Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network

[1]Abdul Malik Badshah, [1]Jamil Ahmad, [1]Nasir Rahim, [1,*]Sung Wook Baik
[1]*College of Electronics and Information Engineering, Sejong University, Seoul, Republic of Korea*
sbaik@sejong.ac.kr

***Abstract* – This paper presents a method for speech emotion recognition using spectrograms and deep convolutional neural network (CNN). Spectrograms generated from the speech signals are input to the deep CNN. The proposed model consisting of three convolutional layers and three fully connected layers extract discriminative features from spectrogram images and outputs predictions for the seven emotions. In this study, we trained the proposed model on spectrograms obtained from Berlin emotions dataset. Furthermore, we also investigated the effectiveness of transfer learning for emotions recognition using a pre-trained AlexNet model. Preliminary results indicate that the proposed approach based on freshly trained model is better than the fine-tuned model, and is capable of predicting emotions accurately and efficiently.**

Keywords—speech; emotions; convolutional neural network

I. INTRODUCTION

Speech signal is the most natural way to communicate among human beings. Researchers are constantly working to apply this mode in the domain of human-machine interaction. However, it require machines to interpret human spoken phrases intelligently and understand it semantically. Despite the great progress made in speech recognition, this process still requires lot of efforts to it make it *natural interaction* between man and machine. One significant challenge in realizing this goal is the inability of machines to understand the emotional state hidden behind spoken words [1]. In this context, speech emotion recognition (SER) refers to recognition of the emotional state of a speaker by analyzing his/her speech [2]. It is believed that, SER can be used to extract useful semantics from speech, as well as improve the performance of speech recognition system [3].

SER plays an effective role in areas of natural man-machine interactions such as web movies recommendation and computer tutorial applications where the system response depends upon the emotions of the user [1]. It can also be used for in-car board system where information of the mental condition of the driver may be provided to the system to initiate safety procedures, if required [4]. Furthermore, medical use of speech emotion recognition includes diagnostic tool for therapists [5]. In aircraft cockpits, speech recognition system that were trained using stressed speech can exhibit better results compared to those trained by normal speech [6]. Further applications include determination of situational seriousness in emergency call centers based of human emotional analysis from speech data and other mobile communication areas [7]. The main theme of SER systems is to detect particular characteristics of speaker's voice in varying emotional conditions .

Typical SER systems work by extracting features from speech, followed by a classification procedure to predict emotions. There are several challenges being faced by researchers which include selection of appropriate speech features, robustness to tone changes, speaking styles, speaking rates, and the way emotions are expressed in different cultures and environments. One of the major research issues is the extraction of discriminative, robust, and affect-salient features from speech. Features used for SER systems are generally categorized into acoustic, linguistic, context information, and hybrid features combining acoustic and other features [1]. Reliable recognition of emotion heavily depend on the extracted features because they represent the acoustic contents of speech. Besides hand-engineered features including prosodic features (pitch, energy, zero-crossings) [8-10], spectral features (linear predictor coefficients (LPC), linear predictor cepstral coefficients (LPCC), mel-frequency cepstral coefficients (MFCC) [11, 12], and non-linear features like Teager-energy-operator (TEO) [13], automated features extraction methods related to deep learning has also been used. These methods have shown promising results in the fields of speech recognition, emotion recognition, and other speech analysis applications [14-16]. Inspired by the success of these methods, we propose a convolutional neural network (CNN) architecture to extract salient discriminative features from spectrograms for performing SER. Different aspects of the features extraction, content representation, and classification are analyzed and discussed in the context of SER applications.

II. PROPOSED METHODOLOGY

The proposed framework attempts to utilize feature learning schemes for spectrograms generated from speech. This mode of feature learning paradigm bypasses the traditional feature engineering pipeline. Robust and

[*] ***Corresponding Author***

978-1-5090-5140-3/17/$31.00 ©2017 IEEE

discriminative features are learnt from spectrograms automatically forming the basis for SER. The main components of the proposed scheme are illustrated in the subsequent sections.

