Reconocimiento de emociones en la voz empleando redes neuronales y su integración en frameworks multimodales de educción emocional

Jorge Ierache, Fernando Elkfury

Instituto de Sistemas Inteligentes y Enseñanza Experimental de la Robótica (ISIER)

Escuela superior de Ingeniería, Informática y Ciencias Agroalimentaria

Secretaria de Ciencia y Tecnología Universidad de Morón Cabildo 134, Buenos Aires, Argentina

jierache@unimoron.edu.ar

Resumen

El habla es una de las formas naturales para que los humanos expresen sus emociones. Es fácil de obtener y procesar en escenarios en tiempo real, pero, sin embargo, el reconocimiento automático del habla emocional 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 describe el diseño y desarrollo de redes neuronales para la clasificación de emociones en el discurso hablado (voz), se proponen diferentes métodos para convertir un enfoque categórico de clasificación de emociones a uno dimensional y la integración del clasificador con frameworks multimodales de captura de emociones.

Palabras clave: Aprendizaje automático, redes neuronales, reconocimiento de emociones en la voz

Contexto

Esta investigación aplicada se desarrolla en el contexto del Proyecto de Investigación Científica Tecnológica Orientado (PICTO) aprobado por la Agencia Nacional de promoción de la investigación, el desarrollo tecnológico y la innovación(ANPCyT),denominado "Influencias del estado biométrico emocional de personas interactuando en contextos de entornos simulados, reales e interactivos con robots". El mismo se desarrolla dentro del Instituto de Sistemas Inteligentes y Enseñanza Experimental de la Robótica de la Universidad de Morón (ISIER-UM) y está auspiciado por la Secretaría de Ciencia y Tecnología con una duración de tres años. Este trabajo se sustenta sobre las bases iniciales de las investigaciones realizadas en el marco de los proyectos: "Influencias del estado biométrico-emocional de personas interactuando en contextos de entornos virtuales" Ping/17-03-JI-004,(2017-2019), el proyecto denominado "Explotación de datos EEG y parámetros fisiológicos de usuarios interactuando en contextos virtuales" (DC diálogo con las ciencias 2018-2020) - UM-2019 código 80020190100007 UM y el proyecto presentado "Valoración Emocional Multimodal aplicada en contextos gastronómicos" convocatoria PIO 2019 UM.

Introducción

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 pueden transmitir las personas

en un momento dado [1].

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, dada su capacidad para detectar la estructura compleja y sus características asociadas, sin la necesidad de extracción y ajuste manual de estas. Lo cual es un aspecto clave en el desarrollo, 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.

En este trabajo se optó por el uso de espectrogramas, en los cuales las frecuencias fueron convertidas a escala de Mel [2] [Figura 1], y se evalúa el desempeño de redes neuronales convolucionales (CNN) [3] [4] y redes neuronales recurrentes (RNN) [5] para la construcción de un clasificador de emociones.

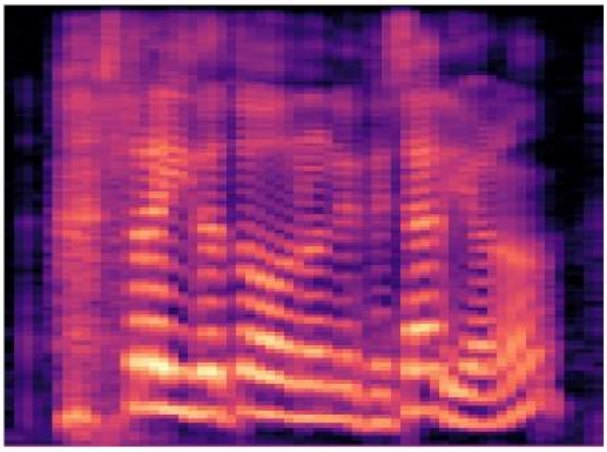

Figura 1. Ejemplo de espectrograma con las frecuencias en escala de Mel.

Líneas de Investigación, Desarrollo e Innovación

Este proyecto se desarrolla en base al aporte potencial que puede tener la educción emocional a partir del discurso hablado (voz) en un framework multimodal de análisis emocional, como el visto en [6] [7] [8] [9] [10] [11] [12] que está asociado al contexto de esta investigación.

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. [13].

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 se adapte a la variante de español latino rio platense. 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 que nos permita trabajar la voz en un framework multimodal de análisis emocional.

