion.

Table 4 Confusion matrix for emotion prediction of each file using CNN with rectangular kernels

Emotion Class	Angry	Boredom	Disgust	Fear	Happy	Neutral	Sad
Angry	**99.32%**	0	0	0	0.78%	0	0
Boredom	0	**76.25%**	0	0	0	21.25%	2.5%
Disgust	2.1%	0	**95.65%**	0	0	2.1%	0
Fear	24.24%	3.0%	7.5%	**62.12%**	0	0	3.0%
Happy	45.94%	0	2.94%	0	**52.45%**	0	0
Neutral	2.56%	11.5%	0	0	0	**84.61%**	1.28%
Sad	0	3.22%	0	0	0	1.61%	**95.16%**

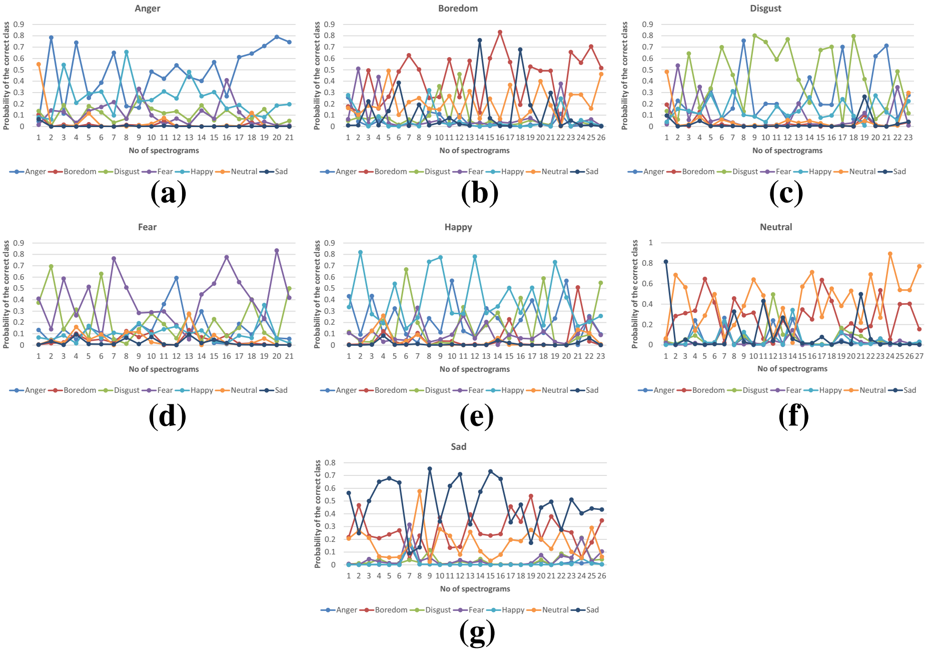


Fig. 5 Prediction performance for various emotions single file containing multiple spectrograms (**a**) Anger, (**b**) Boredom, (**c**) Disgust, (**d**) Fear, (**e**) Happy, (**f**) Neutral, (**g**) Sad

4.3 SER performance comparison with state-of-the-art

In comparison with other state-of-the art approaches, we picked Qirong Mao *et al.* [34], as they have also used spectrogram of the speech signal as the input of CNN. They compared their feature representation technique with other well-established feature representation schemes: spectrogram representation (“RAW” features), TEO [46], acoustic features extracted in [16] (A1), local invariant features (LIF), salient discriminative feature analysis (SDFA) and have shown the average accuracy and standard deviation for these techniques. Moreover, they have used 4 different datasets and we picked the average accuracy for Emo-DB as we have used the same database for our experiments. They have concluded that the learned features (i.e. LIF and SDFA)) outperform the baseline ones (RAW, TEO and A1) on Emo-DB.

Figure 6 compares and shows the average recognition accuracy of five feature extraction methods with the proposed method. The results revealed that they have obtained higher accuracy on SDFA among all the methods under consideration, which is about 72.12%. The proposed model achieved 80.79% average accuracy on the same dataset by using rectangular shaped kernels and a deeper model.

