Towards a classifier to recognize emotions using voice to improve recommendations

José Manuel Fuentes, Joaquin Taverner, J.A. Rincon, and Vicente Botti

Valencian Research Institute for Artificial Intelligence (VRAIN)
Universitat Politècnica de València, Valencia, Spain
{jofuelo1, joataap, jrincon, vbotti}@dsic.upv.es

Abstract. The recognition of emotions in tone voice is currently a tool with a high potential when it comes to making recommendations, since it allows to personalize recommendations using the mood of the users as information. However, recognizing emotions using tone of voice is a complex task since it is necessary to pre-process the signal and subsequently recognize the emotion. Most of the current proposals use recurrent networks based on sequences with a temporal relationship. The disadvantage of these networks is that they have a high runtime, which makes it difficult to use in real-time applications. On the other hand, when defining this type of classifier, culture and language must be taken into account, since the tone of voice for the same emotion can vary depending on these cultural factors. In this work we propose a culturally adapted model for recognizing emotions from the voice tone using convolutional neural networks. This type of network has a relatively short execution time allowing its use in real time applications. The results we have obtained improve the current state of the art, reaching 93.6% success over the validation set.

Keywords: Emotion recognition, voice analysis, recommendation system.

1 Introduction

Emotions play an important role in our social interactions. Our ability to recognize emotions is based on the ability to recognize different verbal and non-verbal communication acts such as gestures or facial expression. The voice tone is one of the non-verbal communications that allows to identify the emotion. When we refer to the recognition of the voice tone we are referring to the set of prosodic characteristics such as the tone, energy, or speech speed, but not the spoken message itself. Depending on the voice tone in which a person speaks different emotions can be expressed. For example, a low voice tone with low frequencies may be related to sadness, while a tone with ups and downs may be identified as joy. The underlying problem that must be faced when analyzing emotions in the voice tone is the dependence of the voice tone on factors such as culture or language [6]. Different cultures may interpret the voice tone differently [4].

In addition, the musicality of the voice tone varies depending on the language. Therefore, when designing models capable of recognising emotion from the voice tone, these cultural and idiomatic factors must be taken into account.

In the area affective computing [8] different models have been proposed to recognize emotions from variations in the voice tone. Currently most of these models are based on the use of Neural Networks (NN). More specifically in the use of Recurrent Neural Networks (RNN). These networks allow the voice tone to be analysed sequentially. However, these networks have a high computing time for classification which makes them not useful in real time applications. In this work we propose a cultural adapted classifier to recognize emotions in real time from the voice tone for Spanish speakers which in future works can be used as input for a recommendation system. Our model uses a Convolutional Neural Network to analyze the characteristics of the voice tone. This type of NN obtain's good results in less time than the RNN. In addition, thanks to the pre-processing of the signal, our model is tolerant to failures.

2 Related work

Intelligent user-based recommendation systems are commonly based on extracting information about the user to perform a customized recommendation. In recent years, an increasing number of models consider different affective characteristics, such as emotions or mood, to improve the personalization of recommendations. For example, in [15] a model for recommending recipes according to the user's mood is presented. This model is based on tests to consult the user's mood and recommend a recipe that fits that mood. The problem with this model is that the user must enter his mood manually. In contrast, models that are able to recognize emotions or mood automatically are less invasive to the user. For example the model proposed in [9] allows to recommend songs using the emotions expressed by the user. For this, the system captures an image of the face and extracting the emotional state through neural networks. One of the main challenges when generating this type of system is to generate models that allow for real-time emotion recognition. One of the most common ways to capture user emotion in present literature is voice tone analysis. The recognition of emotions through the tone of voice is a difficult challenge. This is due in part to both the cultural and linguistic dependence of the voice tone and to the pre-processing of the audio signal for subsequent analysis. At present there are different techniques that allow to extract characteristics of the voice tone such as the duration or the energy of the signal. *Mel Frequency Cepstral Coefficients* (MFCC) [11] are the most commonly used features in automatic audio recognition. This technique assumes that in short periods of time the characteristics of the audio remain relatively stable. Therefore, by a preprocess, which subdivides the audio signal into smaller fragments, relatively static characteristics can be extracted for each fragment of the audio signal. On the other hand, *Spectral Rolloff* technique [1] allows to obtain characteristics about the relationship between energy and frequency. Its use in conjunction with other features, such as

MFCCs, has been shown to improve the overall performance of speech recognition systems [5]. The *Log Filterbank Energies* spectrogram model is also widely used to extract characteristics from the voice tone. This spectrogram is obtained by applying filter banks to the signal periodgram. They are an intermediate step in obtaining MFCCs. Therefore, they have a higher correlation than MFCCs, but in some cases they retain a greater amount of information from the original signal [14].

