Emotions Understanding Model from Spoken Language using Deep Neural Networks and Mel-Frequency Cepstral Coefficients

Marco Giuseppe de Pinto
University of Bari Aldo Moro
via E. Orabona 4 Bari, Italy
marcogiuseppe.depinto@uniba.it

Marco Polignano
University of Bari Aldo Moro
via E. Orabona 4 Bari, Italy
marco.polignano@uniba.it

Pasquale Lops
University of Bari Aldo Moro
via E. Orabona 4 Bari, Italy
pasquale.lops@uniba.it

Giovanni Semeraro
University of Bari Aldo Moro
via E. Orabona 4 Bari, Italy
giovanni.semeraro@uniba.it

***Abstract*—The ability to understand people through spoken language is a skill that many human beings take for granted. On the contrary, the same task is not as easy for machines, as consequences of a large number of variables which vary the speaking sound wave while people are talking to each other. A sub-task of speeches understanding is about the detection of the emotions elicited by the speaker while talking, and this is the main focus of our contribution. In particular, we are presenting a classification model of emotions elicited by speeches based on deep neural networks (CNNs). For the purpose, we focused on the audio recordings available in the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. The model has been trained to classify eight different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprise) which correspond to the ones proposed by Ekman plus the neutral and calm ones. We considered as evaluation metric the F1 score, obtaining a weighted average of 0.91 on the test set and the best performances on the "Angry" class with a score of 0.95. Our worst results have been observed for the sad class with a score of 0.87 that is nevertheless better than the state-of-the-art. In order to support future development and the replicability of results, the source code of the proposed model is available on the following GitHub repository: https://github.com/marcogdepinto/Emotion-Classification-Ravdess**

***Keywords*—emotion detection, natural language understanding, sentiment analysis, deep learning, machine learning, classification, mel-frequency cepstral coefficients, cnn, ravdess.**

I. Introduction

The human communication through the spoken language is the base for information exchange and it is the main aspect of the society since the first human settlements. In the same way, emotions go back to a primordial instinct prior to the spoken language that we know today and that can be considered as the first natural communication strategies. In today's common languages, the word is followed by emotions in order to make communication more direct, clear and understandable to all. The approach usually actioned, when it is required to face emotion predictions, is more based on human brain then machines. As humans, we consider facial expressions, volume and tone of the speaker' voice and many other factors, while our brain works on connecting all the different informations that we will use to interact with other people. In the area of human-machine interaction, we have tried for a long time to provide the ability to understand the language spoken on a computer. Today, many tools come close to this ability, even if the interpretation of emotions during the dialogue is a characteristic often overlooked. Due to the importance of including this aspect in a more complex and complete model of human-machine interaction, in this work we focused on an efficient strategy to be able to identify the main emotion expressed by a subject during the dialogue. In the literature it is shown how each of us expresses more than one basic emotion [4] (six Ekman's emotions [2]) at a time but our opinion is that it is extremely difficult, both for the speaker and for the listener, to be able to recognize which and in what percentage of the emotions are mixed. In this regard it was decided to create a model that aims to identify only the emotion that has a greater value in the audio track. Different approaches have been tried to have a machine classify feelings like computer vision or text analytics. In this work, our goal is to use pure audio data considering Mel-frequency cepstral coefficients[10] (from now on MFCC).

II. Related Work

In this field of research, many classification strategies have been presented in the last years. One of the system, proposed by Iqbal et al.[6] used Gradient Boosting, KNN and SVM to work on a granular classification on the RAVDESS dataset used in this work to identify differences based on gender with approximately between 40% and 80% overall accuracy, depending on the specific task.
In particular, the proposed classifiers performed differently in different datasets but we considered only the work done on RAVDESS for the scope of this paper. In Iqbal et al.[6] work, three types of datasets have been created including only male recordings, only female recordings and a combined one. In RAVDESS (male) SVM and KNN have 100% accuracy in both anger and neutral, but in happiness and sadness Gradient Boosting performs better than SVM and KNN. In RAVDESS (female) SVM achieves 100% accuracy in anger as same as male part. SVM has overall good performance except in sadness. Performance of KNN is also good in anger and neutral like 87% and 100% respectively. In anger and

