Deep Learning as Feature Encoding for Emotion Recognition

Bhalaji Nagarajan, V. Ramana Murthy Oruganti
Dept. of Electrical and Electronics Engineering
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India

Abstract—Deep learning is popular as an end-to-end framework extracting the prominent features and performing the classification also. In this paper, we intensively investigate deep networks as an alternate to feature encoding technique of low-level descriptors for emotion recognition on the benchmark EmoDB dataset. Fusion performance with such obtained encoded features with other available features is also investigated. Highest performance to date in the literature is observed.

Index Terms—Emotion recognition, feature encoding, OpenSMILE, LSTM, BILSTM

I. INTRODUCTION

Speech emotion recognition has been a growing area of research as more and more intelligent voice assistants and spoken dialogue systems are made available to the common man. It is necessary for these intelligent systems to understand the human emotions as it will enable them to understand the human behavior and provide better interactions with the users. Human emotions can be classified into six basic emotions anger, surprise, disgust, fear, happiness and sadness. Each of these emotions can be differentiated using the tone of the speech. The tone and energy of the speech is modified by the speaker to produce different emotions [1]. These changes are easily detected by humans even if emotions are expressed at a subtle level. But it is challenging for a computer system to recognize these emotions automatic [2].

Currently, deep net features are observed to outperform handcrafted features in audio and vision community, including emotion recognition. These deep nets are trained directly over the source audio or video files. However, in several situations, the source files cannot be made public due to confidentiality agreements with the participants. Handcrafted features can be made public for furthering research in that field. In that perspective, this paper investigates extensively the following two performances

1) Can deep nets be used as an alternate to traditional feature encoding of local descriptors?
2) Can such deep net encoded features complement other available features for enhancing overall performance?

The rest of the paper is organized as follows. Latest works in 2018 are reviewed briefly in Section 2. The outline of the proposed technique is shown in Section 3. The experimental setup, results and the inferences are presented in Section 4 followed by the Conclusion in the last section.

II. RELATED WORKS

This sub-section reviews handcrafted features for emotion recognition. In [4], spectral features – linear prediction coefficients and Mel frequency cepstral coefficients were extracted from the speech input. Fuzzy C-means (FCM) was used for identifying Cluster centers. Success rate of 92.86% using the SVM classifier was reported on the EmoDB dataset. [9] investigated the significance of the vocal tract for emotion recognition from the speech input by using multi-resolution analysis (MRA). [10] proposed a novel prominence related feature that complements with traditional acoustic features for emotion classification. [11] proposed a variational mode decomposition (VMD) based approach for the estimation of epochs from the speech input. Epochs are defined as the locations of significant excitation in the vocal tract during the production of voice signal. The estimation of epoch locations enhances emotion analysis.

This sub-section reviews deep learning based features for emotion recognition. Deep features from different descriptors extracted from audio samples and spectrograms are learnt separately and then fused to form a single emotion recognition net [3]. [5] hypothesized that Mel-spectrogram with deltas and delta-deltas preserves the effective emotional information and reduces the effect of emotionally irrelevant frames. Hence they proposed an attentional model – a 3D attention-based convolutional recurrent neural networks (RNN) to learn discriminative features for speech emotion using the deltas and delta-deltas. [8] proposed a novel method of learning the hidden representations between multimodal – speech and text data using convolutional attention networks. [7] investigated pre-trained Imagenet models for speech feature extraction. Further training on the target emotional speech database such as EmoDB substantially enhanced recognition performance.

Experiments in this paper are conducted using early version of deep learning network. A new novel feature extractor is needed to enhance recognition performance. In such situation, it is ideal to use a pre-trained deep learning network for the target emotion database. It has resulted in enhanced performance of such pre-trained deep learning network. It is expected that the recognition performance of such pre-trained deep learning network will be further enhanced.

The rest of the paper is organized as follows. Latest works in 2018 are reviewed briefly in Section 2. The outline of the proposed technique is shown in Section 3. The experimental setup, results and the inferences are presented in Section 4 followed by the Conclusion in the last section.
proposed semi-supervised auto-encoders to improve speech emotion recognition to get the benefit from the the labelled and unlabelled data.