A. Spectrograms

A spectrogram is the visual representation of a signal strength over time at different frequencies present in certain waveform. It is represented by a two-dimensional graph in which time is shown along the horizontal axis, frequency along the vertical axis, and the amplitude of the frequency components at a particular time is indicated by the intensity or color of that point in the graph. Low amplitudes are represented by dark blue colors and stronger (or louder) amplitudes are represented by brighter colors up through red. It is computed from the speech signal by applying Fast Fourier transform (FFT) to speech signal, which form time-frequency representation. In order to discover the frequencies at a moment in the signal, it is divided into small chunks and FFT is applied to the speech waveform for each chunk. Sample spectrograms are shown in Fig. 1.

Spectrograms have been used in a variety of speech analysis tasks including sound event classification [17], speaker recognition [18], SER [14], and speech recognition [16]. Their suitability for acoustic content representation has been exhibited in these prior works. In this work, the aim is to use spectrograms to represent audio from telephone speech at emergency centers. Phone calls in emergency situations are often made in unconstrained environmental conditions. Consequently, significant background noise, low transmission quality, and Lombard effect, makes SER more challenging. In this preliminary part of the work, we aim to perform SER based on existing emotional speech datasets and later enhance the model to perform SER using telephone speech data.

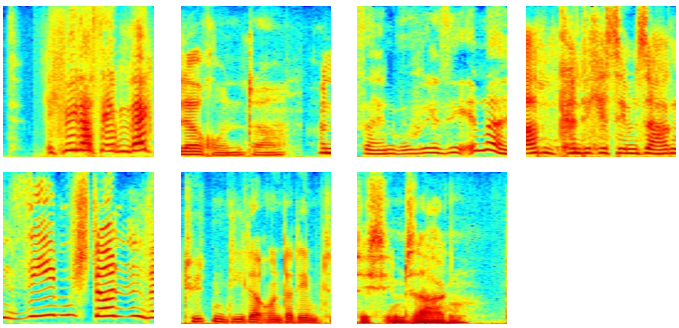

Fig. 1 Sample speech spectrograms

B. Convolutional Neural Networks

Convolutional neural network is a hierarchical neural network which consist of a variety of layers in sequence. A typical model usually consist of several convolutional layers where the visual contents (i.e. spectrograms) are represented as a set of feature maps obtained after convolving the input with a variety of filters which are learned during the training phase. Pooling layers may be introduced after convolutional layers to accumulate maximum activation features from convolutional feature maps. As a result of pooling, spatial resolution of these maps is reduced. Furthermore, CNNs may also contain fully connected (FC) layers where each neuron of the input layer is connected with every neuron in the layer. A sequence of convolutional, pooling, and FC layers form the features extraction pipeline which models the input data in abstract form. Finally, a Softmax layer performs the final classification task based on this representation.

C. Model Architecture

The proposed CNN model, shown in Fig. 2 consist of three convolutional layers, three fully connected layers and a Softmax layer. The input of the network is a 256 x 256 spectrogram generated from emotional speech signals. The initial convolutional layers extract features from these spectrograms using convolution operations. Layer C1 has 120 (11 x 11) kernels which are applied at a stride setting of 4 pixels. It is followed by rectified linear units (ReLU) and a max pooling layer of size 3 x 3 with stride 2. ReLU act as activation functions instead of the typical sigmoid functions which improves efficiency of the training process. Layer C2 has 256 kernels of size 5 x 5 and they are applied on the input with a stride 1. Similarly, C3 has 384 kernels of size 3 x 3. Each of these conv. layers are followed by ReLU units. Layer C3 is followed by three FC layers having 2048, 2048, and 7 neurons, respectively. In order to avoid overfitting, the first two FC layers are followed by dropout layers having a dropout ratio of 50%. Compared to the natural images, it is relatively difficult to extract discriminative features from spectrograms for robust emotions recognition.