978-1-7281-4384-2/20/$31.00

neutral, Gradient Boosting performs poorly. KNN performance is very poor in happiness and sadness comparing with other classifiers. In male and female combined dataset, performances of SVM and KNN are really good in anger and neutral rather than Gradient Boosting. KNN's performance is really poor in happiness and sadness. Average performances of classifiers in male dataset are better than female dataset except SVM. In combined database, SVM get high accuracy than gender based datasets. Another approach presented by Jannat et al.[7] achieved 66.41% accuracy on audio data and more than 90% accuracy mixing audio and video data. In particular, given pre-processed image data that includes faces and audio waveforms, Jannat et al.[7] trained 3 separately deep networks: one network only on image data, another only on the plotted audio waveforms, and the third on both image and waveform data. One of the first approaches that used the RAVDESS dataset, but classifying only some of the emotions available, was published by Zhang et al.[16], reaching an overall accuracy score higher than the model proposed in this work but using less classes. Giving more details, in Zhang et al.[16] three shared emotion recognition models for speech and song have been proposed: a simple model, a single-task hierarchical model and a multi-task hierarchical model. The simple model creates a single classifier, independent of domain. The two hierarchical models use domain during training. The single-task model trains a separate emotion classifier for each domain. The multi-task model trains a multi-task classifier to jointly predict emotion across both domains. In the testing phase, the testing data are separated based on the predicted domain. The data are analyzed using the classifier corresponding to the estimated domain. The work have been conducted adopting the directed acyclic graph SVM (DAGSVM)[14].

III. The Proposed Model

The model of classification of emotions here proposed is based on a deep learning strategy base on convolutional neural networks (CNN) and dense layers [8]. The key idea is considering the Mel-frequency cepstral coefficients[1], [10] (MFCC), commonly referred to as the "spectrum of a spectrum", as the only feature to train the model. MFCC is a different interpretation of the Mel-frequency cepstrum (MFC), and it has been demonstrated to be the state of the art of sound formalization in automatic speech recognition task [11]. The MFC coefficients have mainly been used has the consequence of their capability to represent the amplitude spectrum of the sound wave in a compact vectorial form. As described in [10], the audio file is divided into frames, usually using a fixed window size, in order to obtain statistically stationary waves. On the small frames obtained, the Discrete Fourier Transformation is applied, and only the logarithm of the amplitude spectrum is kept. The amplitude spectrum is normalized with a reduction of the "Mel" frequency scale. This operation is performed for empathizing the frequency more meaningful for a significant reconstruction of the wave as the human auditory system can perceive. For each audio file, 40 features have been extracted. The feature has been generated converting each audio file to a floating point time series. Then, a MFCC sequence has been created from the time series. The MFCC array has been transposed and the arithmetic mean has been calculated on its horizontal axis. The MFCC calculations are deeply explained in the article of Davis S. et al.[1] and in the book of Huang et al.[5].

Layer (type)	Output Shape	Param #
conv1d_3 (Conv1D)	(None, 40, 128)	768
activation_4 (Activation)	(None, 40, 128)	0
dropout_3 (Dropout)	(None, 40, 128)	0
max_pooling1d_2 (MaxPooling1	(None, 5, 128)	0
conv1d_4 (Conv1D)	(None, 5, 128)	82048
activation_5 (Activation)	(None, 5, 128)	0
dropout_4 (Dropout)	(None, 5, 128)	0
flatten_2 (Flatten)	(None, 640)	0
dense_2 (Dense)	(None, 8)	5128
activation_6 (Activation)	(None, 8)	0