III. PROPOSED WORK

The performance of classifiers depends on the discriminative power of the available features. In this work, we explore extensively three types of features – handcrafted, pre-trained net and the deep net features. An overview of the proposed framework is presented in Figure 1. Information from different layers of pre-trained deep net models are used to yield pre-trained net features. CNN+LSTM network is trained using spectrograms extracted from the EmoDB dataset. This network yields deep net features. Both the pre-trained net and deep net feature vectors are concatenated [14] for final emotion recognition task. Details of each module are described in the following sections

A. Handcrafted Features

The input audio signals are converted into N segments of fixed length. The INTERSPEECH 2013 ComParE Vocalization Challenge feature set [20] (IS13 ComParE) configuration is used in OpenSMILE toolkit to extract Low-level Descriptors (LLDs) of 141D using 250ms frame size and 10ms step size. Further, the segments are converted to spectrograms using a Hamming window of 25ms width and 10ms step. These spectrograms are used as the input to the pre-trained net framework described in next section.

B. Pre-trained Net Features

Often, very deep network trained on large scale datasets are made available to research community. Although these networks are trained for objectives such as speaker identification, sound classification, the information of different layers can be used as features in current application. Pre-trained Net features are extracted using VoxCeleb pre-trained network [15]. VoxCeleb is a large-scale speaker identification dataset created using a fully automated computer vision pipeline. The baseline performance of the dataset was calculated using a CNN based architecture and is available as an open source net. The net is trained with real world utterances of over 1000 celebrities collected from open source media. We used that pre-trained net – VoxCeleb Net – to extract features from the audio signals in the current experiment. It is well established in the literature [13] that different layers of deep net contain characteristic information of the source input. The shallow layers of the deep net represent low dimensional features which are generally local characteristics and the deeper layers represent higher dimensional features which are global properties of the source input. In this work, we extract high dimensional features from FC6 layers which is 4096D and FC7 layer which is 1024D. We also extract information from from POOL5 layer and term it as latent spatial descriptors (LSDs) [13]. LSDs were observed to contain high dimensional spatial information [13], where each dimension of LSD represents response of a specific latent concept. Flattening the LSDs would produce very high dimensional data which are computation intensive. Therefore, the LSDs are encoded to reduce the descriptor dimensions as follows.

Feature Encoding – Fisher Vector

Fisher Vector [13, 22] is used to encode the LSDs that are extracted from the POOL5 layer of VoxCeleb pre-trained net. Gaussian Mixture Model (GMM) is used on randomly selected 1,50,000 descriptors. The parameters obtained from the GMM fitting are termed as \( \theta = (\pi_j, m_j, \Sigma_j; j = 1, 2, k) \), where \( m_j, \pi_j, \) and \( \Sigma_j \) are the mean, prior probabilities and covariance of each distribution. The GMM describes each descriptor \( x_i \) to a mode \( j \) in the mixture with a strength defined by its corresponding posterior probability, as computed in Equation 1.

\[
q_{ij} = \frac{(x_i - m_j)^T \Sigma_j^{-1}(x_i - m_j)}{\sum_{t=1}^{k} (x_i - m_t)^T \Sigma_t^{-1}(x_i - m_t)} \tag{1}
\]

The mean \( (u_{jk}) \) and deviation vectors \( (v_{jk}) \) in each mode \( k \) are computed as follows

\[
u_{jk} = \frac{1}{N \sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \left( \frac{x_{ij} - \mu_{jk}}{\sigma_{jk}} \right) \tag{2}
\]
\[ v_{jk} = \frac{1}{N\sqrt{2\pi}} \sum_{i=1}^{N} q_{ik} \left[ \left( \frac{x_{ji} - \mu_{jk}}{\sigma_{jk}} \right)^2 - 1 \right] \]  

where \( j = 1, 2, \ldots, D \) spans the local descriptor vector dimensions. The resultant vectors \( \{u_{jk}\} \) and \( \{v_{jk}\} \) for each of the \( k \) modes in the Gaussian mixtures are concatenated to produce a \( D \) dimensional vector, where \( D = kxd \). The resultant descriptors are first normalized using power-law normalization [21]. Each component \( v_{j}, j = 1to D \) is modified as \( v_{j} = |v_{j}|^{\alpha} \text{sign}(v_{j}) \), where \( \alpha < 1 \). In this work, the values of \( k = 256 \) and \( \alpha = 0.2 \) are used. The resultant vector is L2-normalized again.

