Speech Emotion Recognition using pre-trained and fine-tuned transfer learning approaches

Adil CHAKHTOUNA[1], Sara SEKKATE[2], Abdellah ADIB[1]

[1] *Team Computer Science, Artificial Intelligence & Big Data, MCSA Laboratory, Faculty of Sciences and Technologies of Mohammedia, Hassan II University of Casablanca*; adilchakhtouna10@gmail.com; abdellah.adib@fstm.ac.ma
[2] *Higher National School of Arts and Crafts of Casablanca*; sarasekkate@gmail.com

Abstract. Speech Emotion Recognition (SER) seems to be an exciting task due to its promising uses in many areas. A large number of studies have extracted acoustic descriptors from input speech and used them as features to train and test emotion recognition models. On the other hand, dimensional representations of speech signal such as spectrograms have been explored to appraise deep neural networks and more specifically the Convolutional Neural Network (CNN). To address this last issue, the current study aims to take advantage of the implementation of transfer learning architectures by using Mel spectrograms feature extraction mechanism to transform speech cues into images and then forward them to the pre-trained models. As well as make a comparative analysis of the performance using four different pre-trained architectures (VGG-16, VGG-19, EfficientNet B0 and EfficientNetV2 B0) both including and excluding the fine-tuning process. The experimental results were performed with two different configurations, pre-trained and fine-tuned models, the VGG-19 has obtained the highest accuracy of 79.83% on the held-out test. An evaluation carried out on the Italian Database of Elicited Mood in Speech (DEMoS) proves the effectiveness of the suggested model, which outperforms the state-of-the-art in SER.

Keywords: Transfer learning, Speech Emotion Recognition, Mel spectrogram, VGG-16, VGG-19, EfficientNet B0, EfficientNetV2 B0, DEMoS.

1 Introduction

One of the primary forms of communication between humans around the world is speech. Emotions can be characterized throughout speech and they play an influential part in the human communication process. Speech Emotion Recognition (SER) appears on the horizon as a promising task by virtue of giving computers the ability to understand and recognize various emotional aspects in speech. The future advances in human-computer communication technologies are anticipated to enhance emotional competence through the diversity of its multiple pathways, such as body language [1], facial expressions, electroencephalography (EEG) [2] and speech [3].

So far, SER has created a growing worldwide interest from researchers and scientists as a result of its significance in exploring the emotional aspects of humans and their relationship towards the development of a variety of multidisciplinary fields including psychological sciences, knowledge systems and health informatics. The critical issue facing investigators in the field of SER relates to the way of extracting reliable and discriminative features from the vocal content of a spoken utterance. This task has been approached by a large number of studies, each in its own way. Spectral and prosodic features including fundamental frequency, energy, zero-crossing rate, Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coefficients (LPC) are part of the most widely used ones in SER [4]. Furthermore, there are several kinds of machine learning techniques that are applied to understand the connection between the extracted features of speech and the predefined emotion labels. Regarding modeling approaches in SER, Support Vector Machine (SVM) [5], ensemble learning [6], Hidden Markov Model (HMM) [7] and, more recently, the involvement of deep neural networks such as Recurrent Neural Network (RNN) [8] and Convolutional Neural Network (CNN) [9] have been addressed.

New investigative approaches have been recently proposed to analyze speech signals based on the transfer learning technique. The primary goal of transfer learning is to take the learning gained from the baseline task and use it to help the target task learn. the current study aims to take advantage of the implementation of well-known transfer learning architectures and addresses the SER task as an image classification problem by using Mel spectrograms feature extraction mechanism to transform speech cues into images and then forward them to the pre-trained and fine-tuned models.

The body of this document is structured in the following manner: Section 2 briefly reviews the existing works in SER. Section 3 outlines the proposal's methodology. Section 4 explains in detail the different experiences and presents the main results achieved in this study. Finally, in Section 5, we summarize the most relevant findings and give future guidelines for investigation in SER.

