risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 19/03/2021
Aceitação/Acceptance: 28/05/2021

Clasificación y representación de emociones en el discurso hablado en español empleando Deep Learning

Fernando Elkfury[1], Jorge Ierache[1,2]

{felkfury; jierache}@unimoron.edu.ar

[1]Instituto de Sistemas Inteligentes y Enseñanza Experimental de la Robótica (ISIER), Secretaria de Ciencia y Tecnología, Escuela Superior de Ingeniería, Informática, Ciencias Agroalimentarias Universidad de Morón (1708) Morón Argentina.

[2]Laboratorio de Sistemas Información Avanzados (LSIA) Facultad de Ingeniería Departamento Computación Universidad de Buenos Aires (C1063) Ciudad Autónoma de Buenos Aires, Argentina.

DOI: 10.17013/risti.42.78–92

Resumen: Inferir emociones a partir de la voz de las personas implica muchos problemas que necesitan ser estudiados cuidadosamente, tales como: qué emociones podemos identificar realmente, definir concretamente qué se entiende por cada emoción descripta, cuáles son las mejores características para la identificación y qué clasificadores dan el mejor rendimiento. En este trabajo se comparan dos modelos de redes neuronales para la clasificación de emociones en el discurso hablado (voz) y se propone un método para la transformación de enfoque categórico de clasificación de emociones a uno dimensional para la integración del clasificador con frameworks multimodales de captura de emociones.

Palabras-clave: Reconocimiento de emociones en la voz, Aprendizaje automático, redes neuronales, frameworks multimodales.

Emotion classification and representation of emotion in spanish spoken speech using Deep Learning

Abstract: Inferring emotions from people's voices involves many problems that need to be studied, such as: what emotions can we really define, specifically define what is meant by each described emotion, what are the best features to extract, and which classifiers perform the best. In this work we compare two neural networks models for the classification of emotions in spoken speech (voice) and a method is proposed for the transformation of the categorical approach of emotion classification to a dimensional one for the integration of the develop classifier with an emotional inference multimodal framework.

Keywords: Deep Learning, emotion recognition, framework, neuronal network.

1. Introducción

En el marco de las tecnologías de la información se ve un crecimiento en el uso de la componente afectiva en la interacción humano-maquina, en este orden la computación afectiva presentó un rápido crecimiento en distintos campos como educación (López et al., 2016), turismos (Chanchí et al., 2020), redes sociales, con el fin de analizar las emociones y/o sentimientos que expresan los usuarios a partir de sus publicaciones (Chanchí & Córdoba, 2019), entre otros; sin embargo, pocos desarrollos adoptan la voz como una herramienta para la inferencia emocional. Nuestro enfoque propone la integración de clasificadores emocionales basados en técnicas de Deep Learning para la contrastación y validación de datos en un framework de educción emocional multimodal. El análisis de las emociones en la voz humana es una tarea poco trivial, incluso para el propio ser humano. Si bien el habla es la forma tradicional de comunicación, no es una característica sensible a los cambios emocionales y por lo tanto la educción emocional a partir de la misma, cuando no se posee contexto semántico ni de otra clase, resulta parcial. Según Albert Meherabin, el tono de la voz expresa solo un 38% de las emociones que puedan transmitir las personas en un momento dado (Mehrabian, 2017) Deep Learning ha sido considerado como un campo de investigación emergente en el aprendizaje automático y ha ganado más atención en los últimos años. Las técnicas de aprendizaje profundo para los sistemas de reconocimiento de emociones tienen varias ventajas sobre los métodos tradicionales, incluida su capacidad para detectar la estructura compleja y las características sin la necesidad de extracción y ajuste manual de estas. Lo cual es un aspecto clave en el desarrollo de estos dado que la precisión de los clasificadores suele estar ligada a la selección de las características de la voz que se usaran para el entrenamiento.

Entre los recientes trabajos de investigación en el área, se destacan: a) “Deep Learning for Emotional Speech Recognition”, desarrollado por Sanchez-Gutierrez, et al. donde se comparan resultados del entrenamiento de modelos usando Restricted Boltzmann Machines (RBM) (Sánchez-Gutiérrez et al., 2014) y Deep Belief Networks (DBN) (Patterson & Gibson, 2017). Minimizaron el número de variables involucradas eligiendo las emociones de alegría, tristeza, ira, miedo, asco y sorpresa junto con neutral. La base de datos fue creada por el Centro de Tecnologías y Aplicaciones del Lenguaje y del Habla (TALP) de la Universidad Politécnica de Cataluña (UPC) con el propósito de investigar el discurso emocional. La base de datos era parte de un proyecto más grande, INTERFACE, que involucraba cuatro idiomas, inglés, francés, esloveno y español. (ELRA, 2011); b) “Recognition of Emotions in Mexican Spanish Speech: An Approach Based on Acoustic Modelling of Emotion-Specific Vowels” (Caballero-Morales, 2013) propone la construcción de un corpus en español mexicano y el desarrollo de un clasificador emocional del discurso a partir del modelado de vocales asociadas con cada estado emocional; c) Leila Kerkeni, et al. en su artículo “Automatic Speech Emotion Recognition Using Machine Learning” presenta una comparativa entre las técnicas MLR, SVM y RNN haciendo uso set Inters1p (ELRA, 2011) con extracción de MFCC y MS (Davis & Mermelstein, 1980) evaluando las combinaciones de estas características, además de combinarlo con técnicas de selección de características, LR-RFE. (Kerkeni et al., 2020); d) Mustaqeem y Soonil Kwon en su artículo proponen un framework basado en CNN que utiliza espectrogramas para educir las emociones a partir del habla de un sujeto. (Mustaqeem & Kwon, 2019); e) Abdul Malik Badshah1 et al. emplea el uso de espectrogramas (similar al autor anterior) obtenidos por medio de la transformada de Fourier para alimentar un modelo CNN aplicando una arquitectura con kernels rectangulares. (Badshah et al., 2017)