Objetivos y Resultados Obtenidos

Con intención de dar respuesta a los problemas planteados se propuso:

1. Desarrollar un clasificador de emociones que pueda reconocer emociones en el discurso hablado en español rioplatense.
2. Establecer una comparativa de performance entre clasificadores basados en RRN y CNN.
3. Integrar el clasificador con un framework multimodal de captura de emociones.
    a. Desarrollar una API que facilite la explotación del modelo diseñado.
    b. Proponer e implementar un método para pasar del enfoque categórico a un enfoque dimensional.

 RedUNCI - UNdeC. ISBN: 978-987-24611-3-3

4. Sentar las bases para la mejora continua a partir de la adquisición de nuevos datos para el entrenamiento del modelo elegido.

Los modelos de redes neuronales construidos y entrenados con muestras de audios provenientes de dos sets de datos, INTERS1P (ELRA) [14] y el EMOFILM [15], obtuvieron resultados a la altura del estado del arte [16] [17] [18], en relación al valor de precisión en el set de pruebas.

A continuación, se presenta una tabla con los valores de precisión obtenidos para un modelo puramente convolucional y otro que combina capas convolucionales con capas LSTM [5]:

Modelo	Prec. En el set de pruebas
1 CNN	92.53
2 CNN+LSTM	82.62

Tabla 1. Precisión de ambos modelos de clasificadores construidos

Uno de los desafíos de integración con frameworks multimodales es pasar de un enfoque categórico a uno dimensional que nos permita contrastar equitativamente los datos de todos los métodos de educción emocional.

En este aspecto se trabajó con 2 enfoques. El primero se basa en obtener valores de valencia a partir de la resta de la probabilidad obtenida para la etiqueta "alegría" y la emoción negativa más probable, como lo plantean Leanne Loijens et al. [19] Para los valores de excitación se plantea utilizar la diferencia entre los valores medios de dB de las subsecuentes muestras tomadas durante una sesión de evaluación de un sujeto y los valores promedio de dB de la muestra particular evaluada

Otra forma presentada, llamada vectorial, es usar el circunflejo de Russel [20] para tomar 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 2 emociones siguientes más probables predichas.

Para esto se optó por elegir una expansión del modelo de Russel presentada por Klaus Sherer [21], de forma de tener las coordenadas necesarias para asociar las 8 emociones con las cuales trabaja el clasificador en base al modelo categórico de Ekman.

Solo se extrajeron las emociones de las cual carece el circunflejo de Russel por lo que el modelo utilizado para trabajar es el que se muestra en la figura 2 debajo.

Figura 2 Circunflejo basado en publicaciones Russel y Sherer

Dado que este trabajo se desarrolla bajo las líneas de investigación del ISIER UM [22], para las pruebas se implementó el API del clasificador con el framework visto en [6] [7] [8] [9] [10] [11] [12]. Haciendo uso de la última versión en desarrollo del framework, se hicieron pruebas con una metodología muy similar a la planteada en los artículos mencionados. Se pretende capturar la voz del sujeto de prueba y extraer su estado emocional a partir de la misma mientras se lo estimula con imágenes obtenidas del IAPS [23].

De esta forma se pueden obtener datos sincronizados de excitación valencia desde diferentes sensores para su contrastación y consecuente validación.

Para evaluar la capacidad de generalización del clasificador incorporado se recolectaron 22 audios de novelas argentinas y se las clasificó con un grupo de 8 personas por medio de encuestas SAM [24].Se obtuvo un 72% de precisión, que, si bien es alentador, está sujeto a la necesidad de una evaluación más extensiva dado el volumen reducido de muestras disponibles para las pruebas.

Este proyecto sienta las bases para la continuación de la línea de investigación que vincula la educción de emociones en el discurso hablado (voz) con un framework multimodal de análisis emocional. Se abre el camino para la mejora continua de clasificadores basados en redes neuronales a partir de la recolección de muestra de audio en castellano rioplatense, como también para validar los avances hasta ahora conseguidos.

Formación de Recursos Humanos

El grupo de investigación se compone de un investigador formado y tres investigadores en formación. En el marco de la investigación se finalizó una tesis de grado en ingeniería en informática y se encuentra en etapa inicial el proyecto de una tesis de doctorado.

Referencias

[1] A. Mehrabian, «Communication Without Words,» de *communication theory*, Routledge, 2017, p. 193–200.