2 Related works

Over the last few decades, several surveys have been conducted on the SER, each according to its own point of view. The traditional pioneer approaches start by extracting key features from the speech signals and then ranking various classes of emotions depending on the feature values. Alternatively, emerging deep learning methods accomplish the SER task by merging the two steps into a single process. Authors in [10], introduce a multi-resolution feature extraction based on MFCC features, which are obtained from the Discrete Wavelet Transform (DWT) sub-band coefficients and mixed with standard MFCC and pitch features. The feature set was reduced using Linear Discrimination Analysis (LDA) before its input fed into a naive Bayes classifier. The proposed methodology is evaluated on the German Emotional Speech Database (EMO-DB) with a variety of noise types such as restaurant, train, exhibition, street, car, bab-

ble and airport. The average accuracies of 80.10% and 89.09% were achieved in the speaker-independent and speaker-dependent configurations, respectively, at a Signal-to-Noise-Ratio (SNR) of 30dB.

In [11], the author conducted a combination of deep features extracted from deep learning architectures including VGG-16, ResNet18, ResNet50, ResNet101, SqueezeNet, and DenseNet201 and acoustic features such as Root Mean Square energy (RMS), MFCC and zero-crossing rate. The ReliefF feature selection algorithm was used to choose the most effective features from the combined feature set. The chosen attributes were given to SVM-based modeling to perform the classification task. The highest accuracies achieved with ResNet101 deep features + acoustic features were 79.41% and 90.21% for the RAVDESS and EMO-DB datasets, respectively. While for the IEMOCAP database, an accuracy of 85.37% with VGG-16 deep features + acoustic features was achieved. In [12], an end-to-end trained model to recognize continuous emotions was employed, it consists of a stacked CNN with 2 layers of Long Short-Term Memory (LSTM). The idea is mainly based on retrieving features from the raw signal rather than on extracting hand-created features. The presented model outperforms the state-of-the-art on the RECOLA database having the concordance correlation coefficients of 0.815 and 0.502 for the excitation and valence dimensions, respectively, on the validation set.

The SER system proposed by Sun and Zou [13] was built with a DNN-decision tree SVM framework, the authors' contribution is based on the calculation of the degree of emotional confusion. At each leaf of the tree, different DNNs were trained to extract deep characteristics to train every SVM classifier in the decision tree. The experimental performance conducted on the Chinese CASIA database indicates that the overall recognition rate has increased by 2.91% and 6.25% compared to the DNN-SVM and conventional SVM classification schemes, respectively. To improve the efficiency of SER systems in cross-language and cross-corpus situations, a transfer learning method based on Deep Belief Network (DBN) has been proposed in [14]. The eGeMAPS [15] feature set extracted with openSMILE[1] toolbox were used. Different experimental setups including within corpus and multi-language schemes, were undertaken to test the performance of the DBN in comparison with the sparse autoencoder (AE) with SVM. Five corpora of German, English and Italian languages were investigated. The proposed method will prove helpful for developing strong SER system involving data from multiple languages.

The main contribution of this study is manifested by taking benefit of well-known CNN architectures such as VGG and EfficientNet, which are pre-trained on the large ImageNet database, and also addresses the issue of SER as an image classification task. The impact of transfer learning in SER is demonstrated by performing the fine-tuning process.

[1] https://www.audeering.com/research/opensmile/

3 Methodology

In this section, we propose a SER system that relies on the use of Mel spectrogram to model the speech signal as an image. The proposed SER framework has two major stages, as depicted in Fig. 1: it starts by transforming all speech utterances into Mel spectrogram representations and apply then transfer learning configurations to recognize several emotions. Further details are provided in the following subsections.

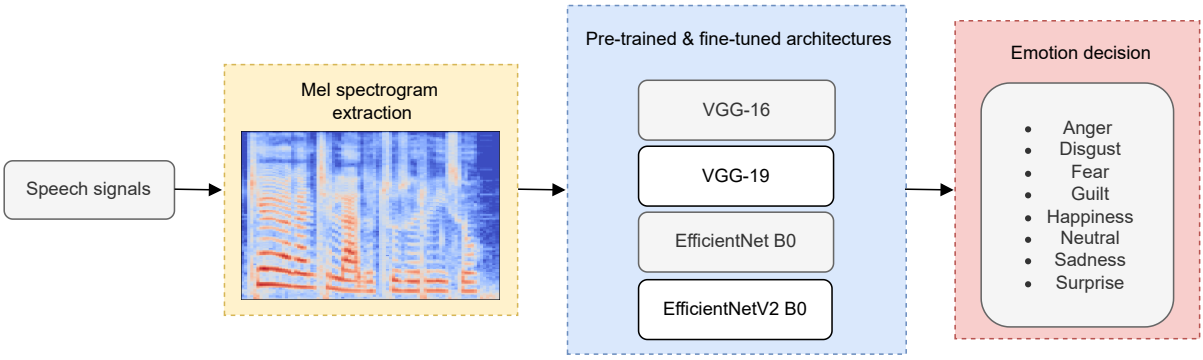

Fig. 1. Diagram of the proposed SER system.