4.4 Affective state analysis in emergency calls

From the previous experiments, it was observed that the proposed framework was able to achieve high recognition rates for most of the emotions using the CNN with rectangular kernels. However, the speech in Emo-DB dataset was uttered by actors with minimal possible noise in the background. To evaluate the effectiveness of our framework, we tried to use it for a more challenging dataset with real life telephone recordings. These audio recordings were

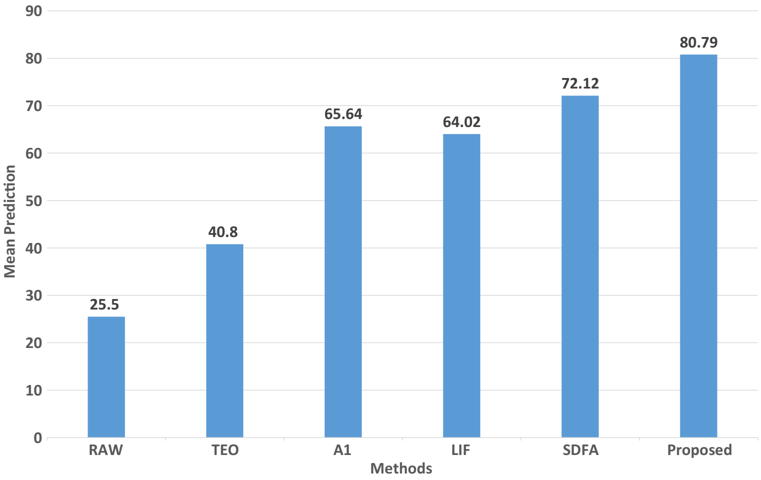


Fig. 6 Comparision of prediction accuracy with other state-of-the-art SER methods

obtained from a Korean emergency call center, and were annotated by experts into two categories i.e. normal and emotional speech. Due to the complex and challenging nature of this data, shallow classification schemes were not able to yield acceptable accuracies. Therefore, we evaluated the performance on emergency calls and conducted several experiments to compare the proposed architecture using rectangular kernels with other state-of-the-art features and classifiers. In these experiments, we used support vector machine (SVM), random forest (RF) and decision tree classifiers to compare the classification accuracy of these classifiers with the proposed CNN framework. Mel Frequency Cepstral Coefficients (MFCCs) features were extracted from the speech signals which were used to train and evaluate these classifiers in separating normal speech from emotional speech [4]. Figure 7 shows the classification accuracy of classifiers using hand-crafted features and the proposed CNN architecture.

Results revealed that our proposed framework was able to achieve highest accuracy among the classifiers used in separating normal speech from emotional speech with 89.46% and 86.54% prediction rate for normal and emotional speech, respectively. SVM predicted with 71.12% recognition rate for normal and 61.48% for emotional speech which was significantly lower

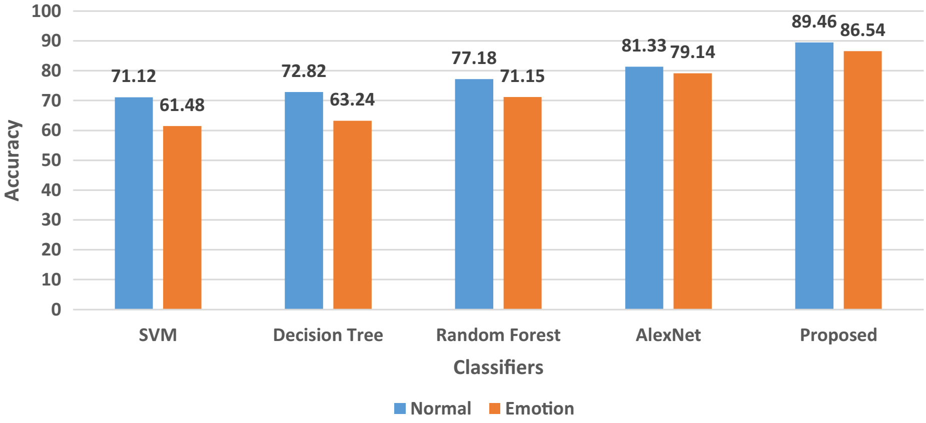


Fig. 7 Prediction accuracy of multiple classifiers

among the classifiers used in the conducted experiments. Classification accuracy for emotional speech using decision tree was 63.24%, while 72.82% accuracy was achieved for normal speech. Random forest achieves better accuracy rates among the classifiers with MFCC features, 77.18% and 71.15% accuracy for normal and emotional speech, respectively. 81.33% and 79.14% accuracy rates were predicted for normal and emotional speech using AlexNet model.

5 Conclusions and future work

In this paper, we presented a method to recognize emotions in speech using convolutional neural network with rectangular kernels. Speech signals are represented as spectrograms which are generated with a 50% overlap. Generated spectrograms were resized to fit the needs of the CNNs during training and evaluation. Two different CNNs were trained on the spectrograms having different kernel sizes and pooling approaches. In the first CNN, the default architecture, similar to AlexNet was used. Whereas, the second CNN was obtained by modifying the kernel sizes and pool neighborhoods from square to rectangular, in order to make it more suitable for spectrograms. Both the CNNs were trained using the same dataset and similar parameters. For each spectrogram, the trained CNNs generated probabilities. For classifying a particular speech segment, majority voting scheme was used.