The characteristics obtained through the different pre-processes are then used to perform the recognition of the audio signal using different machine learning techniques. Within the area of recognition of emotions in the tone of voice, different classification techniques have been used. The first approximations made in the area of the recognition of emotions in the tone of the voice were based on the use of Hidden Markov Models (HMM) or Support Vector Machines (SVM) [12]. At present, with the rise of neural networks, most proposals use models based on Recurrent Neural Networks (RNN). This type of networks allow developers to work with data based on time sequences. Therefore, they are suitable for continuous audio analysis. One of the most used RNNs in this field are the *Long Short Term Memory* (LSTM) [2]. LSTM allow to solve the problem of gradient vanishing or exploding. This type of neural network has proven effective in improving the accuracy of recognition in different tasks. For example, in [3] the authors compare two classical models, *Multivariate Linear Regression* (MLR) and SVM, with a modern LSTM obtaining the best results with the use of the LSTM. However, RNNs have the disadvantage that their processing time is high. Therefore, RNNs are not highly recommended for applications designed to be used in real time. This is why, in recent years, there have been models that propose to use Convolutional Neural Networks (CNN). CNNs are a type of neural networks specialized in image analysis. CNN has also proven to be very effective in the task of recognition patterns in audio. To do this, the network is fed with images of the spectral frequencies of the audio. This type of technique has proven to be effective in recognizing emotions in tone of voice with less compute time than RNN. For example, in [16], a CNN is compared with a classical SVM. The authors tested both models on two types of tasks: image classification and audio emotion recognition. In both tasks the CNN improves the accuracy of the SVM. Specially in the audio task, where the authors get an accuracy of 97.6% with the CNN, which is much higher than the 46.6% obtained with the SVM. On the other hand, in this type of classification task we have to take into account factors such as culture and language. Over the years different authors have theorized about the dementia between emotions, culture and language. Constructivist psychology holds that emotions depend on these cultural and idiomatic factors and that it is therefore impossible to correctly identify an emotion without taking them into account [10]. In the case of emotion analysis based on the tone of voice, cultural and idiomatic factors take on greater importance since substantial differences can appear in the variation of acoustic frequency when expressing emotions [7]. Therefore, this type of recognition systems must be adapted through the use of corpus specialized in the culture and language in which they are going to be

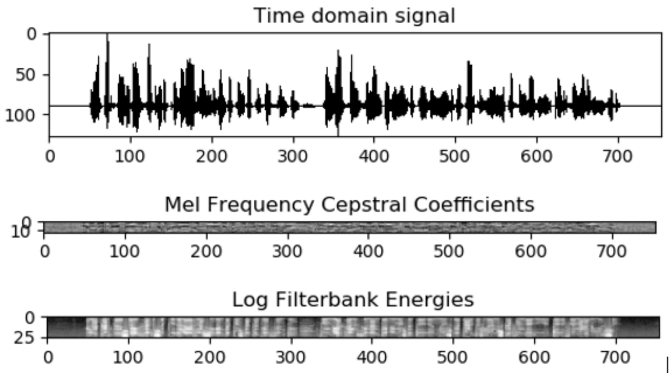

Fig. 1: MFCC and LFE extracted from the signal represented in the time domain

used. Using general models or models trained in other languages can lead to errors in predicting emotion [13]. At present, however, most of the proposals in this area are focused on emotion analysis for English speakers.

3 Proposal

The proposal presented in this paper would allow the use of the voice to recognize emotional states. The cues would serve as an input to a recommendation system. however, due to the complexity of performing emotion classification using the voice, in this proposal we focus on how to classify emotions. because of this complexity and the high dependence that the tone of voice has on language and culture discussed above, we propose a model of emotion recognition in the tone of voice adapted to Spanish speakers. Our proposal is based on CNNs, since, as mentioned above, there are precedents of improving the state of the art by applying the advantages of CNNs to the voice recognition task. However, as CNNs are specialized in the processing of data matrices, it is necessary to pre-process the audio signal. In this pre-process we have extracted two main characteristics of the audio: the *Mel Frequency Cepstral Coefficients* (MFCC) and the *Log Mel-Filterbank Energies* (LFE). Then we fed this two characteristics to the network as individual images. We considered also to use the *Time-domain Signal Representation* and the *Spectral Subband Centroids*, but adding them offered worse results.

An example of the features extracted from an acoustic signal along with the corresponding time domain signal is shown in Figure 1. On the other hand, when working with audio signals we must consider the time factor since the audio files can have different duration. Therefore, it is not possible to apply padding techniques in a direct way to these type of samples it is necessary to regularize the size. To do that, we have made an split of the samples into fragments of a fixed size (13x200 pixels for the MFCC and 26x200 pixels for the LFE). We have applied padding to the last fragment of each audio to adjust it to the same size.