En este trabajo se optó por el uso de espectrogramas, en los cuales las frecuencias fueron convertidas a escala de Mel (Volkmann et al., 1937), y se evalúa el desempeño de redes neuronales convolucionales (CNN) (Fukushima, 1980) (Shafkat, 2018) y redes neuronales recurrentes (RNN) (Josh Patterson & Adam Gibson, 2017) en la construcción de un clasificador de emociones.

En la sección dos a continuación se resumen métodos de representación de emociones comúnmente utilizados. En la sección tres se describe el problema y se proponen una solución. La sección cuatro expone un conjunto de pruebas basales para la evaluación de las soluciones propuestas. La sección cinco plantea una discusión al respecto enfoques y tecnologías utilizadas por diversas investigaciones recientes. Por último, en la sección seis se discuten resultados y próximos pasos.

2. Representación de las emociones

Las emociones se pueden representar y definir de diversas formas, determinarlas y establecer una taxonomía de estas no es una tarea simple. En 1994 el psicólogo Paul Ekman propuso un conjunto de 6 emociones básicas que no están influenciadas por la cultura de las personas. Las emociones propuestas inicialmente fueron alegría (joy), miedo (fear), tristeza (sadness), ira (anger), disgusto (disgust) y sorpresa (surprise). Se las considera básicas por estar ligadas a la supervivencia de los individuos y en base a patrones evolutivos (Garcia, 2013).

Además de estas seis emociones pueden existir muchas otras secundarias y derivadas de las básicas, aunque originadas por la influencia cultural. En publicaciones posteriores Ekman añadió una séptima emoción básica llamada desprecio (contempt). También se suele considerar una octava emoción denominada neutral. La mayor parte de las teorías emocionales coinciden que las emociones básicas son menores a 10 (Garcia, 2013).

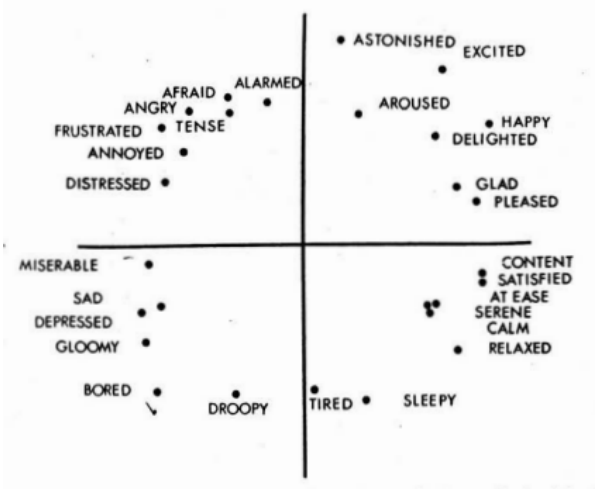

Figura 1 – Circunflejo de Russel

Dada la dificultad de definir y trabajar con extensas listas de emociones, algunos autores platearon trabajar en enfoques continuos o dimensionales. En el enfoque dimensional uno de los modelos más aceptados es el que presento el psicólogo James Russel (Russell, 1980). En el cual plantea un eje bidimensional que representa la valencia ("Valence") en el eje x, y la excitación ("Arousal") en el eje y donde las etiquetas emocionales se asocian en su propuesta conocida como el circunflejo de Russel. Siendo la valencia el grado de placer o de disgusto de la emoción manifestada y la excitación representa efectivamente el grado de relajación o excitación del sujeto. En la figura 1 se muestra el modelo de circunflejo propuesto por Russel.

Considerando varios estudios, el número de dimensiones del espacio emocional para la representación de emociones suele limitarse a dos o tres. (Garcia, 2013). La tercera dimensión, menos frecuente, se denomina control o dominancia y define el grado de control que tiene la persona sobre la emoción manifestada. Si bien existen otros mecanismos dimensionales como la Rueda de Ginebra (GEW, 2005), estos no se aplicaron en esta etapa en razón que contiene una mayor cantidad de etiquetas de distinción emocional y se dificulta su integración con las emociones resultantes del enfoque categórico en particular en voz.

