Speech emotion representation: A method to convert discrete to dimensional emotional models for emotional inference multimodal frameworks

Fernando Elkfury[1][0000-0003-2131-604X], Jorge Ierache[1,2] [0000-0002-1772-9186]

[1] Instituto de Sistemas Inteligentes y Enseñanza Experimental de la Robótica, Universidad de Morón (1708) Morón Argentina.
[2] Laboratorio de Sistemas Información Avanzados Universidad de Buenos Aires (C1063) Ciudad Autónoma de Buenos Aires, Argentina.
{felkfury,jierache}@unimoron.edu.ar

Abstract. Computer-Human interaction is more frequent now than ever before, thus the main goal of this research area is to improve communication with computers, so it becomes as natural as possible. A key aspect to achieve such interaction is the affective component often missing from last decade developments. To improve computer human interaction in this paper we present a method to convert discrete or categorical data from a CNN emotion classifier trained with Mel scale spectrograms to a two-dimensional model, pursuing integration of the human voice as a feature for emotional inference multimodal frameworks. Lastly, we discuss preliminary results obtained from presenting audiovisual stimuli to different subject and comparing dimensional arousal-valence results and it's SAM surveys

Keywords: Emotions, Multimodal Framework, Affective computing.

1 Introduction

Even though speech is the most traditional way of human communication, as a feature for emotion recognition it is not as expressive as one may think. According to Albert Mehrabian [1] voice tone can only transmit a 38% of the emotions a person might feel at a given time. Despite being a feature with a low percentage of expressiveness, in a multimodal environment of emotion analysis it is meaningful for correctly inferring the emotional state of a person by comparing and correcting data from other sensors or methods of emotion assessment. Human-computer interaction is now more and more frequent due to accelerated technological development, although, they are often lacking an affective component. Thus, one of the main goals of the recent computer-human communication development is to improve user experience through making interaction between computers and humans as natural as it is between persons [2]. Current works in this field, such as [3] [4] [5] [6], infer emotion in a categorical manner, usually partially matching Ekman's model [7]. This work, additionally to the categorical approach, provides a preliminary architecture to obtain valence and arousal from a voice sample in the context of a multimodal emotional dimensional approach for emotion elicitation and representation, based on the use of a CNN [8] classifier for determining valence,

and a method to calculate arousal based on the measurement of the voice source dB. The CNN classifier from recent associated projects [9][10] is capable of up to 92% accuracy for a Spanish dataset. In the following section we present and develop our classifiers and the proposed conversion method. In the third section we discuss our preliminary tests. And lastly in the fourth section we discuss partial results and conclusions.

2 Solution development

To improve and facilitate integration of the human voice as a component in an emotional inference multimodal framework we develop a convolutional neuronal network (CNN) classifier that is trained with Mel scale [11] spectrograms from the audio samples in the ELRA [12] emotional data set. We work with the seven emotional labels proposed by Ekman, joy, fear, sadness, anger, disgust, surprise plus a neutral emotion considered by most data sets and tools available. To represent emotions dimensionally we use a Russell's circumflex of emotion [13] updated in the work of Sherer [14] so we can keep working with eight emotions. (See Fig. 1.)

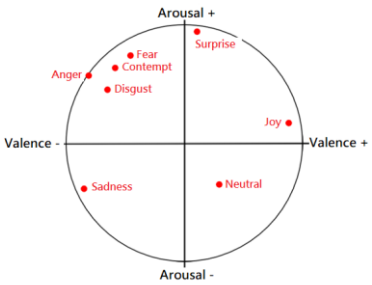

Fig. 1. Circumflex based on Russel and Sherer's work.

To convert emotional models, we start with the subtraction of the probability obtained for the "joy" label of the classifier and the probability of the most negative emotion as is proposed by Leanne in [15]. For example, for a given prediction of the classifier where "joy" equals to 0.8 anger 0.2 "fear" 0.1 and "saddens" 0.3, valence value would be in this case 0.5 from the subtraction 0.8 - 0.3. According to Leanne the "surprise" emotion is not taken in consideration for the calculation, so we add up to that statement saying that "neutral" also should not be considered. We now must find an associable feature to arousal values. We propose using the difference between the mean dB values of subsequent samples taken during a subject's testing session and the mean dB values of the sample tested. In formula 1 we see a brief description of what is proposed, "x" is the average dB value of the previous samples, "y" is the dB value of the current sample from which we want to obtain the value of arousal and "n" is the number of samples taken in the session so far including the current one. Then the difference between the current sample and the previous average is calculated and rescaled to place it where it corresponds in the circumflex.

$$E = \frac{(\sum_{i=1}^{n} x)}{n} - y \quad (1)$$

The figure 3 below demonstrates the proposed architecture's workflow to obtain Arousal/valence values from a speech voice sample.