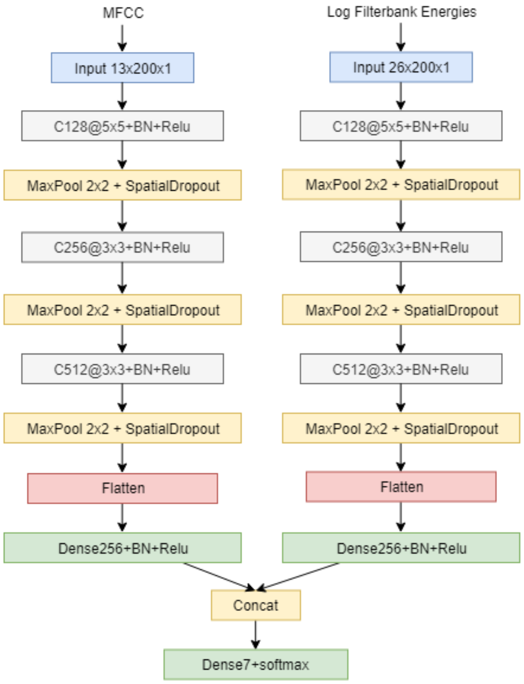

Fig. 2: Proposed convolutional neural network for the audio problem

We design a non-sequential CNN with two parallel convolutional branches: one to analyze MFCC audio characteristic and the other to analyze the LFE audio characteristic. We have used the same structure for both branches:

- 1 convolutional layer of 128 5x5 sized filters.
- A max pooling layer with 2x2 filters and stride 2 followed by a spatial dropout layer.
- 1 convolutional layer of 256 5x5 sized filters.
- A max pooling layer with 2x2 filters and stride 2 followed by a spatial dropout layer.
- 1 convolutional layer of 512 5x5 sized filters.
- A max pooling layer with 2x2 filters and stride 2 followed by a spatial dropout layer.
- A flatten layer
- 1 fully connected layer of 256 neurons.

Then, the results of both fully connected layers are concatenated to a fully connected output layer of 7 neurons one for each emotion including the neutral state. The resulting schema of the CNN is shown in Figure 2.

Table 1: Number of samples per class in the voice task

Class	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral
Number of samples	725	732	735	732	728	730	1658

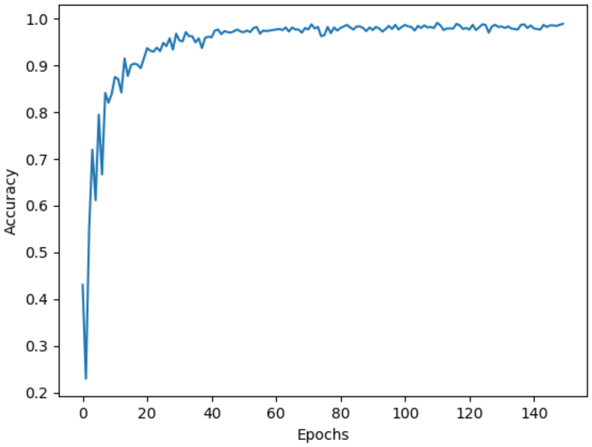

Fig. 3: Accuracy of the CNN throughout iterations for the validation set.

3.1 The voice data set

As mentioned above, as the voice tone is cultural and language dependant, we had to use a corpus adapted to the language in which the model is going to be used that in this case is Spanish. However, despite there are a large number of data sets for voice classification in English, to our knowledge the only data set made for this proposal in Spanish is the *INTER1SP Spanish Emotional Database* [1]. This data set contains 6040 samples of audio pronounced, in a silent studio, by two actors: a man and a woman. Table 1 summarizes the distribution of the samples for the six emotion and the neutral state.

We decided to divide the data set into three parts: training set, validation set, and test set. The percentages assigned follow the classic distribution of 70% for training, 15% for validation, and 15% for test.

3.2 Results

The accuracy obtained by this model is 93.6% for the validation test (see Table 2). The current state of the art for this corpus is at a precision of 90.05%, which corresponds to the work presented in [3]. In that proposal the authors combined melfrequency cepstrum coefficients with modulation spectral features.

[1] http://catalog.elra.info/en-us/repository/browse/ELRA-S0329/

Table 2: Confusion matrix for the validation test corresponding to the 15% of the samples.

