Artificial Intelligence Research and Development
I. Sanz et al. (Eds.)

doi:10.3233/FAIA230675

Zero-Shot Prediction for Emotion Recognition Using Deep Spectrum Features

Elena ORTEGA BELTRÁN [a,1], Ismael BENITO-ALTAMIRANO [a,b] and Carles VENTURA [a]
[a] *Universitat Oberta de Catalunya*
[b] *Universitat de Barcelona, Departament d'Enginyeria Electrònica i Biomèdica*

Abstract. In this work we evaluate the zero-shot performance for emotion classification in audio recording datasets using a series of deep spectrum feature extractors which have been pre-trained on snore-analysus datasets.We evaluate here if those features are suitable for emotion-analysis on a zero-shot premise (no re-training of the CNNs). Once the feature extraction is complete, we evaluate training conventional classification models, i.e. an SVC, to classify a series of Spanish datasets which consist of male and female recordings expressing a set of emotions.

Keywords. Paralinguistics, Emotion recognition, Deep spectrum features

1. Introduction

One of the fields of Artificial Intelligence is the recognition of emotions from various sources. Among its applications we have, for example, social robots that are specifically designed to interact with humans in fields such as medicine[1,2]. In this work, we focus on the specific field of paralinguistics, the discipline that studies paralanguage, which is the set of phenomena that are not strictly verbal, of a vocal or gestural nature, that accompany the use of language and complete its meaning [3]. More specifically, this work focuses on the audio modality, where we propose to use a feature-based extractor, followed by a supported vector machine classifier (SVC), with pre-trained weights on another sound-related problem (snore analysis) [4], and we evaluate if zero-shot prediction is possible to solve our paralinguistic problem of emotion classification.

Results indicate that the proposed pipeline is suitable for emotion classification, achieving good results in zero-shot prediction. Moreover, we show how splitting the training of the SVC over gender –male/female– could improve our proposed solution.

2. Materials and methods

Our proposal is to use the so-called DeepSpectrum toolkit, a Python-based toolkit which extracts features from audio data using a pipeline that consists of converting audio data into a spectrum representation of this data, and then using pre-trained CNNs to extract

[1]Corresponding Author: Elena Ortega Beltrán, e-mail: eortegabeltran@uoc.edu.

features from those images [4]. Later, we propose using an SVM classifier, and train it to classify the feature representation of the audio data against the classes of the ELRA-S0329 dataset [5], which contains more than 6,000 recordings of Spanish recordings from one male and one female professional speakers. These recordings contain annotations of read text material for 11 classes: in neutral style with five different speech styles –loud, normal, slow, fast and soft– plus six expressions –anger, disgust, fear, joy, sadness and surprise–. The dataset is divided into train (80%) and test (20%) partitions.

We considered two experiments. On the one hand, we compared the results of training the SVC with features extracted with two pre-trained models: VGG16 and Xception. On the other hand, we divided the datasets by gender, and we trained the SVC model –using the VGG16 features– and we tested both trained models (male/female) against both subsets of the test dataset (male/female) as well as the merged one.

3. Results

The accuracy of the first test case (VGG16 architecture) is 82,4%, whereas the accuracy obtained when Xception is used is significantly lower (75%). For both implementations, results indicate that our proposed model (DeepSpectrum + SVC) presented a good response distinguishing the six emotion classes of the dataset (Anger, Disgust, Fear, Joy, Sadness and Surprise); however, it performs poorly dividing the neutral emotion tones (Fast, Slow, Loud, Soft and Normal). Futures works might include a separate training for the subset of neutral classes.

In the second experiment, we train the system both in the separated gender datasets and the complete one. We compare the different performance of both classifiers in the female and male datasets. We observe that the classifier trained in a specific gender subset achieves a better accuracy (83.88% and 83.19% for female and male respectively) than the one trained in the complete dataset (78.45% and 77.37% for female and male respectively).

Acknowledgements

This work is part of the project PLEC2021-007868, funded by MCIN and by the EU, and the project PID2022-138721NB-I00.

References

[1] Makiuchi MR, Uto K, Shinoda K. Multimodal emotion recognition with high-level speech and text features. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE; 2021. p. 350-7.

[2] Papakostas GA, Sidiropoulos GK, Papadopoulou CI, Vrochidou E, Kaburlasos VG, Papadopoulou MT, et al. Social robots in special education: a systematic review. Electronics. 2021;10(12):1398.

[3] Poyatos F. The multichannel reality of discourse: language-paralanguage-kinesics and the totality of communicative systems. Language Sciences. 1984;6(2):307-37.

[4] Amiriparian S, Gerczuk M, Ottl S, Cummins N, Freitag M, Pugachevskiy S, et al. Snore sound classification using image-based deep spectrum features. 2017.

[5] ELRA, Emotional speech synthesis database,;. http://catalog.elra.info/en-us/repository/browse/ELRA-S0329/.