Retraction

Retraction: Feature Specific Hybrid Framework on composition of Deep learning architecture for speech emotion recognition (*J. Phys. Conf. Ser.* 1916 012094)

Published 23 February 2022

This article (and all articles in the proceedings volume relating to the same conference) has been retracted by IOP Publishing following an extensive investigation in line with the COPE guidelines. This investigation has uncovered evidence of systematic manipulation of the publication process and considerable citation manipulation.

IOP Publishing respectfully requests that readers consider all work within this volume potentially unreliable, as the volume has not been through a credible peer review process.

IOP Publishing regrets that our usual quality checks did not identify these issues before publication, and have since put additional measures in place to try to prevent these issues from reoccurring. IOP Publishing wishes to credit anonymous whistleblowers and the Problematic Paper Screener [1] for bringing some of the above issues to our attention, prompting us to investigate further.

[1] Cabanac G, Labbé C and Magazinov A 2021 arXiv:2107.06751v1

Retraction published: 23 February 2022
Feature Specific Hybrid Framework on composition of Deep learning architecture for speech emotion recognition

Mansoor Hussain\textsuperscript{1}, Abishek S\textsuperscript{1}, Ashwanth K P\textsuperscript{1}, Bharanidharan C\textsuperscript{1}, Girish S\textsuperscript{1}

\textsuperscript{1}Computer Science and Engineering, Sri Krishna College of Engineering and Technology
ashwanthpalanisamy@gmail.com

Abstract. Speech cues may be used to identify human emotions using deep learning model of speech emotion recognition using supervised learning or unsupervised learning as machine learning concepts, and then it build the speech emotion databases for test data prediction. Despite of many advantageous, still it suffers from accuracy and other aspects. In order to mitigate those issues, we propose a new feature specific hybrid framework on composition of deep learning architecture such as recurrent neural network and convolution neural network for speech emotion recognition. It analyses different characteristics to make a better description of speech emotion. Initially it uses feature extraction technique using bag-of-Audio-word model to Mel-frequency cepstral factor characteristics and a pack of acoustic words composed of emotion features to feed the hybrid deep learning architecture to result in high classification and prediction accuracy. In addition, the proposed hybrid networks' output is concatenated and loaded into this layer of softmax, which produces a for speech recognition, a categorical classification statistic is used. The proposed model is based on the Ryerson Audio-Visual Database of Emotional Speech and Song audio (RAVDESS) dataset, which comprises eight emotional groups. Experimental results on dataset prove that proposed framework performs better in terms of 89.5% recognition rate and 98% accuracy against state of art approaches.

Keywords: Speech Emotion Recognition, Convolution Neural Network, Recurrent Neural Network, Mel frequency cepstral coefficients

1. Introduction

Speech emotional recognition is still a complicated subject. Acknowledging a entity or emotion as a result of its properties is a particularly challenging job in pattern recognition and artificial intelligence. Deep learning (DL) has shown great promise in a variety of fields, including social network analysis and forensics. Furthermore, DL has aided major developments in recognition science, such as speech and object recognition \[1\]. Because of their ability to learn strong structures from actual data, deep learning models have also been used as feature vectors and effectively identify details \[2\]. Furthermore, the emotional quality of a patient's voice is often used in medical diagnostics for a variety of disorders.

However, there are some drawbacks to using a machine learning model to recognise speech emotion, including the following. First, there's function analysis received far less attention in emotion
recognition than in speech recognition, resulting in a lack of consensus among researchers on the features are best for feature extraction. Furthermore, due to a lack of cooperation among researchers and there aren't any benchmarking databases that researchers can share, the identical errors were made replicated in recording for various emotional speech databases.

Consequently, current paper focuses on emotional speech recognition using feature specific hybrid deep learning architecture such as CNN [3] and RNN [4] on extracted features. Initially Features are extracted using Bag of Acoustics Words. Extracted feature taken by CNN and RNN for prediction and classification of speech emotions. The remainder of this document is laid out as follows. The Section II explains previous work on speech emotion identification. Section III goes through the specifics of the proposed systems as well as the context information, including a review of the emotion recognition methodology. The implementation and performance assessment of the systems are covered in Section IV. Finally, conclusions are discussed in section V.

2. Related works

On various aspects of prediction characteristics, the most up-to-date for Emotional recognition of voice using machine learning and deep learning techniques is listed in this section. On various aspects of prediction characteristics, the most up-to-date in emotion recognition using machine learning algorithms techniques is listed in this section.