3.1 Emotional database

The use of database of Elicited Mood in Speech (DEMoS) [16] was proposed in this study. It is an extensive new database with an Italian emotional speech. The access to the data requires the signature of an academic license agreement. DEMoS was generated by 68 native speakers, including 23 females and 45 males, comprising 9697 induced emotional samples from eight emotions, the 'big six' according to Ekman [17]: happiness, anger, sadness, surprise, fear, disgust, the secondary emotion guilt and the neutral state.

3.2 Mel spectrogram extraction

From all previous work related to SER, we can infer that finding the appropriate input features can make a difference in the performance of emotion recognition. Speech properties can be modeled using a one-dimensional vector or two- or three-dimensional representations, depending on the type of considered features. Among the most common ways to simultaneously observe frequency, time, and intensity variations is via speech spectrogram, which maps the energy of speech over time and frequency. Mel spectrogram provide the time-frequency representation of speech signal.

To obtain the Mel spectrogram, the magnitude spectrogram is first computed, and then mapped onto the Mel scale. Furthermore, each speech signal is divided into equal number of frames, and each frame is produced by overlapping the preceding frame by half. The windowing process is applied to each frame in order to prevent the discontinuities arising at the extremities of each frame. The widely recommended Hamming window function is employed in this study. Hence, the Fourier spectrum is revealed on each shorter segment to transform it from time domain to frequency domain. Finally, the power spectrograms of each signal processed by the Fourier transform are extracted. In order to get the same dimension for all Mel spectrograms, short speech signals must be padded by a segment of their part to reach the maximum length of the longer statement in the dataset. Because the speech utterances in the DEMoS database have not the same length.

3.3 Transfer learning architectures

In our study, the aim of representing speech signals as Mel spectrograms (images) is to take advantage of the use of advanced deep learning algorithms dedicated to the field of computer vision, in particular CNN architectures. In this section, we discuss four existing CNN architectures: VGG-16, VGG-19, EfficientNet B0 and EfficientNetV2 B0.

A- VGG-16 and VGG-19

VGG-16 [18] and VGG-19 [18] are two commonly known CNN architectures used in a variety of image classification tasks. The principal concept behind the VGG model is to increase the overall performance of the network by expanding the depth of its layers, therefore its strategy relies on the application of convolution kernels to transform the input into several small volumes of kernels, thereby minimizing the number of parameters in the model, and allowing the network to be more discriminative.

The structure of the pre-trained VGG-16 architecture used on the ImageNet dataset, comprising a total of 21 main layers, includes 13 convolutional layers, 5 max pooling layers and 3 Fully Connected (FC) layers. In addition, a deeper network than the VGG-16, called the VGG-19, is introduced. The VGG-19 possesses 24 fundamental layers, with 16 convolutional layers, five max pooling layers and three FC layers.

B- EfficientNet B0 and EfficientNetV2 B0

EfficientNet B0 [19] provides a rule-based approach for scaling patterns and increasing the performance of the CNN. It is generally accepted that improving CNN performance necessitates dimensional scaling in terms of depth, width, or resolution. The basic EfficientNet B0 is then scaled using a compound scaling approach to produce a family of patterns from B1 to B7. EfficientNetV2 B0 [20], a novel class of CNN that exhibit faster learning speed and better parameter effectiveness than earlier models.

The two architectures contain 7 blocks. These blocks also vary in the number of sub-blocks, which increases as we move from EfficientNet B0 to EfficientNetV2 B0. As a result, the total number of layers in EfficientNet B0 is 237, whereas in EfficientNetV2 B0 is 255.

In this paper, we explore how well the four architectures including VGG-16, VGG-19, EfficientNet B0, and EfficientNetV2 B0 transfer to the SER classification task by comparing their performance with other related studies.

4 Experiments and results

4.1 Implementation details

In this subsection, we provide further details on the implementation parameters used in the different experimental setups. We employ the Google Colab environment to train several fine-tuned models. To secure the validity of our results, the Mel spectrogram images extracted from the DEMoS dataset were divided into 70% for training and 30% for testing. The architecture of the used FC layers added in the top of all the four pre-trained models. The FC layers block consists of three dense layers and and dropout layer, the detailed parameters of the FC layers are shown in Table 1.