3. Relevancia del problema y soluciones propuestas

La comunicación entre las personas y las máquinas o sistemas es cada vez más frecuente por los avances tecnológicos, sin embargo, los desarrollos son en su mayoría carentes de la componente afectiva. Por tanto, uno de los objetivos recientes de la comunicación persona-máquina es la mejora de la experiencia de usuario, intentando conseguir que esta comunicación sea lo más parecida a la interacción entre personas. (Garcia, 2013). Dicho esto, se plantean los siguientes problemas para los cuales se propuso y desarrollaron soluciones En primer lugar, la falta de una arquitectura de reconocimiento de emociones en el discurso que reconozca emociones en el discurso hablado en español. Seguido de la ausencia de un API moderna de uso libre que pueda ser utilizada para la investigación y desarrollo de software. Y, por último, la falta de un método de conversión de enfoques categóricos a dimensionales, que nos permita Integrar la emoción determinada por la fuente de voz en un framework multimodal de análisis emocional. Con intención de dar respuesta a los problemas planteados se desarrolló un clasificador de emociones que pueda reconocer emociones en el discurso, además, estableciendo una comparativa de performance entre clasificadores basados en RRN y CNN, para luego fijar la integración del clasificador con un framework multimodal de captura de emociones, en particular con el aporte del desarrollo de una API que facilite la explotación del modelo diseñado; proponiendo e implementando un método para pasar del enfoque categórico a un enfoque dimensional.

Para el desarrollo propuesto se aplicaron dos sets de datos que están etiquetados con emociones básicas y parcialmente coincidentes con el enfoque categórico de Ekman. El primero de ellos, "Interes1p" (ELRA, 2011) contiene las grabaciones de un hombre y una mujer hispanohablantes profesionales españoles (español de Castilla) grabados en una sala con ruido reducido. Consiste en grabaciones y anotaciones de material de texto leído en estilo neutral más seis emociones; tristeza, enojo, temor, asco y sorpresa; todo en estilos de voz rápidos, lentos, suaves y fuertes. El material de texto se compone de 184 elementos que incluyen oraciones fonéticamente equilibradas, dígitos y palabras

aisladas. El material fue el mismo para todos los modos y estilos, dando un total de 3h 59min de voz grabada para el hablante masculino y 3h 53min para el hablante femenino.

El segundo set de datos “EmoFilm” (Parada-Cabaleiro, Costantini, Batliner, Baird, & Schuller , 2018), es un corpus que compone de datos de habla emocional multilingüe que comprende 1115 instancias de audio producidas en inglés, italiano y español. Los clips de audio (con una longitud media de 3,5 segundos y 1,2 segundos) se extrajeron en formato de onda (sin comprimir, mono, frecuencia de muestreo de 48 kHz y 16 bits) de 43 películas (original en inglés y su sobre doblado en italiano y versiones en español). Se consideraron géneros como comedia, drama, horror y suspenso. Se extraen ira, desprecio, felicidad, miedo y tristeza. (Parada-Cabaleiro, Costantini, Batliner, Baird, & Schuller , 2018). Para poder trabajar con las siete emociones que propone Ekman más la emoción neutral articulada en la mayoría de las herramientas y sets de datos, se añade al set de datos Inters1p el corpus EmoFilm.

3.1. Clasificadores

Se optó por el uso de espectrogramas, como los que se ven en la figura 2, como dato para alimentar las redes neuronales. La hipótesis es que las capas convolucionales sean suficiente para encontrar patrones adecuados para la generalización en el espectrograma. Para la generación de espectrogramas se utilizó la librería de Python LibRosa (librosa, 2020) y para la implementación en general de los modelos de redes neuronales se utilizó el framwork de machine learning TensorFlow (Google, 2020) dada su simpleza, comunidad activa y soporte constante.

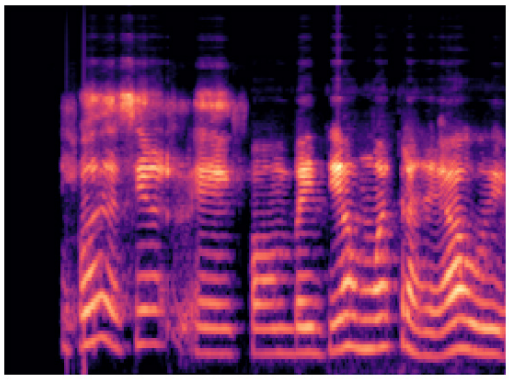

Figura 2 – Espectrograma en escala de Mel

Los espectrogramas se calculan con referencia en dB relativa al valor más alto de la serie temporal. Se calcula los espectrogramas en escala de Mel, con la intención de que variaciones en frecuencias muy alejadas a al rengo vocal de las personas tengan poco impacto en la imagen final que se alimenta a la red.

Proponemos dos modelos inspirados en la popular arquitectura AlexNet (Krizhevsky, Sutskever, & Hinton, 2017). El primero es completamente convolucional y tiene una estructura como se lo muestra en la figura 3 a continuación.