2.1. Bidirectional Long Short-Term Memory Network with Adulation Convolution Skip for Speech Emotion Recognition

Attention-Based Convolution is used in this form. To recognise speech emotion, a SCBAMM stands for skip convolution dependent on attention long short-term memory that can be used in both directions, a deep-learning acoustic model, is used. A Bi-LSTM layer, a convolutional layer, a skip layer, a mask layer, and a convolutional layer, and an attention layer are among the thick layers, and a pooling layer are among the eight hidden layers. SCBAMM allows greater use of spatiotemporal data and more efficiently captures features associated with emotion. Furthermore, it partially addresses the problems of gradient bursting and gradient disappearing in deep learning [5].

2.2. Sparse Kernel Reduced-Rank Regression for Speech Emotion Recognition

In this method, Sparse Kernel Reduced-Rank Regression for Speech Emotion Recognition has been employed as machine learning model. In this analysis, SMILE The scale symmetric and the function extractor feature descriptor feature transformation has been used as feature extraction technique for speech emotion on various modalities [6]. The extracted emotions have been classified and predicted using Sparse Kernel Reduced-Rank Regression to find the best solution for the matrices of coefficients with good recognition rate. For instruction it produces the 5 types based on two separate datasets of the classifier.

3. Proposed model

In this section, a novel feature specific hybrid deep learning architecture has been proposed for speech recognition has been designed with following functionalities

3.1. Signal Pre-processing

Segmentation based pre-processing has been applied to the speech signal according to a silenced and un-silenced part of the speech signal in order to reduce the amount of data supplied to the input of the neural network, which considerably improves its input data sensitivity. The methods suggested are based on the use of the three edit law [7].

3.2. Feature Extraction

BoAW is the famous commonly utilised exemplification techniques in order to recognise emotional expression since there is no standard pattern for acoustic input. It's the only pejoratives of acoustic low-level that can only be obtained from audio signals. BoAW was then accustomed to portray a sound’s two-dimensional simple power range. A mel sizing $f_{\text{mel}}$ synthesises the frequency sensed by the human auditory system and is used to apply the frequency scale in physical terms $f$ (Hz) human ears have heard to all this. The role of the mel scale was expressed as

$$F_{\text{mel}} = 2595 \log_{10}(1 + (f/700))$$

The number of extracted MFCC vectors in the input audio signals varies depending on the duration of the audio signal. As a result, BoAW is used to solve this problem by using a clustering algorithm to construct a vector with a fixed duration from a variable length audio signal. The frequency of each disk's existence were then computed as characteristics used when developing a graph of data of acoustic words [8].

3.3. Speech Emotion classification using CNN

The Extracted Feature Vector are constructed as Matrix. For the Convolution layers, a layer for feedback is combined with N grading moderators, every one with a scale $S \times S$. The model's weights are learned during the learning process. AlexNet's model, which has layers 64 filters of size 7 X 7 which is followed by 128 5 X 5 size filters, and ended with 192 3 X 3 size filters, has achieved extremely high image recognition tasks with amazing precision. In general, progressively the quantity was raised to let us collect more data and a visual image of the data in a larger space.

Max pooling layer

A template discretionary mechanism is used in the max pooling layer that is usually placed between successive Convolution layers [9]. Its aim is to measure the highest, or each function map's greatest value on every group while keeping the scale of the depth unchanged. Which is achieved by applying a max buffer to most of the main input's sub-regions that do not overlap. The main objective of the max pooling layer is to resolve the problem of computational complexity.

![Figure 1. Speech Emotion Recognition using CNN](image)

Figure 1 represents the construction of CNN for emotional recognition of voice. The most popular method of max pooling is pooling layers with a filter size of 2 X 2. It's normally, for each spatial position in the feedback, a two-down-sample leash was added, twice both in terms of width and height.

Batch Normalization layer
The batch normalisation rule states that you can solve this problem by ensuring that the input given to the more advanced layers is only about 0 and 1. The mean and variance of the batch can be used to normalise it.

**Drop out layer**

The model is prohibited from deciphering or removing the over-fitting of the training data n percent randomly from the weights between layers. The random dropout of n % of weights in - batch helps to avoid excessive conformity and accelerates the method of education. Both of the weights are used in the calibration process with no dropouts.

### 3.4. Speech Emotion Classification using RNN

Each neuron in the RNN structure was linked to the next layer’s neuron. RNN networks require a fixed-size input and do not operate with variable-length sequential data. A RNN network is made up of three layers, each of which computes a vector. The count of nodes in the feedback layer is corresponding to the dimensionality of BoAW vectors. In softmax layer for classification and estimation, there were eight nodes in the output layer [10]. Nonlinear functions are used to describe hidden layers, which are made up of one or more layers Figure 2.