Experiments revealed that rectangular kernels and max pooling operations in rectangular neighborhoods are more suitable for SER using spectrograms. The format of encoded information in the spectrograms favor rectangular kernels as compared to square shaped kernels. The proposed method can be further enhanced if more labelled data can be collected and a much deeper CNN having rectangular kernels could be effectively trained.

Acknowledgements This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. R0126-15- 1119, Development of a solution for situation-awareness based on the analysis of speech and environmental sounds).

References

1. Abdelgawad H, Shalaby A, Abdulhai B, Gutub AAA (2014) Microscopic modeling of large-scale pedestrian–vehicle conflicts in the city of Madinah, Saudi Arabia. J Adv Transp 48:507–525
2. Ahmad J, Muhammad K, Kwon S-I, Baik SW, Rho S (2016) Dempster-Shafer Fusion Based Gender Recognition for Speech Analysis Applications. In: Platform Technology and Service (PlatCon), 2016 International Conference on, pp 1–4
3. Ahmad J, Sajjad M, Rho S, Kwon S-I, Lee MY, Baik SW (2016) Determining speaker attributes from stress-affected speech in emergency situations with hybrid SVM-DNN architecture. Multimed Tools Appl 1–25. https://doi.org/10.1007/s11042-016-4041-7
4. Ahmad J, Fiaz M, Kwon S-I, Sodanil M, Vo B, Baik SW (2016) Gender Identification using MFCC for Telephone Applications-A Comparative Study. International Journal of Computer Science and Electronics Engineering 3.5 (2015):351–355
5. Aly SA, AlGhamdi TA, Salim M, Amin HH, Gutub AA (2014) Information Gathering Schemes For Collaborative Sensor Devices. Procedia Comput Sci 32:1141–1146
6. Badshah AM, Ahmad J, Rahim N, Baik SW (2017) Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In: Platform Technology and Service (PlatCon), 2017 International Conference on, pp 1–5
7. Banse R, Scherer KR (1996) Acoustic profiles in vocal emotion expression. J Pers Soc Psychol 70:614
8. Bengio Y, Courville A, Vincent P (2013) Representation learning: A review and new perspectives. IEEE Trans Pattern Anal Mach Intell 35:1798–1828

9. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B (2005) A database of German emotional speech. In: Interspeech, pp 1517–1520
10. Curtis S, Zafar B, Gutub A, Manocha D (2013) Right of way. Vis Comput 29:1277–1292
11. Deng L, Seltzer ML, Yu D, Acero A, Mohamed A-R, Hinton GE (2010) Binary coding of speech spectrograms using a deep auto-encoder. In: Interspeech, pp 1692–1695
12. Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on, pp 511–516
13. Dennis J, Tran HD, Li H (2011) Spectrogram image feature for sound event classification in mismatched conditions. IEEE Signal Process Lett 18:130–133
14. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recogn 44:572–587
15. Engberg IS, Hansen AV, Andersen O, Dalsgaard P (1997) Design, recording and verification of a danish emotional speech database. In: Eurospeech
16. Eyben F, Wöllmer M, Schuller B (2009) OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit. In: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pp 1–6
17. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes M (2000) Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 47:829–837
18. Gharavian D, Sheikhan M, Nazerieh A, Garoucy S (2012) Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. Neural Comput & Applic 21:2115–2126
19. Guo Z, Wang ZJ (2013) An unsupervised hierarchical feature learning framework for one-shot image recognition. IEEE Trans Multimedia 15:621–632
20. Gutub A, Alharthi N (2011) Improving Hajj and Umrah Services Utilizing Exploratory Data Visualization Techniques. Inf Vis 10:356–371
21. Guven E, Bock P (2010) Speech emotion recognition using a backward context. In: Applied Imagery Pattern Recognition Workshop (AIPR), 2010 I.E. 39th, pp 1–5
22. Haq S, Jackson PJ, Edge J (2009) Speaker-dependent audio-visual emotion recognition. In: AVSP, pp 53–58
23. Hu H, Xu M-X, Wu W (2007) Fusion of global statistical and segmental spectral features for speech emotion recognition. In: INTERSPEECH, pp 2269–2272
24. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R et al (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, pp 675–678
25. Kaysi I, Sayour M, Alshalalfah B, Gutub A (2012) Rapid transit service in the unique context of Holy Makkah: assessing the first year of operation during the 2010 pilgrimage season. Urban Transp XVIII Urban Transp Environ 21st Century 18:253
26. Kaysi I, Alshalalfah B, Shalaby A, Sayegh A, Sayour M, Gutub A (2013) Users' Evaluation of Rail Systems in Mass Events: Case Study in Mecca, Saudi Arabia. Transp Res Rec J Transp Res Board 2350:111–118
27. Khan MK, Zakariah M, Malik H, Choo K-KR (2017) A novel audio forensic data-set for digital multimedia forensics. Aust J Forensic Sci 1–18. http://doi.org/10.1080/00450618.2017.1296186
28. Kim S, Guy SJ, Hillesland K, Zafar B, Gutub AA-A, Manocha D (2015) Velocity-based modeling of physical interactions in dense crowds. Vis Comput 31:541–555
29. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
30. Krothapalli SR, Koolagudi SG (2013) Emotion recognition using vocal tract information. In: Emotion Recognition Using Speech Features, ed. Springer, pp 67–78
31. Liu P, Choo K-KR, Wang L, Huang F (2016) SVM or deep learning? A comparative study on remote sensing image classification. Soft Comput 1–13. https://doi.org/10.1007/s00500-016-2247-2
32. Lugger M, Janoir M-E, Yang B (2009) Combining classifiers with diverse feature sets for robust speaker independent emotion recognition. In: Signal Processing Conference, 2009 17th European, pp 1225–1229
33. Mao Q, Wang X, Zhan Y (2010) Speech emotion recognition method based on improved decision tree and layered feature selection. Int J Humanoid Rob 7:245–261
34. Mao Q, Dong M, Huang Z, Zhan Y (2014) Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans Multimedia 16:2203–2213
35. Morrison D, Wang R, De Silva LC (2007) Ensemble methods for spoken emotion recognition in call-centres. Speech Comm 49:98–112
36. Nanda A, Sa PK, Choudhury SK, Bakshi S, Majhi B (2017) A Neuromorphic Person Re-Identification Framework for Video Surveillance. IEEE Access 5:6471–6482
37. Pao T-L, Chen Y-T, Yeh J-H, Cheng Y-M, Lin Y-Y (2007) A comparative study of different weighting schemes on KNN-based emotion recognition in Mandarin speech. Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues, pp 997–1005

38. Ramakrishnan S, El Emary IM (2013) Speech emotion recognition approaches in human computer interaction. Telecommun Syst 52(3):1467–1478
39. Raman R, Sa PK, Majhi B, Bakshi S (2016) Direction Estimation for Pedestrian Monitoring System in Smart Cities: An HMM Based Approach. IEEE Access 4:5788–5808
40. Rao KS, Koolagudi SG, Vempada RR (2013) Emotion recognition from speech using global and local prosodic features. Int Journal Speech Technol 16:143–160
41. Rout JK, Choo K-KR, Dash AK, Bakshi S, Jena SK, Williams KL (2017) A model for sentiment and emotion analysis of unstructured social media text. Electron Commer Res 1–19. https://doi.org/10.1007/s10660-017-9257-8
42. Schmidt EM, Kim YE (2011) Learning emotion-based acoustic features with deep belief networks. In: Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 I.E. Workshop on, pp 65–68
43. Schuller B, Rigoll G, Lang M (2004) Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04). IEEE International Conference on, pp I-577
44. Srivastava N, Hinton GE, Krizhevsky A, Sutskever I, Salakhutdinov R (1929-1958) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:2014
45. Stuhlsatz A, Meyer C, Eyben F, Zielke T, Meier G, Schuller B (2011) Deep neural networks for acoustic emotion recognition: raising the benchmarks. In: Acoustics, Speech and Signal Processing (ICASSP), 2011 I.E. International Conference on, pp 5688–5691
46. Sun R, Moore E (2011) Investigating glottal parameters and teager energy operators in emotion recognition. Affective computing and intelligent interaction, pp 425-434
47. Ververidis D, Kotropoulos C (2006) Emotional speech recognition: Resources, features, and methods. Speech Comm 48:1162–1181
48. Wöllmer M, Metallinou A, Katsamanis N, Schuller B, Narayanan S (2012) Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions. In: Acoustics, Speech and Signal Processing (ICASSP), 2012 I.E. International Conference on, pp 4157–4160
49. Xia M, Lijiang C (2010) Speech emotion recognition based on parametric filter and fractal dimension. IEICE Trans Inf Syst 93:2324–2326
50. Xu Z, Luo X, Liu Y, Choo K-KR, Sugumaran V, Yen N et al (2016) From latency, through outbreak, to decline: detecting different states of emergency events using web resources. IEEE Trans Big Data PP:1 https://doi.org/10.1109/TBDATA.2016.2599935
51. Yen N, Zhang H, Wei X, Lu Z, Choo K-KR, Mei L et al (2017) Social Sensors Based Online Attention Computing of Public Safety Events. IEEE Trans Emerg Top Comput 5(3):403–411
52. Yu D, Seltzer ML, Li J, Huang J-T, Seide F (2013) Feature learning in deep neural networks-studies on speech recognition tasks. Published at ICLR 2013. https://sites.google.com/site/representationlearning2013/
53. Yun S, Yoo CD (2012) Loss-scaled large-margin Gaussian mixture models for speech emotion classification. IEEE Trans Audio Speech Lang Process 20:585–598
54. (2017, 4–5-2017). NVIDIA/DIGITS. Available: https://github.com/NVIDIA/DIGITS