Table 1. Parameters of the FC layers.

Layer	**Parameter value**	**Activation function**
Dense	512 (nodes)	Relu
Dropout	0.5 (probability)	-
Dense	256 (nodes)	Relu
Dense	8 (nodes)	Softmax

4.2 Fine-tuned models

The fine-tuning technique is the main part in the transfer learning approach. It allows to carefully choose the layers to be fine-tuned in the model to obtain a better result. To improve the accuracy of our pre-trained models, we conducted two configurations for fine tuning using our Mel spectrogram images for the SER classification task. In the first configuration (VGG-16-C1, VGG-19-C1, EfficientNet B0-C1, and EfficientNetV2 B0-C1), we froze all pre-trained layers of the used architectures and replaced the last FC layers used on the ImageNet dataset with our FC layers. Alternatively, a part of the pre-trained model can be considered in the fine-tuning stage together with the FC layers. In the second configuration (VGG-16-C2, VGG-19-C2, EfficientNet B0-C2, and EfficientNetV2 B0-C2), in addition to the FC layers, we have fine-tuned the fourth and fifth convolutional blocks of the VGG-16-C2 and VGG-19-C2. Furthermore, we fine-tuned 32 layers for both EfficientNet B0-C2 and EfficientNetV2 B0-C2.

4.3 Results and discussion

In this section, we provide all the results obtained using our pre-trained models. Table 2 displays the results obtained on the test set using several measures, including precision, recall, F1-score, and accuracy. According to Table 2, the VGG-19-C2 architecture outperforms the other architectures in all evaluation metrics. It achieved the accuracy of 80% followed by the VGG-16-C2, EfficientNet B0-C2, EfficientNetV2 B0-C1, EfficientNet B0-C1, EfficientNetV2 B0-C2, VGG-19-C1 and VGG-16-C1 with respective accuracies of 71%, 68%, 66%, 65%, 64%, 57% and 55%. We can confirm that for all the pre-trained models, the second fine-tuning configuration gives better results compared to the first one, except for the EfficientNetV2 B0 model which has the opposite case.

Table 2. The obtained results of the pre-trained models using both fine-tuning configurations.

Model	Precision	Recall	F1-score	Accuracy
VGG-16-C1	0.57	0.53	0.54	0.55
VGG-16-C2	0.70	0.69	0.69	0.71
VGG-19-C1	0.59	0.55	0.56	0.57
VGG-19-C2	**0.79**	**0.80**	**0.79**	**0.80**
EfficientNet B0-C1	0.66	0.62	0.63	0.65
EfficientNet B0-C2	0.69	0.66	0.67	0.68
EfficientNetV2 B0-C1	0.70	0.63	0.65	0.66
EfficientNetV2 B0-C2	0.66	0.63	0.64	0.64

The confusion matrix of the best architecture among the used models (VGG-19-C2) is presented in Table 3. The most recognized samples for all emotions are placed on the diagonal of the confusion matrix. The recognition rates of the majority of the emotions were anticipated with remarkable percentages. Recall values of 79%, 85%, 78%, 73%, 79%, 80%, 83% and 81% were reached for anger, disgust, fear, guilt, happiness, neutral, sadness and surprise, respectively. The highest misclassified part comes from guilt emotion, in which 9% of its samples are incorrectly classified as sadness.

4.4 Comparison with other studies

In this section, the only related work we found for the DEMoS dataset was proposed by [21], whose model was trained on the DEMoS dataset without the neutral emotion. To make a fair comparison, we repeated all the experiments without considering the neutral state, then we chose the best pre-trained architecture among the four used models for comparison with the existing work.

Table 3. The confusion matrix of the VGG-19-C2 architecture.