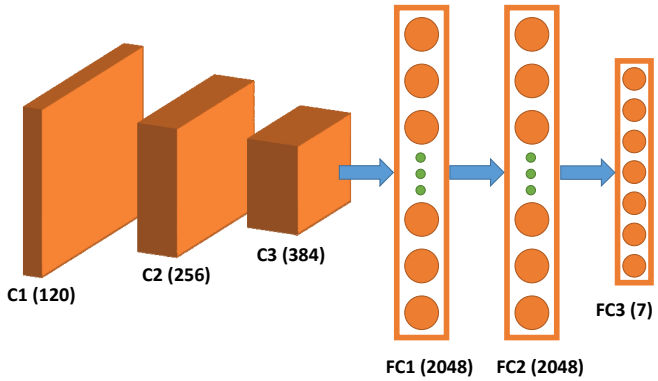

Fig.2 Proposed CNN architecture for SER using spectrograms

D. Model Training

The proposed CNN model was specified in Caffe [19], using NVidia DIGITS 3.0 as a frontend [20]. The spectrogram images were generated in MATLAB from the Berlin Speech Emotion dataset and resized to 256 x 256. More than 3000 spectrograms were generated from all the audio files in the dataset. Seventy five percent of this data was used for training and the rest was used in the testing phase. For each emotion, we had collected about 500 images in the dataset.

The training process was run for 30 epochs with a batch size set to 100. Initial learning rate was set to 0.01 with a decay of 0.1 after each 10 epochs. The training was performed on a single NVidia GeForce GTX Titan X GPU with 12 GB onboard memory. The training took around 40 minutes and the best accuracy was achieved after 28 epochs. On the training set, a loss of 0.71 was achieved, whereas 0.95 loss was recorded on the test set. An accuracy of 61.75 % was achieved per spectrogram. It is important to notice here that the overall accuracy is very low. However, each audio file consist of multiple spectrograms, and if our method can classify more than 30% of the spectrograms correctly, then the chances of predicting emotions in the entire file are sufficiently high. An aggregation mechanism governed is employed to combine the individual predictions into an overall prediction result for the entire audio call.

E. Emotion Prediction using CNN

The trained model is used to obtain predictions for each spectrogram generated for the incoming speech signals. The Softmax layer of the model outputs predictions for the seven different emotions in the form of probabilities which are used in the mean prediction based reasoning process to obtain overall prediction scores for these emotions. Prediction reported by the model for each individual spectrogram act as evidence to update the belief values for all emotions. In the current scenario with seven classes, if roughly 25-30% predictions for the audio files are made correctly, then there is a good chance that the particular emotion may be predicted accurately based on the collected evidence from multiple spectrograms. The predictions from CNN model are accumulated to determine the probabilities for individual emotions by computing mean predictions from the gathered evidence.

III. Experimental Results & Analysis

This section provides details about the experimental setup, conducted experiments, and analysis of the results.

A. Dataset

Berlin dataset [21] was used to assess performance of the proposed SER system. The dataset consist of expert annotated speech data from four different users. Every audio file is annotated using one of the seven different emotions including *neutral*, *fear*, *anger*, *happy*, *sadness*, *disgust*, and *boredom*. Seventy five percent of this data was used for training and the remaining was used for testing. Five folds cross validation was performed to obtain the results.

B. Experiments

Several experiments were carried out to assess the suitability of proposed method for emotion recognition. Two different set of experiments were performed. In the first experiment, the training dataset was used to train a fresh CNN model and its prediction performance is assessed. In the second experiment, transfer learning approach is explored to determine its suitability for the task of emotions prediction using spectrograms.

1. Fresh Trained CNN based SER

The model presented in Fig. 2 was trained on the spectrograms generated from the Berlin emotions dataset. Results of the trained model on the test set are shown in Table 1. It can be seen that the model was able to perform predictions with accuracy above 50% for emotions like anger, boredom, disgust, and sad. However, prediction performance for fear, happy, and neutral emotions were lower than 50%. Fear emotions are often confused with anger, disgust, and happy. Though the confusion rate (around 19% in each case) is lower than the correct predictions (25.33%). This makes it more probable that the fear emotion may be correctly predicted if sufficient evidence can be gathered from the speech stream. On the other hand neutral emotions are heavily confused with boredom. Similarly, happy emotion is confused with anger which adversely affect it prediction performance. In case of these three emotions, further work needs to be done to decrease their confusion with the other emotions.