[2] J. Volkmann, S. S. Stevens y E. B. Newman, «A Scale for the Measurement of the Psychological Magnitude Pitch,» *The Journal of the Acoustical Society of America,* vol. 8, p. 208–208, 1 1937.

[3] K. Fukushima, «Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,» *Biological Cybernetics,* vol. 36, p. 193–202, 4 1980.

[4] I. Shafkat, «Intuitively Understanding Convolutions for Deep Learning,» Junio 2018. [En línea]. Available: https://towardsdatascience.com/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1. [Último acceso: 10 2020].

[5] A. G. Josh Patterson, «Deep Learning,» [En línea]. [Último acceso: 10 2020].

[6] J. Ierache, G. Ponce, R. Nicolosi, C. Cervino y E. Eszter, «Registro emocional de las personas interactuando en contextos de entornos virtuales,» *CACIC 2018*, La Plata, Plata, pp 877-886, ISBN: 978-950-658-472-6

[7] J. Ierache, G. Ponce, R. Nicolosi, I. Sattolo y G. Chapperon, «Valoración del grado de atención en contextos áulicos con el empleo de interface cerebro-computadora en el marco de la computación afectiva,» *CACIC 2019*, Rio Cuarto, pp: 417-426, ISBN:978-987-688-377-1

[8] J. Ierache, I. Sattolo, G. Chapperon, R. Ierache, F. Nervo, F. Elkfury, G. Ponce y R. Nicolosi, «Computación afectiva aplicada a la valoración emocional en contextos gastronómicos,» *WICC 2020*, El Calafate, pp: 664-668. ISBN: 978-987-3714-82-5

[9] J. Ierache, F. Nervo, I. Sattolo y G. Chapperon, «Propuesta de un Modelo Multimodal de valoración emocional en el marco de la computación afectiva aplicado en ambientes gastronómicos,» *CACIC 2020*, La Matanza. En Prensa.

[10] C. Barrionuevo, J. Ierache y I. Sattolo, «Reconocimiento de emociones a través de expresiones faciales con el empleo de aprendizaje supervisado aplicando regresión logística,» *CACIC 2020* , La

Matanza. En prensa.

[11] J. Ierache, I. Sattolo, G. Chapperon, R. Ierache, F. Elkfury, C. Barrionuevo y F. Nervo, «Captura multimodal de estados emocionales aplicado a contextos de computación Afectiva,» *WICC 2021*, Chilecito, 2021. Comunicado 24/2/2021.

[12] J. Ierache, I. Sattolo y G. Chapperon, «Framework multimodal emocional en el contexto de ambientes dinámicos». DOI. 10.17013/risti 40. 45-59 ISSN: 1646-9895.

[13] S. P. Garcia, *reconocimiento afectivo automático mediante el análisis de parámetros acústicos y lingüísticos del habla espontánea.*

[14] ELRA, «Emotional speech synthesis database,» [En línea]. Available: http://catalog.elra.info/en-us/repository/browse/ELRA-S0329/.

[15] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird y B. Schuller , «EmoFilm - A multilingual emotional speech corpus,» [En línea]. Available: https://zenodo.org/record/1326428#.XoyMIIgzbcs.

[16] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub y C. Cleder, «Automatic Speech Emotion Recognition Using Machine Learning,» 2018.

[17] Mustaqeem y S. Kwon, «A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition,» *Sensors,* vol. 20, p. 183, 12 2019.

[18] M. Sánchez-Gutiérrez, E. M. Albornoz, F. Martinez, L. Rufiner y J. Goddard, «Deep Learning for Emotional Speech,» 2014.

[19] L. Loijens y O. Krips, «FaceReader Methodology Note,» [En línea]. Available: https://www.noldus.com/facereader/resources. [Último acceso: 10 noviembre 2020].

[20] J. A. Russell, «A circumplex model of affect.,» *Journal of Personality and Social Psychology,* vol. 39, p. 1161–1178, 1980.

[21] K. R. Scherer, «What are emotions? And how can they be measured?,» *Social Science Information,* vol. 44, p. 695–729, 12 2005.

[22] «Instituto De Sistemas Inteligentes y Enseñanza De La Robótica (ISIER),» [En línea]. Available: http://isierum.c1.biz/.

[23] P. J. Lang, M. M. Bradley y B. N. Cuthbert, *International Affective Picture System,* American Psychological Association (APA), 2005.

[24] P. J. Lang, «The Cognitive Psychophysiology of Emotion,» de *Anxiety and the Anxiety Disorders*, Routledge, 2019, p. 131–170.