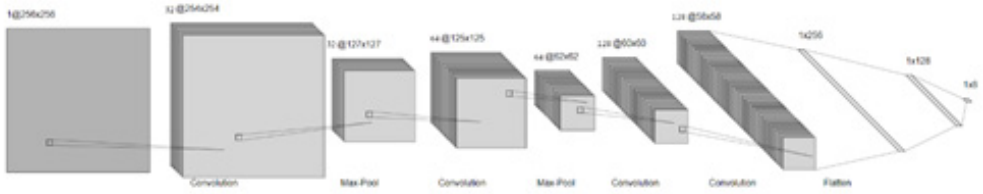

Figura 3 – Modelo de red neuronal con capas convolucionales

El segundo, es una variación del primero donde se incluye además una capa LSTM. En la figura 4 debajo se muestra la estructura del segundo modelo propuesto.

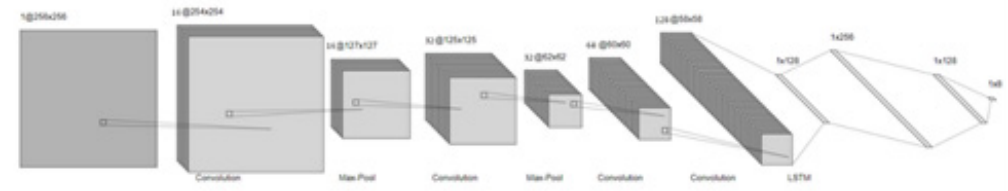

Figura 4 – Modelo de red neuronal con capas convolucionales y LSTM

Para ambos modelos se optó por kernles cuadrados y de tamaño constante, 3x3, con un stride de (1,1). Tanto el incremento del kernel como el tamaño del max pooling o la frecuencia del mismo, no dieron buenos resultados. Para minimizar los tiempos de entrenamiento y maximizar los resultados obtenidos se empleó batch normalization como método de optimización y resultó esencial para reducir los tiempos de entrenamiento y alcanzar los altos niveles de precisión que más adelante se exponen. Otros métodos de normalización como el uso de drop out, wheigt normalization y layer normalization no fueron tan efectivos.

El entrenamiento se realiza de forma clásica, separando en 20% del set de datos para pruebas y se aplicaron técnicas de data augmentation con la intención de romper patrones semánticos que puedan ser capturados, producto de que el set de datos contiene varias oraciones iguales. Los espectrogramas son redimensionados de su resolución original de 1890 x 1410 a 256 x 256. Después se aplica un volteo horizontal aleatorio con probabilidad del 50% y modificaciones del ancho aleatorias también con probabilidad del 50%. Alteraciones en la altura de la imagen no fueron efectivas. A continuación, se presenta la tabla 1 con los resultados de precisión en el set de pruebas luego del entrenamiento:

Modelo	**precisión en el set de pruebas**
1 CNN	92.53
2 CNN+LSTM	82.62

Tabla 1 – Tabla comparativa del desempeño de ambos clasificadores propuestos

En la figura 5 se muestran las matrices de confusión de ambos modelos luego de las pruebas.

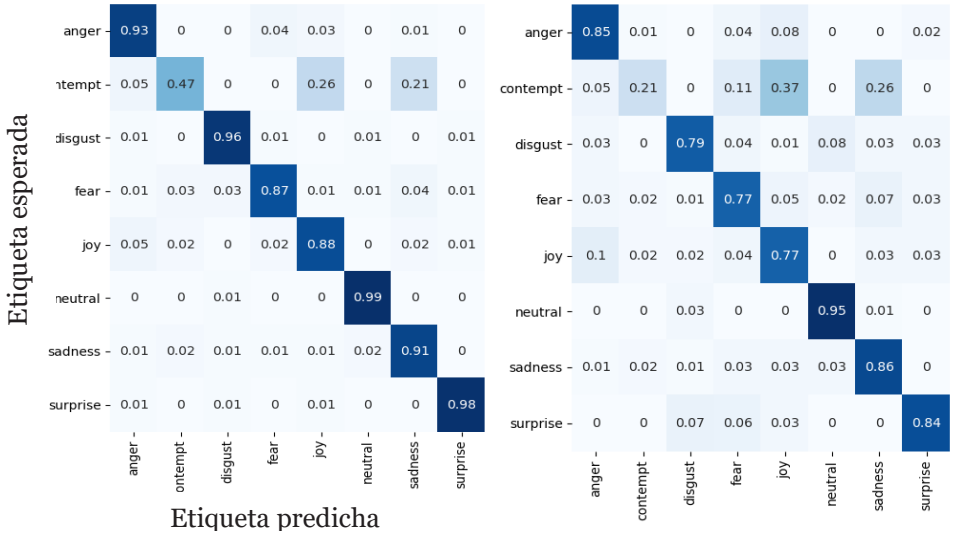

Figura 5 – Matriz de confusión para el primer modelo (CNN)(Izquierda).
Matriz de confusión para el modelo CNN+LSTM (Derecha)

3.2. Integración de resultados del clasificador bajo el enfoque dimensional

Para la integración de los resultados del clasificador, en particular se desarrolló un método de transformación vectorial aplicado a las predicciones categóricas provistas por el clasificador para su representación dimensional, concretamente se confecciona un API que recibe muestras de audio de entre 3 y 4 segundos preferentemente, dado que esa longitud de muestra es la más representativa dentro del set de entrenamiento mixto mencionado en secciones anteriores. El desarrollo produce los espectrogramas en escala Mel y los pasa por el modelo neuronal puramente convolucional para devolver un objeto JSON con las ocho etiquetas emocionales alegria (joy), miedo (fear), tristeza (sadness), ira (anger), disgusto (disgust), sorpresa (surprise), desprecio (contempt), neutral (neutral) y su correspondiente valor de probabilidad predicho, valores de excitación-valencia asociados y promedio de dB de la muestra como se ve a continuación:

{
“anger”: “0.0”,
“surprise”: “0.0”,
“disgust”: “0.1”,
“fear”: “0.3”,
“sadness”: “4.6”,

“neutral”: “6.0”,
“contempt”: “20.7”,
“joy”: “68.3”,
“valence”: “-2.81735”,
“arousal”: “0.836”,
“sampledB”: “-51.525”
}

El API fue desarrollado con Flask (Ronacher, 2015), un framework para desarrollo de servicios web en Python y funciona como una interfaz para ser explotada por otras aplicaciones. Implementa los clasificadores que fueron presentados en la sección 3.1 y en particular el algoritmo de transformación dimensional(excitación-valencia) que se presenta en el cuerpo principal de esta sección.

Este sistema contribuye con el desafío de integración en contexto multimodales que integran diversas fuentes de emociones, en particular la trasformación de enfoques categóricos a dimensiones facilitando la fusión de datos emocionales de diversas fuentes (rostro, voz, EEG, ECG, HR, entre otras). La representación dimensional de los valores de excitación y valencia predichos se desarrolla con el empleo del circunflejo de Russel, en este orden se toma una coordenada de origen asociada a la emoción más probable según el clasificador y, a partir de ahí, desviarse en dirección y magnitud proporcional (a la probabilidad correspondiente) hacia los puntos asociados a las dos emociones siguientes más probables.

Para esto se optó por elegir una expansión del modelo de Russel presentada por Klaus Sherer en la publicación “What are emotions? And how can they be mesaure?” (Scherer, 2005) para tener las coordenadas de base necesarias para asociar las ocho emociones que trabaja el clasificador en base al modelo categórico de Ekman.

A diferencia del modelo original de Russel, Sherer no solo lo extiende, sino que lo presenta con una orientación inversa. En este trabajo a fines de compatibilidad con el framework presentado en (Ierache, Ponce, Nicolosi, Sattolo, & Chapperon, 2019) (Ierache, Nervo, Sattolo, & Chapperon, 2020) (Ierache, Sattolo, & Chapperon, 2020) se mantiene el enfoque original de Russel en relación a la orientación de los valores de positividad y negatividad de valencia, es decir, positivos del lado derecho y negativos del lado izquierdo.

Además, solo se extraen las emociones de las cual carece el circunflejo de Russel por lo que el modelo utilizado para trabajar es uno como el que muestra la figura 7.

A continuación, se detalla el algoritmo propuesto para obtener los valores de excitación y valencia predichos, este método se basa en una estrategia de desplazamiento vectorial:

1. Construir un mapa de emociones asociando las coordenadas extraídas del circunflejo de Russel (Russell, 1980) extendido por Sherer (Scherer, 2005) con las 7 emociones planteadas por Ekman (Ekman, 2005) y neutral, tal como se ve en la tabla 2.

Figura 7 – Circunflejo basado en publicaciones de Russel y Sherer

Emoción	Coordenada (x,y)
Anger	(-3.2, 2.3)
Contempt	(-2.4, 2.6)
Disgust	(-2.6, 1.9)
Fear	(-1.8, 3)
Joy	(3.5, 0.7)
Neutral	(1.1, -1.3)
Sadness	(-3.4, -1.5)
Surprise	(0.4, 3.8)

Tabla 2 – Mapa de coordenadas

2. Analizar una muestra de audio y extraer las 3 etiquetas emocionales más probables predichas (las de mayor valor).
3. Consultar el mapa del punto 1 y tomar (x_1, y_1) como la coordenada asociada a la emoción más probable, (x_2, y_2) la coordenada de la segunda más probable, y (x_3, y_3) la coordenada de la tercera.
4. Aplicar la fórmula 1 para extraer (x_{r2}, y_{r2}) usando la probabilidad predicha para la segunda emoción más probable y remplazando (x_x, y_x) con (x_2, y_2).
5. Aplicar la fórmula 1 para extraer (x_{r3}, y_{r3}) usando la probabilidad predicha para la tercera emoción más probable y remplazando (x_x, y_x) con (x_3, y_3)

$$x_{ri} = x_1 + \left(prob.predicha\ para\ i * (x_i - x_1)\right)$$
$$y_{ri} = y_1 + \left(prob.predicha\ para\ i * (y_i - y_1)\right)$$

6. Promediar x_2 con x_{r3} e y_{r2} con y_{r3} para obtener una única coordenada resultante (x_{r4}, y_{r4})
7. Los valores x_{r4} e y_{r4} obtenidos luego del promedio son asociados directamente con valores de excitación y valencia.