Table 1. Confusion matrix for emotion prediction using fresh trained CNN

	Anger	**Boredom**	**Disgust**	**Fear**	**Happy**	**Neutral**	**Sad**
Anger	84.21	0.00	1.75	3.51	10.53	0.00	0.00
Boredom	0.00	53.91	7.83	4.35	2.61	14.78	16.52
Disgust	3.80	7.59	68.35	7.59	7.59	3.80	1.27
Fear	18.67	14.67	16.00	25.33	18.67	2.67	4.00
Happy	46.59	0.00	4.55	10.23	36.36	1.14	1.14
Neutral	0.00	41.05	1.05	3.16	2.11	42.11	10.53
Sad	0.00	13.85	0.00	0.00	0.00	3.08	83.08

2. Fine-tuned CNN based SER

In the recent years, transfer learning has been used to utilize the learning from a relatively different domain and apply it to solve a particular problem more effectively. In such a case, the learned weights from the pre-trained model are used to initialize the model before tuning the parameters according to the new dataset. Learning rate is usually kept very small (one tenth of the normal learning rate) so that the weights are adjusted slightly. Though in several cases, transfer learning may help achieve better performance than the fresh trained model, in our case, it wasn't very helpful because of the huge difference in the types of datasets used to train or fine-tune the model. The confusion matrix obtained by fine-tuned AlexNet model [22] (trained on ImageNet data [23]) for emotions recognition is provided in Table 2. This fine-tuned model improves the prediction performance in case of anger, neutral, and sad emotions. However, prediction performance in case of the other four emotions got decreased. In the current scenario, we opted to use the freshly trained CNN due to lesser complexity and better performance.

Table 2. Confusion matrix for emotion prediction using fine-tuned CNN

	Anger	Boredom	Disgust	Fear	Happy	Neutral	Sad
Anger	92.98	0.00	0.00	3.51	1.75	1.75	0.00
Boredom	3.48	37.39	6.96	0.00	0.00	41.74	10.43
Disgust	22.78	2.53	49.37	3.80	3.80	11.39	6.33
Fear	34.67	2.67	0.00	46.67	1.33	13.33	1.33
Happy	73.86	1.14	3.41	3.41	17.05	1.14	0.00
Neutral	5.26	8.42	1.05	2.11	3.16	75.79	4.21
Sad	0.00	10.00	1.54	2.31	0.00	10.77	75.38

C. Prediction Performance

Prediction results (i.e. probabilities of true class) for spectrograms generated for various emotional audio files from the test set are provided in Fig. 3. Results reveal that the proposed framework is capable of correctly predicting most of the emotions by generating high confidence outputs more than 50% of the time. In case of fear and happy emotions, the performance is acceptable but not very good because some of the spectrograms are confused other emotions.

The mean predictions for anger, disgust, and sad emotions are above 0.68, whereas for the boredom, fear, happy, and neutral are 0.48, 0.33, 0.35, and 0.44 respectively. In all the cases mean prediction value for all the emotions was greater than the other emotions. Overall, the proposed method helped us achieve an accuracy of 84.3% for all the speakers on the test set.

IV. Conclusions

In this paper, we attempt to solve the problem of SER using feature learning scheme based on deep convolutional neural networks. Speech signal is represented as spectrograms which act as the input to deep CNNs. The CNN model consisting of three convolutional and three fully connected layers extract features from these spectrograms and output predictions for the seven emotion classes. In this regard, two different set of experiments were performed. In the first experiment, we trained a fresh CNN model based on the spectrograms generated for the Berlin emotions dataset. Satisfactory results were achieved using this model for most of the emotions except fear. In the second experiment, we fine-tuned a pre-trained AlexNet model to determine the suitability of transfer learning in solving the problem of SER. However, the results were not satisfactory. Further work is needed to improve the proposed framework so that all emotions are recognized effectively in a robust manner. We plan to use more data with relatively complex models to improve the SER performance even further.

Acknowledgement

This work was supported by the ICT R&D program of MSIP/IITP. (No. R0126-15-1119, Development of a solution for situation-awareness based on the analysis of speech and environmental sounds).