C. Deep Net Features

Deep net features are extracted from the deep nets trained in current experiment. The 141D LLDs extracted using OpenSMILE are used for training this deep net. LLDs are sequence data and therefore recurrent convolutional nets is selected as the deep net architecture. Deep net is implemented using Keras framework and trained on a NVIDIA GeForce GTX 1050 Ti 4GB RAM GPU. The deep nets are trained using the default values of Adam optimizer and categorical cross entropy loss as the cost function. The training is carried out with 50 epochs and a validation split of 25. The input data is shuffled during each training epoch to make the net learn the features better. Numerous experiments are conducted to obtain the best model with changes to the architecture and varying parameters as follows

- Varying the number of CNN layers before the LSTM layer.
- Using Leaky ReLu [19], instead of ReLu as the activation function of the CNN layers.
- Using Batch Normalization (BN) layer before the fully connected layer.
- Using Bidirectional LSTM instead of normal LSTM.
- Other experiments were carried out with GRU and simple RNN in place of LSTM.

Best results are obtained with a combination of CNN and LSTM layers. Two layers of convolutional layers followed by max pool layers are used before the inputs are given to a LSTM layer. The output of the LSTM layer is flattened and two fully connected layers of dimension 1024 and 7 are used. The last layer is a SoftMax layer to get the classification of emotion. Once the classification net is trained, the deep net features are extracted from the fully connected layer with 1024 units. Table I gives the complete architecture of the best model.

| Layer   | Arguments | Output Shape | Parameters |
|---------|-----------|--------------|------------|
| Input   |           | (None, 896, 141) | 0          |
| Conv. Layer | # Filters = 32, Kernel = 3 | (None, 896, 32) | 15368      |
| MaxPooling | Pooling size = 2, Stride = 2 | (None, 448, 32) | 0          |
| Conv. Layer | # Filters = 64, Kernel = 3 | (None, 448, 64) | 6208       |
| MaxPooling | Pooling size = 2, Stride = 2 | (None, 224, 64) | 0          |
| LSTM | Units = 7 | (None, 224, 7) | 2016       |
| Flatten |           | (None, 1568) | 0          |
| Dense | Units = 1024 | (None, 1024) | 1606656    |
| Dense | Units = 7 | (None, 7) | 7175        |
| Activation |         | (None, 7) | 0          |

D. Classification

Normalization is carried out on the fused features and are trained with a linear SVM classifier. One-vs.-all approach is used for the classification where the highest score is used to determine the class of the input sample.

E. Dataset

Experiments are carried out on the benchmark EmoDB [17] dataset. The dataset consists of 535 utterances in German language recorded at 48 kHz, which are later down-sampled to 16 kHz. The utterances were recorded from 10 people of equal gender split up, each speaking from a predefined group of 10 sentences. The utterances were categorized into 7 emotions. 10-fold cross validation [18] is used to evaluate the performance of the proposed approach. Accuracy, F1 score, Precision and Recall performance measures are computed for comparison with other works in the literature.

IV. RESULTS AND DISCUSSION

In this Section, the results obtained from the experiments conducted are presented and analysed from the perspective of the two questions raised in the Introduction.

A. Can deep nets be used as an alternate to traditional feature encoding of local descriptors?

The performance of Fisher Vector encoded LLDs is taken as baseline in Table [1] (Sl. no. 1). The performance of pre-trained VoxCeleb net features is presented in Sl. Nos. 2-4. It is observed that FC6 features perform best compared to the other two layers information. Deep net with different types of architecture is investigated as an alternate encoding technique e.g. Fisher vector encoding in Sl nos. 5-12. The best performance 84.07% is nearly 8% (absolute) more than the Fisher vector encoded LLDs, 78.36%. This clearly shows the usage of deep net as alternate encoding technique to handcrafted features, where source files are not available to the public.