El procedimiento anterior se puede visualizar gráficamente con el siguiente ejemplo. Suponemos una predicción del clasificador donde las tres emociones más probables fueron, Neutral con 0.50, Alegría con 0.25 y Tristeza con 0.25. En la figura 8, se tomó el punto O (x_1, y_1) y se calcularon los puntos A (x_{r2}, y_{r2}) y B (x_{r3}, y_{r3}) usando la fórmula 1 como se explicó en el apartado anterior. El vector w, es el promedio de los vectores u y v. Finalmente del punto C, asociado al vector w, se extraen valores de excitación y valencia final. Como se mencionó al inicio del apartado, lo que ocurre esencialmente es una desviación del punto O, que representa la emoción más probable, en dirección y magnitud proporcionales a los puntos A y B, que representan la segunda y tercera emoción más probable respectivamente considerando su origen sobre la emoción más probable.

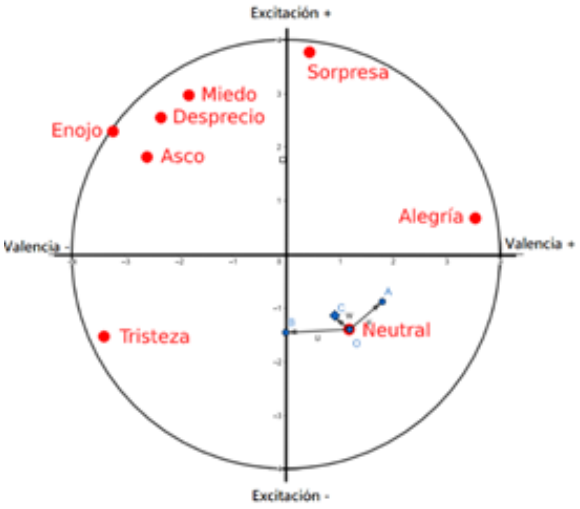

Figura 8 – Ejemplo grafico del método de transformación donde el punto C, denotado con un rombo, es el resultado desde el cual se extraen valores de excitación y valencia

Para los valores de predicción ya mencionados, los puntos resultan:

O = (1.1, -1.3), A = (1.7, -0.8), B = (0.0, -1.3), C= (0.85, -1.05)

Siendo Excitación = -1.05 y Valencia = 0.85 en relación con el punto C.

4. Pruebas del clasificador y del método de transformación con set de datos independiente (español rioplatense)

La falta de sets de datos en diferentes variaciones del español es un problema para la generalización del clasificador. Sin embargo, gracias a las técnicas de data argumentación y a la naturaleza de la extracción de características de un espectrograma, el clasificador desarrollado demostró inicialmente la capacidad de generalización. Para evaluar esto se recolectaron, 22 audios de novelas argentinas (Elkfury, Set de datos experimental, 2020) y se las clasifico en forma independiente con un grupo de ocho personas por medio de encuestas SAM (Self-Assessment Manikin) figura 9, que les permite a los sujetos calificar los audios en términos de excitación y valencia (Bradley & Lang, 1994).

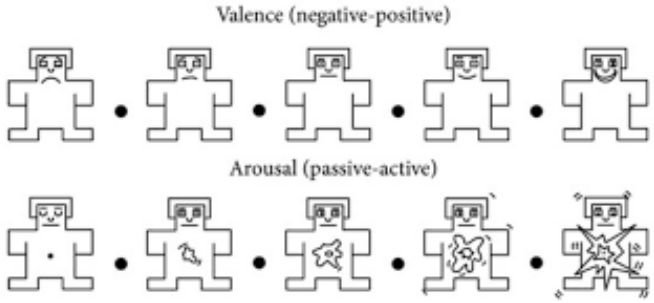

Figura 9 – SAM (Self-Assessment Mankin)

La figura 10 a y 10 b muestran la representación en el circunflejo de Russel, de los resultados de clasificación en particular de dos audios recolectados, donde los puntos negros muestran las clasificaciones de los sujetos con el empleo de la encuesta SAM y el rombo azul la clasificación predicha por la red convolucional utilizando el método de transformación vectorial antes presentado y entrenada con el set de datos mixto. Se puede apreciar en particular para estos dos audios representados una coincidencia en los cuadrantes de pertinencia entre los reportes SAM de los ocho usuarios y la determinada por la representación vectorial propuesta a partir del resultado del clasificador. Sobre la experimentación inicial con los veintidós audios de videos se obtuvo un 72% de coincidencia de cuadrante, entre los valores representados con el enfoque vectorial propuesto sobre la base de la predicción del clasificador y los valores de los encuestados (SAM).