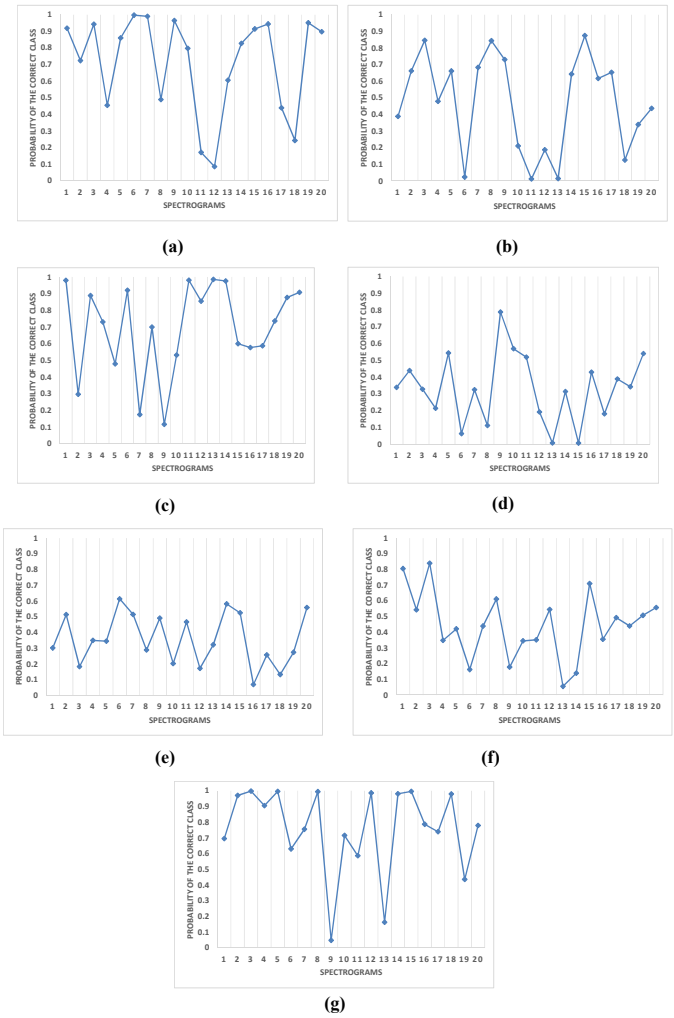

Fig. 3. Prediction performance on test audio files for (a) anger, (b) boredom, (c) disgust, (d) fear, (e) happy, (f) neutral, and (g) sad.

References

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition,* vol. 44, pp. 572-587, 2011.

[2] Z. Zhao and X. Ma, "Active learning for speech emotion recognition using conditional random fields," in *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2013 14th ACIS International Conference on*, 2013, pp. 127-131.

[3] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies,* vol. 59, pp. 157-183, 2003.

[4] B. Schuller, G. Rigoll, and M. Lang, "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*, 2004, pp. I-577-80 vol. 1.

[5] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE transactions on Biomedical Engineering,* vol. 47, pp. 829-837, 2000.

[6] J. H. Hansen and D. A. Cairns, "Icarus: Source generator based real-time recognition of speech in noisy stressful and lombard

effect environments☆," *Speech Communication,* vol. 16, pp. 391-422, 1995.

[7] J. Ahmad, K. Muhammad, S. i. Kwon, S. W. Baik, and S. Rho, "Dempster-Shafer Fusion Based Gender Recognition for Speech Analysis Applications," in *2016 International Conference on Platform Technology and Service (PlatCon)*, 2016, pp. 1-4.

[8] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language,* vol. 27, pp. 151-167, 2013.

[9] H. Meinedo and I. Trancoso, "Age and gender classification using fusion of acoustic and prosodic features," in *INTERSPEECH*, 2010, pp. 2818-2821.

[10] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication,* vol. 46, pp. 455-472, 2005.

[11] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, pp. 1142-1158, 2009.

[12] J. Ahmad;, M. Fiaz;, S.-i. Kwon;, M. Sodanil;, B. Vo;, and S. W. Baik, "Gender Identification using MFCC for Telephone Applications - A Comparative Study," *International Journal of Computer Science and Electronics Engineering,* vol. 3, pp. 351-355, 2015.

[13] S.-H. Chen and J.-F. Wang, "Speech enhancement using perceptual wavelet packet decomposition and teager energy operator," in *Real World Speech Processing*, ed: Springer, 2004, pp. 51-65.

[14] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia,* vol. 16, pp. 2203-2213, 2014.

[15] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8614-8618.

[16] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605,* 2013.

[17] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE signal processing letters,* vol. 18, pp. 130-133, 2011.

[18] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, 2009, pp. 1096-1104.

[19] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick*, et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 675-678.

[20] (2016). *NVidia DIGITS*. Available: https://github.com/NVIDIA/DIGITS

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, 2005, pp. 1517-1520.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097-1105.

[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma*, et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision,* vol. 115, pp. 211-252, 2015.