Total params: 87,944
Trainable params: 87,944
Non-trainable params: 0

Fig. 1. Detailed description of the architecture of the proposed classifier

The deep neural network designed for the classification task is reported operationally in Fig. 1. The network is able to work on vectors of 40 features for each audio file provided as input. The 40 values represent the compact numerical form of the audio frame of 2s length. Consequently, we provide as input a of size $< number\_of\_training\_files > x\,40\,x\,1$ on which we performed one round of a 1D CNN with a ReLu activation function [12], dropout of 20% and a max-pooling function 2 x 2. The rectified linear unit (ReLu) can be formalized as $g(z) = max\{0, z\}$, and it allows us to obtain a large value in case of activation by applying this function as a good choice to represent hidden units. Pooling can, in this case, help the model to focus only on principal characteristics of every portion of data, making them invariant by their position. We have run the process described once more by changing the kernel size. Following, we have applied another dropout and then flatten the output to make it compatible with the next layers. Finally, we applied one Dense layer (fully connected layer) with a *softmax* activation function, varying the output size from 640 elements to 8 and estimating the probability distribution of each of the classes properly encoded (0=Neutral; 1= Clam; 2= Happy; Sad=3; Angry=4; Fearful= 5; Disgust=6; Surprised=7).

IV. Evaluation of the Model

The evaluation of the proposed model has been carried out in order to to investigate whether the model produces results of accuracy that are good enough to produce interesting considerations to be used in future work on the subject that include speeches in real noisy domains. Different models of classification have been evaluated beyond that proposed by us so that it is possible to generate baselines for the results obtained. As a first approach, a decision tree (DT) and a random forest (RF) classifiers with 1000 trees have been performed. The two models have been implemented using

Scikit-learn [1] [13] Python library with default parameters.

Dataset. The dataset used in this work is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[9]. This dataset contains 7356 files made by 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound). Ratings were provided by 247 individuals who were characteristic of untrained adult research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity, interrater reliability, and test-retest intrarater reliability were reported.

Each of the 7356 RAVDESS files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 02-01-06-01-02-01-12.mp4). These identifiers define the stimulus characteristics:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only);
- Vocal channel (01 = speech, 02 = song);
- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised);
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion;
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door");
- Repetition (01 = 1st repetition, 02 = 2nd repetition);
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

Enrichment of training data. As deep learning models struggle with little amount of data and this dataset does not have a great amount of features, a pipeline has been created to enrich the training and the test set. In particular, using the video available in the dataset we have extracted with the FFMPEG library a new set of features with a different frequency: the audio files of the dataset have a frequency of 48MHz, while the ones that have been extracted from the videos have a frequency of 44,1MHz, thus introducing some (necessary) noise but still being able to increment the dimension of the training and the test set.

Metrics, data splitting and experimental runs The files have been randomly splitted in two train and test datasets with a test set composed with the 33% of the original full dataset. Consequently, the training set is composed by a 3315 MFCC vector of 40 features obtained thanks to the LibROSA library [2]. The test set shape is 1633 x 40. No cross-validation set has been used for this task. Both the sets are already labeled with the class value encoded as described before. We used as evaluation metric the F1 score as a compact indicator of the quality of the classifier and as standard for comparing our results with them at state of the art. The model has been trained using the sparse categorical cross entropy loss function [3] and rmsprop optimizer for 1000 epochs and the best models have been used for the classification phase. The number of batches has been set to 16 for optimization reasons. During the training we validate the model using the accuracy score as common in deep learning architectures.

V. Discussion of results

The results obtained from the evaluation phase show the effectiveness of the model compared to the baselines and the state of the art on the RAVDESS dataset. In particular, Tab. I shows the values of precision, recall and F1 obtained for each of the emotional classes. These results show us that precision and recall are very balanced, allowing us to obtain F1 values distributed around the value 0.90 for almost all classes. The small variability of F1 results point out the robustness of the model that effectively manages to correctly classify emotions in eight different classes. The classes "Sad" and "Surprised" are the ones in which the model is less accurate, but this result does not surprise us because it is known in the literature that they are the most difficult classes to identify not only by speech but also while observing facial expression or analysing written text [15]. In order to evaluate the effectiveness of the classification of emotions proposed in this work, we decided to compare it with the results obtained from two baselines decision tree (DT) and random forest (RF) and the works of Iqbal et al.[6] and Zhang et al.[16]. The results shown in Tab II allow us to observe how the F1 values of our model are better than baselines and competitors on all classes except "Angry" and "Happy". However, it is necessary to point out that the drop in performance is minimal, and the model works on four classes more than the model proposed by Iqbal et al.[6] and one class more than the model of Zhang et al.[16]. It is therefore well known that as the number of classes increases, the classification task becomes more complex and loses its accuracy. Nevertheless, the CNN-MFCC model proposed here manages to obtain a score of F1 that, on average, is equivalent to that of the two jobs we have been confronted with. A further index of model reliability can be found in Fig. 2 and Fig. 3. In the first one, it is possible to observe how the value of loss (error in the accuracy of the model) tends to decrease both on the test set and on the training set up to the 1000th epoch. The decrease is less evident from the 400th epoch but still perceptible. In Fig. 3, it is reported the average value of accuracy on all the classes that, to the contrary of the loss, increases with the increases of the ages. Such values of loss and accuracy do not differ much among the training and test dataset, allowing us to affirm that the model does not turn out to be overfitted while training. The consequence of this is, in fact, in line with the F1 scores previously observed.