5. Discusión

En el presente trabajo se desarrolló un enfoque categórico al igual que los trabajos indicados en la primera sección, se distinguen entre 4 y 8 emociones, en el contexto de las emociones básicas que propone Ekman. Sin embargo, se diferencia de ellos al proponer un método de transformación de etiquetas categóricas a valores de excitación/valencia para facilitar su contrastación y validación con otros métodos de educción emocional en un contexto multimodal

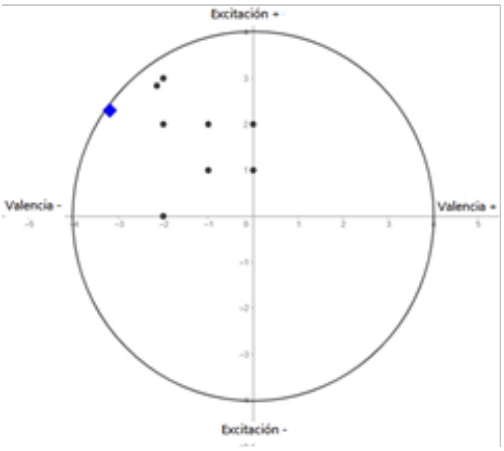

Figura 10a – Audio 1 correspondiente a E+, V -

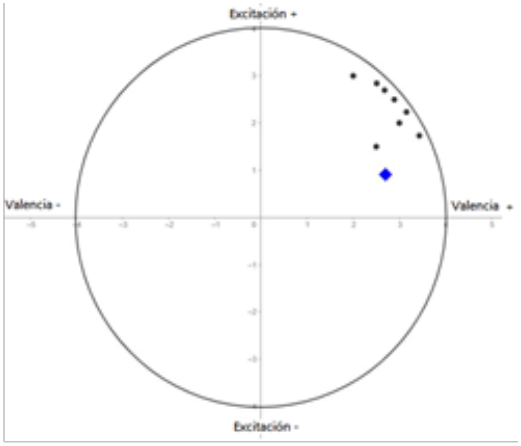

Figura 10b - Audio 2 correspondiente a E+, V+

Se plantea un enfoque similar en relación el uso de arquitecturas de clasificadores con redes neuronales, que adoptaron los trabajos de (Mustaqeem & Kwon, 2019) y (Badshah et al., 2017).

Para los espectrogramas a diferencia de los trabajos anteriores se aplicó para los espectrogramas la escala de Mel, ya que esta realza frecuencias del rango vocal humano, aumentando la resistencia del clasificador al ruido, mejorando la robustez del mismo.

6. Conclusiones

En relación con la confección de un modelo de clasificador de emociones se puede decir que el uso de espectrogramas y redes convolucionales tiene un buen resultado y que las técnicas de data augmentation fueron fundamentales para lograr los altos niveles de precisión obtenidos. Queda por delante la recolección de una mayor cantidad de muestra de audio en castellano rioplatense, en principio para realizar pruebas y determinar el grado de generalización que tiene un clasificador que se entrenó, en su mayoría, con muestras habladas en español de Castilla (Emofilm+Inters1p). Si se tuviera un set de datos lo suficientemente grande podrían combinarse ambos sets para incrementar la versatilidad del clasificador.

El método propuesto para obtener los valores de excitación y valencia predichos contribuye a mejorar la clasificación, cuando la predicción del clasificador tiene niveles parciales de incertidumbre. En futuras líneas de investigación se explorarán métodos dimensionales con más etiquetas emocionales como lo es la Rueda de Ginebra.

En orden a nuestras futuras líneas de trabajo, considerando la precisión obtenida por el modelo CNN, se considera ampliar el set de datos empelado en el entrenamiento con la propuesta de datos rioplatense, como así también se espera ampliar la experimentación con otras técnicas de data augmentation como las vistas en (Yafeng Niu, 2017) y variaciones de los hyperparametros del modelo, por ejemplo, las alternativas a la forma del kernel que propone (Badshah et al., 2017)

La integración con en contextos multimodales emocionales es bastante directa permitiendo explotar y contrastar los datos obtenidos junto con los de los otros sensores y métodos de educción emocional. El método de transformación vectorial aplicado al enfoque categórico para su representación dimensional demostró un resultado alentador, pero queda por delante realizar pruebas abiertas para certificar su validez con encuestas SAM y datos de otros sensores.

Referencias

Badshah, A. M., Rahim, N., Ullah, N., Ahmad, J., Muhammad, K., Lee, M. Y., & Baik, S. W. (2017). Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications, 78*, 5571–5589. https://doi.org/10.1007/s11042-017-5292-7

Caballero-Morales, S.-O. (2013). Recognition of emotions in Mexican Spanish speech: An approach based on acoustic modelling of emotion-specific vowels. *TheScientificWorldJournal,* 2013, 162093-162093. PubMed. https://doi.org/10.1155/2013/162093

Chanchí, G. E. G., & Córdoba, A. E. G. (2019). Análisis de emociones y sentimientos sobre el discurso de firma del acuerdo de paz en Colombia. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (E22), 95–107.

Chanchí, G. E. G., Sierra, L. M. M., & Ospina, M. A. A. (2020). Aplicación de la computación afectiva en el análisis de videos promocionales de turismo de la ciudad de Popayán-Colombia. *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, (E36), 341–354.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366. https://doi.org/10.1109/tassp.1980.1163420

Ekman, P. (2005). Basic Emotions. In *Handbook of Cognition and Emotion* (pp. 45–60). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013494.ch3

Elkfury, F., & Ierache, J. (2020). *Set de datos experimental*. https://doi.org/ 10.17632/v3vm6pf2d3.1

ELRA. (2011). *Emotional speech synthesis database*. Retrieved from http://catalog.elra.info/en-us/repository/browse/ELRA-S0329/

Fukushima, K. (1980, 4). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*, 193–202. https://doi.org/10.1007/bf00344251

Garcia, S. P. (2013). Reconocimiento afectivo automático mediante el análisis de parámetros acústicos y lingüísticos del habla espontánea.