Emotion	**Anger**	**Disgust**	**Fear**	**Guilt**	**Happiness**	**Neutral**	**Sadness**	**Surprise**
Anger	**0.79**	0.06	0.05	0.01	0.02	0.01	0.03	0.02
Disgust	0.03	**0.85**	0.02	0.02	0.03	0.01	0.03	0.01
Fear	0.03	0.05	**0.78**	0.03	0.04	0.03	0.03	0.02
Guilt	0.01	0.07	0.03	**0.73**	0.06	0.01	0.09	0.01
Happiness	0.02	0.06	0.02	0.03	**0.79**	0.02	0.04	0.02
Neutral	0.01	0.02	0.03	0.05	0.04	**0.80**	0.01	0.04
Sadness	0.04	0.05	0.02	0.03	0.02	0.01	**0.83**	0.01
Surprise	0.03	0.03	0.03	0.01	0.03	0.01	0.04	**0.81**

Gerczuk et al. [21] proposed to train the ResNet architecture from scratch on Mel spectrograms extracted from the DEMoS database. They attained a 73.8% recognition rate on the test set. In contrast, we obtained a superior performance of 79.7% in our work. It was better if the authors had employed the fine-tuning technique on the ResNet architecture, this might improve the recognition rate of their study.

Table 4. Results obtained by our proposed method and other state-of-the-art work for SER.

Work	**Model**	**Training Technique**	**Dataset**	**Feature**	**Accuracy**
[21]	ResNet	Trained from scratch	DEMoS without neutral	Mel spectrogram	73.8%
Our	VGG-19-C2	Transfer learning	DEMoS without neutral	Mel spectrogram	**79.7%**

5 Conclusion

In this study, we approached the SER task based on the use of Mel spectrogram feature represented as an image. This approach allowed us to overcome the SER challenge by applying the transfer learning approach that enables us to leverage the use of pre-trained models such as VGG-16, VGG-19, EfficientNet B0 and

EfficientNetV2 B0. Depending on the experimental results, by using the fine-tuning technique, the second fine-tuned configuration of the VGG-19 architecture achieved the highest recognition rate compared to the other architectures.

Acknowledgements

This work was supported by the Ministry of Higher Education, Scientific Research and Innovation, the Digital Development Agency (DDA) and the CNRST of Morocco (Alkhawarizmi/2020/01).

References

1. F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2017.
2. Y. Tan, Z. Sun, F. Duan, J. Solé-Casals, and C. F. Caiafa, "A multimodal emotion recognition method based on facial expressions and electroencephalography," *Biomedical Signal Processing and Control*, vol. 70, p. 103029, 2021.
3. S. Sekkate, M. Khalil, A. Adib, and S. Ben Jebara, "An investigation of a feature-level fusion for noisy speech emotion recognition," *Computers*, vol. 8, no. 4, p. 91, 2019.
4. A. Chakhtouna, S. Sekkate, and A. Adib, "Improving speech emotion recognition system using spectral and prosodic features," in *International Conference on Intelligent Systems Design and Applications*, pp. 399–409, Springer, 2021.
5. M. J. Al Dujaili, A. Ebrahimi-Moghadam, and A. Fatlawi, "Speech emotion recognition based on svm and knn classifications fusion," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, p. 1259, 2021.
6. W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1845–1854, 2021.
7. B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, vol. 2, pp. II–1, Ieee, 2003.
8. D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Systems with Applications*, vol. 173, p. 114683, 2021.
9. A. B. A. Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional neural network (cnn) based speech-emotion recognition," in *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pp. 122–125, IEEE, 2019.
10. S. Sekkate, M. Khalil, A. Adib, and S. Ben Jebara, "A multiresolution-based fusion strategy for improving speech emotion recognition efficiency," in *International Conference on Mobile, Secure, and Programmable Networking*, pp. 96–109, Springer, 2019.
11. M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020.
12. P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5089–5093, IEEE, 2018.

13. L. Sun, B. c, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on dnn-decision tree svm model," *Speech Communication*, vol. 115, pp. 29–37, 2019.
14. S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," *arXiv preprint arXiv:1801.06353*, 2018.
15. F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.
16. E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "Demos: An italian emotional speech corpus," *Language Resources and Evaluation*, vol. 54, no. 2, pp. 341–383, 2020.
17. P. Ekman, "Expression and the nature of emotion," *Approaches to emotion*, vol. 3, no. 19, p. 344, 1984.
18. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
19. M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.
20. M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International Conference on Machine Learning*, pp. 10096–10106, PMLR, 2021.
21. M. Gerczuk, S. Amiriparian, S. Ottl, and B. W. Schuller, "Emonet: a transfer learning framework for multi-corpus speech emotion recognition," *IEEE Transactions on Affective Computing*, 2021.