[1] https://scikit-learn.org/stable/

[2] https://librosa.github.io/librosa/index.html

TABLE I. RESULTS OF THE MODEL ON THE TEST SET PER EACH CLASS.

Emotion	precision	recall	F1-score	support
Angry	0.93	0.91	0.92	134
Happy	0.92	0.93	0.92	251
Neutral	0.91	0.89	0.90	242
Sad	0.84	0.90	0.87	271
Calm	0.96	0.94	0.95	253
Fearful	0.92	0.91	0.91	239
Disgusted	0.95	0.93	0.94	127
Surprised	0.90	0.85	0.88	116
accuracy			*0.91*	*1633*
macro avg	*0.92*	*0.91*	*0.91*	*1633*
weighted avg	*0.91*	*0.91*	*0.91*	*1633*

TABLE II. F1-SCORE FOR EACH CLASS COMPARED TO THE BASELINES (DT, RF) AND THE STATE OF ART.

Class	DT	RF	SVM [6]	SVM [16]	CNN-MFCC
Angry	0.84	0.88	**1.0**	0.86	0.95
Happy	0.76	0.78	0.66	**0.92**	0.90
Neutral	0.77	0.69	0.93	0.76	**0.92**
Sad	0.73	0.78	0.67	0.71	**0.87**
Calm	0.85	0.71	-	0.88	**0.92**
Fearful	0.79	0.75	-	0.86	**0.91**
Disgusted	0.74	0.65	-	-	**0.94**
Surprised	0.75	0.68	-	0.71	**0.88**

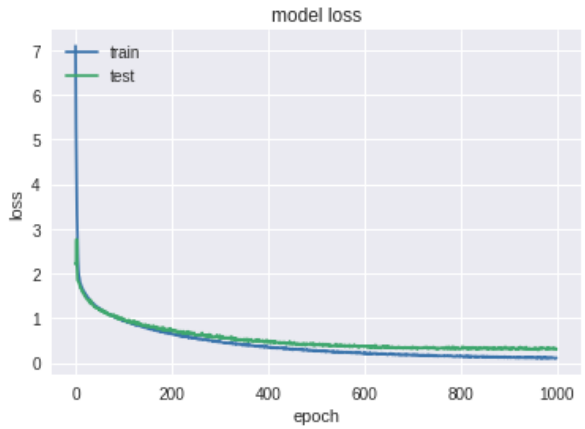

Fig. 2. Trend of the cost function of our deep learning model over 1000 epochs.

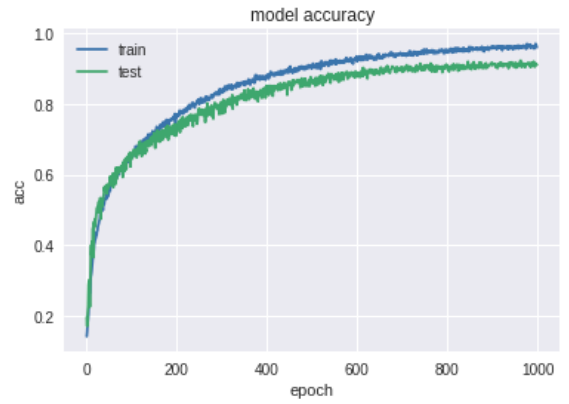

Fig. 3. Trend of the accuracy of our deep learning model over 1000 epochs.