B. Can such deep net encoded features complement other available features for enhancing overall performance?

Information from different layers of VoxCeleb pre-trained network [15] are considered as other available features. Best early fusion was observed by FC6 and FC7 features as seen in Table [1]. These features are fused with best of the deep net features i.e. CNN+LSTM and CNN+BILSTM) network of Table [1] Sl. nos 8, 12. The results obtained are presented in Table [11] Sl. nos 3, 4. The features are observed to be complementing each other enhancing the overall performance.
### TABLE II
**Comparison of Different Feature Extraction Methods**

| Sl. No. | Feature Extraction       | F1 Score | Precision | Recall  |
|---------|--------------------------|----------|-----------|---------|
| 1       | Baseline                 | 78.36    | 86.22     | 77.13   |
| 2       | Pre-trained Net Features | 65.85    | 81.27     | 71.71   |
| 3       | FC6                      | 83.08    | 85.36     | 83.70   |
| 4       | FC7                      | 79.98    | 82.24     | 80.25   |
| 5       | Simple RNN               | 17.41    | 38.58     | 37.06   |
| 6       | LSTM                     | 83.08    | 86.15     | 82.43   |
| 7       | GRU                      | 81.85    | 85.15     | 81.18   |
| 8       | CNN+LSTM                 | 83.59    | 86.31     | 83.52   |
| 9       | CNN(increased filters)+LSTM | 79.72  | 84.10     | 79.35   |
| 10      | CNN+LSTM+BN              | 80.82    | 83.03     | 80.77   |
| 11      | CNN+LSTM+Leaky Relu      | 82.72    | 85.08     | 83.04   |
| 12      | CNN+BILSTM               | 84.07    | 87.19     | 83.83   |

### TABLE III
**Comparison of Different Feature Fusion Techniques**

| Sl. No. | Feature Extraction                  | F1 Score | Precision | Recall  |
|---------|-------------------------------------|----------|-----------|---------|
| 1       | Fusion within pre-trained net features | 81.24    | 83.02     | 81.41   |
| 2       | FC6+FC7                             | 84.07    | 85.97     | 84.19   |
| 3       | FC6+FC7+(CNN+LSTM)                  | 93.62    | 94.36     | 93.51   |
| 4       | FC6+FC7+(CNN+BILSTM)                | 92.55    | 93.85     | 92.37   |

F1 score comparison of individual emotions with different feature extraction techniques is shown in Figure 2.

### C. Comparison with Other Works

Our best results are compared with latest 2018 works in the literature. The best fusion result, Table [IV] S.Nos 3-4, are observed to be better than all the recent most works on EmoDB dataset. Deep learning techniques [5], [7] are observed to perform poorly than handcrafted features [4], [9]. This is due to the small size of dataset. This is also the reason why we did not train a deep network on source files of EmoDB dataset.

The confusion matrix is used to analyze the relation between each emotion. Figure 3 gives the confusion matrix of the (FC6+FC7)+(CNN+LSTM) fusion model. Higher misclassifications can be observed in anger and happiness as they have high values of energy and arousal. This could be reason for the relatively highest misclassification (1%) amongst all possible emotions.

Table [IV] gives the comparison of our best model and the results of different works from related literature.

Our fusion strategy of complementing the segment features extracted from LLDs using deep features performs better than different methods available in the literature. This is evident from the fact that fusion of handcrafted features makes a deep net perform better than conventional methods.

### V. Conclusion

Deep net is observed to successfully function as a feature encoding technique for LLDs in audio signals. Such encoded features are observed to perform better than traditional feature encoding technique viz. Fisher vector. Further, these deep net features are complementary to other available pre-trained network features, enhancing nearly by 8% (absolute).

In current work, pre-trained network was obtained from large scale speaker identification task. Performance of pre-trained network on emotion recognition in conjunction with our proposed deep net features can be undertaken as future work.