Abdul Malik Badshah received his BS degree from Department of Computer Science, Islamia College Peshawar Pakistan. He is currently a Master student at Sejong University, Seoul. His fields of interest include image and signal processing, machine learning and deep learning.

Nasir Rahim received his BS degree in Computer Science from Islamia College, Peshawar, Pakistan in 2014. Currently, he is pursuing Master degree in digital contents from Sejong University, Seoul, South Korea. His research interests include image processing, image classification, and content based image retrieval.

Noor Ullah received his BS degree in Computer Science from Islamia College, Peshawar, Pakistan in 2014. Currently, he is pursuing Master degree in digital contents from Sejong University, Seoul, South Korea. His research interests include image processing, documents classification, and content based image retrieval.

Jamil Ahmad received his BCS degree in Computer Science from the University of Peshawar, Pakistan in 2008 with distinction. He received his Master's degree in 2014 with specialization in Image Processing from Islamia College, Peshawar, Pakistan. He is also a regular faculty member in the Department of Computer Science, Islamia College Peshawar. Currently, he is pursuing PhD degree in Sejong University, Seoul, Korea. His research interests include deep learning, medical image analysis, content-based multimedia retrieval, and computer vision. He has published several journal articles in these areas in reputed journals including Journal of Real-Time Image Processing, Multimedia Tools and Applications, Journal of Visual Communication and Image Representation, PLoS One, Computers and Electrical Engineering, SpringerPlus, Journal of Sensors, and KSII Transactions on Internet and Information Systems. He is also an active reviewer for IET Image Processing, Engineering Applications of Artificial Intelligence, KSII Transactions on Internet and Information Systems, Multimedia Tools and Applications, and IEEE Transactions on Cybernetics.

Khan Muhammad received his BS degree in computer science from Islamia College, Peshawar, Pakistan with research in information security (Image Steganography and Image Encryption). Currently, he is pursuing MS leading to PhD degree in digitals contents from College of Electronics and Information Engineering, Sejong University, Seoul, Republic of Korea. He is working as a researcher at Intelligent Media Laboratory (IM Lab). His research interests include image and video processing, wireless networks, information security, data hiding, image and video steganography, video summarization, diagnostic hysteroscopy, wireless capsule endoscopy, and CCTV video analysis. He is a student member of the IEEE.

Mi Young Lee is a research professor at Sejong University. She received her MS and PhD degrees from Image and Information Engineering, Pusan National University. Her research interests include Interactive Contents, UI, UX and developing Digital Contents.

Soon-il Kwon received his MS and PhD Degrees in Electrical Engineering from University of Southern California, USA, in 2000 and 2005, respectively. His research interests include speech recognition, human computer interaction, affective computing, speech and audio processing.

Sung Wook Baik received the B.S degree in computer science from Seoul National University, Seoul, Korea, in 1987, the M.S. degree in computer science from Northern Illinois University, Dekalb, in 1992, and the Ph.D. degree in information technology engineering from George Mason University, Fairfax, VA, in 1999. He worked at Datamat Systems Research Inc. as a senior scientist of the Intelligent Systems Group from 1997 to 2002. In 2002, he joined the faculty of the College of Electronics and Information Engineering, Sejong University, Seoul, Korea, where he is currently a Full Professor and Dean of Digital Contents. He is also the head of Intelligent Media Laboratory (IM Lab) at Sejong University. He served as professional reviewer for several well-reputed journals such as IEEE Communication Magazine, Senors, Information Fusion, Information Sciences, IEEE TIP, MBEC, MTAP, SIVP and JVCI. His research interests include computer vision, multimedia, pattern recognition, machine learning, data mining, virtual reality, and computer games. He is a member of the IEEE and corresponding author of this paper.