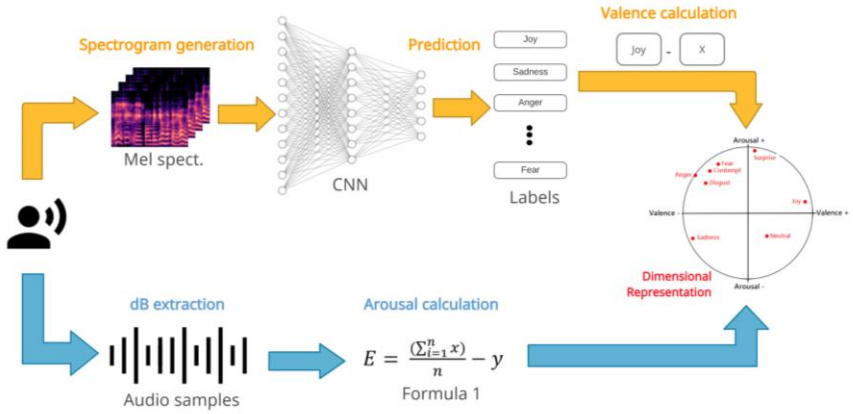

Fig. 2. Example of our architecture's workflow.

3 Test and Results

To evaluate the representation quality of the proposed transformation method, preliminary tests were carried out. In Fig. 4 below, we compare Arousal/Valence values from our CNN classifier and the proposed transformation method (blue mark) with SAM surveys [16] classifications from various subjects (black marks) for a given audio stimuli. Achieving quadrant matching between results.

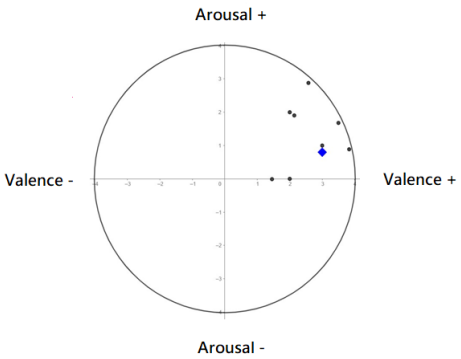

Fig. 3. Example of a test result

4 Conclusions

First of all, dB are a relative measurement unit, in this case they are used for the simple reason of showing the differences in the sample's time series amplitude in a more intuitive and easy-to-work way. The scale may require modifications, which could be determined under empirical tests. Summarizing this arousal values provide us with the visualization of the relationship that might exist between the voice volume and the changes from one emotional state to another. Also, this method relies on the history of a series of samples taken, so the measurement becomes more reliable once a definite trend of the average is established.

References

1. Mehrabian, A.: Communication Without Words. Communication theory, 193-200 (2017).
2. Planet, S.: Reconocimiento afectivo automático mediante el análisis de parámetros acústicos y lingüísticos del habla espontánea. (2013).
3. Sánchez-Gutiérrez, M.E., Albornoz, E.M., Martinez-Licona, F., Rufiner, H.L., Goddard, J.: Deep Learning for Emotional Speech Recognition. Lecture Notes in Computer Science. 311–320 (2014).
4. Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub: Automatic Speech Emotion Recognition Using Machine Learning. Social Media and Machine Learning. (2020).
5. Mustaqeem, Kwon, S.: A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. Sensors (Basel). 20, 183 (2019).
6. Badshah, A.M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M.Y., Kwon, S., Baik, S.W.: Deep features-based speech emotion recognition for smart affective services. Multimedia Tools and Applications. 78, 5571–5589 (2017).
7. Ekman, P.: Basic Emotions. In Handbook of Cognition and Emotion, pp. 45–60. John Wiley & Sons, Ltd (2005)
8. K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, 193–202 (1980).
9. Elkfury, F., Ierache, J.: Reconocimiento de emociones en la voz empleando redes neuronales y su integración en frameworks multimodales de educción emocional. XXIII Workshop de Investigadores en Ciencias de la Computación (WICC 2021). Chilecito, 2021. In press.
10. Elkfury, F., Ierache, J.: Clasificación y representación de emociones en el discurso hablado en español empleando Deep Learning. RISTI - Revista Ibérica de Sistemas e Tecnologías de Información, versión impresa ISSN 1646-9895 n°42. In press.
11. Volkmann J., Stevens S. S., Newman E. B.: A Scale for the Measurement of the Psychological Magnitude Pitch. The Journal of the Acoustical Society of America, 185-190 (1937).
12. ELRA catalog page, http://catalog.elra.info/en-us/repository/browse/ELRA-S0329/, last accessed 2021/4/8.
13. Russell, J. A.: A circumplex model of affect. Journal of Personality and Social Psychology, vol. 39, 1161–1178 (1980).
14. Scherer, K. R.: What are emotions? And how can they be measured? Social Science Information, vol. 44, 695–729, (2005).
15. Loijens L., Krips O.: FaceReader Methodology Note. https://www.noldus.com/facereader/resources, last accessed 2021/4/8.
16. Lang, P. J.: The cognitive psychophysiology of emotion: Fear and anxiety. In A. H. Tuma & J. D. Maser (Eds.), Anxiety and the anxiety disorders, 131–170 (1985).