### REFERENCES

[1] Massaro and Dominic W, “Illusions and issues in bimodal speech perception,” International Conference on Auditory-Visual Speech Processing, 1998.

[2] Swain, Monorama, Aurobinda Routray, and Prithviraj Kabaspaty, “Databases, features and classifiers for speech emotion recognition: a review,” International Journal of Speech Technology, vol. 21.1, pp. 93-120, 2018.

[3] Zhao, Jianfeng, Xia Mao, and Lijiang Chen, “Learning deep features to recognise speech emotion using merged deep CNN,” IET Signal Processing, 2018.

[4] Demircan, Semiye, and Hamar Kahramanli, “Application of fuzzy C-means clustering algorithm to spectral features for emotion classification from speech,” Neural Computing and Applications, vol. 29.8, pp. 59-66, 2018.
Fig. 3. Confusion Matrix of FC6 + Fc7 + CNN + LSTM Model.

[5] Chen, Mingyi, Xuanji He, Jing Yang, and Han Zhang, “3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition,” IEEE Signal Processing Letters, 2018.

[6] Choi, Woo Yong, Kyu Ye Song, and Chan Woo Lee, “Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data,” Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML), 2018.

[7] Zhang, Shiqing, Shiliang Zhang, Tiejun Huang, and Wen Gao, “Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching,” IEEE Transactions on Multimedia, vol. 20.6, pp. 1576-1590, 2018.

[8] Ma, Yaxiong, Yixue Hao, Min Chen, Jincai Chen, Ping Lu, and Andrej Kos, “Audio-Visual Emotion Fusion (AVEF): A Deep Efficient Weighted Approach,” Information Fusion, 2018.

[9] Deb, Suman, and Samarendra Dandapat, “Multiscale Amplitude Feature and Significance of Enhanced Vocal Tract Information for Emotion Classification,” IEEE Transactions on Cybernetics, 2018.

[10] Jing, Shaoling, Xia Mao, and Lijiang Chen, “Prominence features: Effective emotional features for speech emotion recognition,” Digital Signal Processing, vol. 72, pp. 216-231, 2018.

[11] Lal, G. Jyothish, E. A. Gopalakrishnan, and D. Govind, “Epoch Estimation from Emotional Speech Signals Using Variational Mode Decomposition,” Circuits, Systems, and Signal Processing, pp. 1-30, 2018.

[12] Deng Jun, Xinzhou Xu, Zixing Zhang, Sascha Fruholz and Bjorn Schuller, “Semisupervised autoencoders for speech emotion recognition,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 26.1, pp. 31-43, 2018.

[13] Balaji, B., and V. Oruganti. “Multi-level feature fusion for group-level emotion recognition.” Proceedings of the 19th ACM International Conference on Multimodal Interaction. ACM, 2017.

[14] Poria, Soujanya, Erik Cambria, and Alexander Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis.” Proceedings of the conference on Empirical methods in Natural Language Processing. 2015.

[15] Nagrani, Arsha, Joon Son Chung, and Andrew Zisserman. “Voxceleb: a large-scale speaker identification dataset.” Interspeech, 2017.

[16] Florian Eyben, Felix Weninger, Florian Gross, Bjorn Schuller: Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor,” In Proc. ACM Multimedia (MM), pp. 835-838, Oct. 2013.

[17] Burkhardt, Felix, et al. “A database of German emotional speech.” Ninth European Conference on Speech Communication and Technology, 2005.

[18] Longbiao Wang, Jianwu Dang, Linjuan Zhang, Haotian Guan, “A feature fusion method based on extreme learning machine for speech emotion recognition,” IEEE SigPort, 2018.

[19] Maas, Andrew L., Awni Y. Hannun, and Andrew Y. Ng. “Rectifier nonlinearities improve neural network acoustic models.” Proc. ICML., vol. 30.1, 2013.

[20] Schuller, Bjorn, et al. “The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism,” Proceedings INTERSPEECH 2013.

[21] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” ICLR 2015.

[22] F. Perronnin and C. Dance, “Fisher Kernels on Visual Vocabularies for Image Categorization,” Computer Vision and Pattern Recognition (CVPR), 2007.