%	Joy	Surprise	Fear	Anger	Disgust	Sadness	Neutral	Rate
Joy	**97.63**	3.19	0.00	2.00	1.31	0.00	0.87	97.30
Surprise	0.68	**89.36**	1.11	1.00	1.31	0.00	0.00	93.33
Fear	0.00	1.06	**94.44**	0.00	1.31	5.08	0.87	92.39
Anger	0.34	2.13	3.33	**94.00**	2.61	5.08	1.74	86.24
Disgust	0.68	1.06	1.11	3.00	**90.85**	1.69	0.00	94.56
Sadness	0.34	1.06	0.00	0.00	0.65	**81.36**	0.87	92.31
Neutral	0.34	2.13	0.00	0.00	1.96	6.78	**95.65**	91.67
Precision	97.63	89.36	94.44	94.00	90.85	81.36	95.65	

Comparing our results with the results obtained for the same corpus in the current state of the art we can observe that our model improves the accuracy of emotion recognition for this corpus. This success rate is largely due to the absence of environmental noise in the data set. Therefore, the success rate could be reduced in noisy environments. On the other hand, the pre-process carried out to obtain the characteristics divided each audio fragment into different partitions. This partitioning allows the system to make some mistakes as long as the majority of the fragments that compose the audio are correctly classified.

4 Conclusion and future work

In this work we have proposed a model of recognition of emotion in the tone of voice adapted to personal factors such as culture and language which will be used as input for a recommendation system. Our model is weighted to analyze the emotion in the tone of voice in real time. For this we have designed a model based on the use of convolutional neural networks. This type of network makes it possible to recognize emotion in the tone of the voice in less time than recurrent neural networks. In addition, our model is adapted to the culture and language in which it will be used through the use of a Spanish data set. The results we have obtained improve the current state of art on this data set. In addition, the way in which we have pre-processed the signal, dividing each audio fragment into different partitions, allows the system to have a certain tolerance to errors. This is because the resulting emotion is predominant for the set of fragments.

As future work we want to increase the set of samples incorporating another data set. We also want to create another model in a different language to compare cross-cultural variations and to be able to analyze the success rate generated by both classifiers when used in a language different than the language of the data set. Finally, we want to incorporate this model into a recommendation system in order to adapt it to the user's emotion.

Acknowledgements

This work is partially supported by the Spanish Government project TIN2017-89156-R, GVA-CEICE project PROMETEO/2018/002, Generalitat Valenciana and European Social Fund FPI grant ACIF/2017/085, Universitat Politecnica de Valencia research grant (PAID-10-19), and by the Spanish Government (RTI2018-095390-B-C31).

References

1. A. Balakrishnan, Anusha; Rege. Reading emotions from speech using deep neural networks. Technical report, Stanford University, Computer Science Department, 2017.
2. S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.
3. L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. Mahjoub. Speech emotion recognition: Methods and cases study. pages 175–182, 01 2018.
4. K. W. McCluskey, D. C. Albas, R. R. Niemi, C. Cuevas, and C. Ferrer. Cross-cultural differences in the perception of the emotional content of speech: A study of the development of sensitivity in canadian and mexican children. *Developmental Psychology*, 11(5):551, 1975.
5. K. K. Paliwal. Spectral subband centroid features for speech recognition. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, volume 2, pages 617–620. IEEE, 1998.
6. S. Paulmann and A. K. Uskul. Cross-cultural emotional prosody recognition: Evidence from chinese and british listeners. *Cognition & emotion*, 28(2):230–244, 2014.
7. E. Pépiot. Voice, speech and gender:. male-female acoustic differences and cross-language variation in english and french speakers. *Corela. Cognition, représentation, langage*, (HS-16), 2015.
8. R. W. Picard et al. Affective computing. *Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology*, 1995.
9. J. Rincon, F. de la Prieta, D. Zanardini, V. Julian, and C. Carrascosa. Influencing over people with a social emotional model. *Neurocomputing*, 231:47–54, 2017.
10. J. A. Russell, M. Lewicka, and T. Niit. A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, 57(5):848, 1989.
11. B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. volume 2, pages 401–404, 08 2003.
12. B. Schuller, R. Villar, G. Rigoll, and M. Lang. Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. volume 1, pages 325–328, 02 2005.
13. W. Thompson and L.-L. Balkwill. Decoding speech prosody in five languages. *Semiotica*, 2006, 01 2006.
14. V. Tyagi and C. Wellekens. On desensitizing the mel-cepstrum to spurious spectral components for robust speech recognition. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I–529. IEEE, 2005.

15. M. Ueda, Y. Morishita, T. Nakamura, N. Takata, and S. Nakajima. A recipe recommendation system that considers user's mood. In *Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services*, pages 472–476. ACM, 2016.
16. B. Zhang, C. Quan, and F. Ren. Study on cnn in the recognition of emotion in audio and images. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5, June 2016.