![Figure 2. Architecture of RNN for Speech Emotion Recognition](image)

The weights are multiplied by the input vectors of hidden layers, and the output to the next layer is generated using a nonlinear function. During training, linear and nonlinear activation functions are used to propagate the input BoAW vectors bi from the feedback layer to the final layer. The feedback layer uses a simple activation function, although the output and secret layers hire a Softmax nonlinear activation function and rectified linear unit (ReLU), respectively.

### 4. Experimental Results

This segment presents the emotional expression class extraction simulation and synthesis performance using BoAW, RNN, and CNN. The confusion precision, accuracy, and a matrix for disparate factions for proposed and state of art approaches are determined and demonstrated on public dataset RAVDESS using python. The Performance evaluation of the model is represented in the figure 3.
Our studies are based on speech recordings that include eight distinct classes (calm, neutral, angry, happy, sad, fearful, disgusted, and surprised). To get more accurate result and reduce regularisation, the database for training program is categorized up into different divides in dinky tests on concept of multi, which uses (n 1) folds for preparation and one fold for the training case recursively. Table 1 provides the performance analysis of the speech recognition system.

Table 1. Performance Comparison on Speech Emotion Recognition System

| Technique                                      | Recognition Rate | Accuracy |
|-----------------------------------------------|------------------|----------|
| Speech Emotion Recognition using RNN-Existing | 78.6%            | 96.23    |
| Speech Emotion Recognition using Hybrid Deep Learning Architecture- Proposed | 89.5%            | 98.23    |

Furthermore, based on the data in Table 1, proposed hybrid Deep Learning architecture produces highest overall accuracy equal to 98.23 and recognition rate equal to 89.5% facilitating the model's acquisition of data and prediction of emotional groups.

5. Conclusion

Automatic speech emotion detection has recently become a hot topic in academia. Since there are unknown what are the characteristics that successful for the expression of emotion in speech recognition, it is a difficult issue. Proposed architecture named hybrid deep learning architecture a method for obtaining emotional information group as a result of acoustic signals is presented by The platform is built on BoAW, CNN and RNN to categorise the various emotional states. The RAVDESS audio dataset was used to test the proposed architectures for classifying eight emotions in voice. In contrast to cutting-edge techniques, the suggested models outperformed them significantly.

References

[1] X. Jiang, A. Hadid, Y. Pang, E. Granger, and X. Feng, Deep Learning in Object Detection and Recognition. Singapore: Springer, Jan. 2019, doi: 10.1007/978-981-10-5152-4

[2] M. El Ayadi, M. S. Kamel, and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognit., vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020

[3] M. A. Jalal, E. Loweimi, R. K. Moore, and T. Hain, Learning temporal clusters using capsule routing for speech emotion recognition, in Proc. Interspeech, Graz, Austria, Sep. 2019, pp. 1701–1705, doi: 10.21437/Interspeech.2019-3068.
[4] 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 Trans. Biomed. Eng., vol. 47, no. 7, pp. 829–837, Jul. 2000, doi: 10.1109/10.846676.

[5] G. E. Hinton, S. Osindero, and Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput., vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.

[6] T. Danisman and A. Alpkocak, Emotion classification of audio signals using ensemble of support vector machines, in Proc. Int. Tutorial Res. Workshop Perception Interact. Technol. Speech-Based Syst. Berlin, Germany: Springer, 2008, pp. 205–216, doi: 10.1007/978-3-540-69369-7_23.

[7] PrakashDuraisamy, XiaohuiYuan, ElSaba,A. and Sumithra Palanisamy, Contrast enhancement and assessment of OCT images, Proceedings of International Conference on Informatics, Electronics & Vision (ICIEV), 2012 Date: 18–19 May 2012 pp.91-95(Location :Dhaka, Print ISBN: 978-1-4673-1153-3,INSPEC Accession Number: 13058449,Digital Object Identifier :10.1109/ICIEV.2012.6317381).

[8] Sumithra M. G., Thanushkodi, K. and Helan Jenifer Archana ,A. A New Speaker Recognition System with Combined Feature Extraction Techniques, Journal of Computer Science, Vol. 7, Issue 4, pp.459- 465, 2011. (With impact factor SNIP of 0.162 and SJR of0.034).

[9] Balasaraswathi, M., Srinivasan, K., Udayakumar, L., Sivasakthiselvan, S. and Sumithra, M.G., 2020. Big data analytic of contexts and cascading tourism for smart city. Materials Today: Proceedings.

[10] Sivakumar, P., Boopathi, C.S., Sumithra, M.G., Singh, M., Malhotra, J. and Grover, A., 2020. Ultra-high capacity long-haul PDM-16-QAM-based WDM-FSO transmission system using coherent detection and digital signal processing. Optical and Quantum Electronics, 52(11), pp.1-18.