The authors consider the results encouraging: having a dataset bigger than the RAVDESS available, the MFCC can probably be a valid emotion detection feature. We are sure enough that the same model structure can perform similar results also on audio sound files less structured and collected directly in a real noise environment. The MFCC transformation is always applicable, and using strategies of noise reduction and enough training data; the same model could perform well. As a consequence of this, we are working on experimenting with pieces of dialog directly collected from real users as a future extension of this work.

VI. CONCLUSION

In this work, we presented an architecture based on deep neural networks for the classification of emotions using audio recordings from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). The model has been trained to classify seven different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprised) and obtained an overall F1 score of 0.91 with the best performances on the angry class (0.95) and worst on the sad class (0.87). To obtain such result, we extracted the MFCC features (spectrum-of-a-spectrum) from the audio files used for the training. On the above representations of input data, we trained a deep neural network that uses 1D CNNs, max-pooling operations and Dense Layers to estimate the probability of distribution of annotation classes correctly. The approach was tested on the data provided by the RAVDESS dataset. As baseline for our task, we considered is a random forest classifier we trained on the same dataset achieving an average F1 score of 0.75 over the 8 classes.
After the random forest, we trained a decision tree classifier that achieved an F1 score of 0.78.
Our final choice was a deep learning model that obtained a F1 score of 0.91 on the test set.

The good results obtained suggest that such approaches based on deep neural networks are an excellent basis for solving the task. In particular, they are general enough to work in a real application context correctly. Since the result obtained can only be considered a starting point for further extensions, modifications, and improvements of the proposed approach, we have decided to make the product code available to the whole reference community with the hope that it will be useful for future work in the field. The code that implements the presented model can be found at the following GitHub repository: https://github.com/marcogdepinto/Emotion-Classification-Ravdess

ACKNOWLEDGEMENT

This work is funded by project ”DECiSION” codice raggruppamento: BQS5153, under the Apulian INNONETWORK programme, Italy and POR Puglia FESR FSE 2014-2020 - Sub-Azione 1.4.B – progetto “Feel at Home“ codice raggruppamento: UIKTJF3. Dominio di riferimento: Smart Cities & Communities.

REFERENCES

[1] DAVIS, S., AND MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing 28*, 4 (1980), 357–366.

[2] EKMAN, P. Basic emotions. *Handbook of cognition and emotion 98*, 45-60 (1999), 16.

[3] GOODFELLOW, I., BENGIO, Y., COURVILLE, A., AND BENGIO, Y. *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[4] HAYNES, J.-D., AND REES, G. Neuroimaging: decoding mental states from brain activity in humans. *Nature Reviews Neuroscience 7*, 7 (2006), 523.

[5] HUANG, X., ACERO, A., HON, H.-W., AND FOREWORD BY-REDDY, R. *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[6] IQBAL, A., AND BARUA, K. A real-time emotion recognition from speech using gradient boosting. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (2019), IEEE, pp. 1–5.

[7] JANNAT, R., TYNES, I., LIME, L. L., ADORNO, J., AND CANAVAN, S. Ubiquitous emotion recognition using audio and video data. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers* (2018), ACM, pp. 956–959.

[8] LECUN, Y., BENGIO, Y., ET AL. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks 3361*, 10 (1995), 1995.

[9] LIVINGSTONE, S. R., AND RUSSO, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one 13*, 5 (2018), e0196391.

[10] LOGAN, B., ET AL. Mel frequency cepstral coefficients for music modeling. In *ISMIR* (2000), vol. 270, pp. 1–11.

[11] MUDA, L., BEGAM, M., AND ELAMVAZUTHI, I. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083* (2010).

[12] NAIR, V., AND HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), pp. 807–814.

[13] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[14] PLATT, J. C., CRISTIANINI, N., AND SHAWE-TAYLOR, J. Large margin dags for multiclass classification. In *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K. Müller, Eds. MIT Press, 2000, pp. 547–553.

[15] POLIGNANO, M., DE GEMMIS, M., NARDUCCI, F., AND SEMERARO, G. Do you feel blue? detection of negative feeling from social media, 2017.

[16] ZHANG, B., ESSL, G., AND PROVOST, E. M. Recognizing emotion from singing and speaking using shared models. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)* (2015), IEEE, pp. 139–145.