Geneva Emotion Wheel (2005) Retrieved from https://www.unige.ch/cisa/gew

Google. (2020). *TensorFlow*. Retrieved from https://www.tensorflow.org/learn

Ierache, J., Nervo, F., Sattolo, I., & Chapperon, G. (2020). Proposal of a multimodal model for emotional assessment within affective computing in gastronomic settings. *XXVI Congreso Argentino de Ciencias de la Computación - CACIC 2020*, La Matanza. (pp 501-511).

Ierache, J., Ponce, G., Nicolosi, R., Sattolo, I., & Chapperón, G. (2019). Valoración del grado de atención en contextos áulicos con el empleo de interfase cerebrocomputadora. Libro de actas XXV Congreso Argentino de Ciencias de la Computación (CACIC), Universidad Nacional de Río Cuarto, Córdoba (pp 417-426). RedUnci.

Jorge, I., Iris, S., & Gabriela, C. (2020). Framework multimodal emocional en el contexto de ambientes dinámicos. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*, (40), 45–59. https://doi.org/10.17013/risti.40.45–59

Ierache, J., Sattolo, I., Chapperon, G., Ierache, R., Nervo, F., Elkfury, F., . . . Nicolosi, R. (2020). Computación afectiva aplicada a la valoración emocional en contextos gastronómicos. *XXII Workshop de Investigadores en Ciencias de la Computación (WICC 2020, El Calafate, Santa Cruz)*, (pp. 664-668).

Jia, Y. (2020). *Caffe*. Retrieved from http://caffe.berkeleyvision.org/

Johnson, M. K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Retrieved 10 2020, from https://bookdown.org/max/FES/recursive-feature-elimination.html

Josh Patterson, Adam. G. (2017). *Deep Learning*. O'Reilly Media, Inc.

Keras. (2020). Retrieved from https://keras.io/

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Ali Mahjoub, M., & Cleder, C. (2020). Automatic Speech Emotion Recognition Using Machine Learning. *Social Media and Machine Learning*. https://doi.org/10.5772/intechopen.84856

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017, 5). ImageNet classification with deep convolutional neural networks. *Communications of the ACM, 60*, 84–90. https://doi.org/10.1145/3065386

Margaret M. Bradley, Peter J. Lang (1994). Measuring emotion: The self-assessment manikin and the semantic differential (pp. 49-59). Journal of Behavior Therapy and Experimental Psychiatry. https://doi.org/10.1016/0005-7916(94)90063-9

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2005). International Affective Picture System. *International Affective Picture System*. American Psychological Association (APA). https://doi.org/10.1037/t66667-000

librosa. (2020). *Libreria LibRosa*. Retrieved from https://librosa.github.io/librosa/

Loijens, L., & Krips, O. (2020). *FaceReader Methodology Note*. Retrieved noviembre 10, 2020, from https://www.noldus.com/facereader/resources

López, M. B., Montes, A. J. H., Ramírez, R. V., Hernández, G. A., Cabada, R. Z., & Estrada, M. L. B. (2016). EmoRemSys: Sistema de recomendación de recursos educativos basado en detección de emociones. *RISTI-Revista Ibérica de Sistemas e Tecnologias de Informação*, 17, 80–95. https://doi.org/10.17013/risti.17.80–95

Mehrabian, A. (2017). Communication Without Words. In *communication theory* (pp. 193–200). Routledge. https://doi.org/10.4324/9781315080918-15

Mustaqeem, X., & Kwon, S. (2019). A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. *Sensors* (Basel, Switzerland), 20(1), 183. https://doi.org/10.3390/s20010183

Parada-Cabaleiro, E., Costantini, G., Batliner, A., Baird, A., & Schuller, B. (2018). *EmoFilm - A multilingual emotional speech corpus*. Retrieved from https://zenodo.org/record/1326428#.XoyMIIgzbcs

Patterson, J., & Gibson, A. (2017). Deep learning: A practitioner's approach.

Ronacher, A. (2015). *Flask*. Pallets Projects. https://palletsprojects.com/p/flask/

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178. https://doi.org/10.1037/h0077714

Sánchez-Gutiérrez, M. E., Albornoz, E. M., Martinez-Licona, F., Rufiner, H. L., & Goddard, J. (2014). Deep Learning for Emotional Speech Recognition. *Lecture Notes in Computer Science*, 311-320. https://doi.org/10.1007/978-3-319-07491-7_32

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44*, 695–729. https://doi.org/10.1177/0539018405058216

Shafkat, I. (2018). *Intuitively Understanding Convolutions for Deep Learning*. Retrieved 10 2020, from https://towardsdatascience.com/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1

Volkmann, J., Stevens, S. S., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America, 8*, 208–208. https://doi.org/10.1121/1.1901999

Xavier Glorot, A. B. (2011). *Deep sparse rectifier neural networks*.

Yafeng Niu, D. Z. (2017). *A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks*.