ReTEC

Revista Electrónica de Tecnología, Educación y Ciencia
ISSN: 2953-5654
http://retec.unsa.edu.ar
Universidad Nacional de Salta

Trabajo final de grado

Sistema Automático para el Reconocimiento de Estados Emocionales de Personas Analizando su Voz Aplicado al Sistema de Emergencias 911 de Salta (SE911)

Jacobo León Juárez

Facultad de Ciencias Exactas – Universidad Nacional de Salta
leonjuarez777@gmail.com

Revista Electrónica de Tecnología, Educación y Ciencia,

Volumen 1, Número 1, pág.85-87, jun, 2023. ISSN: 2953-5654

Disponible en http://retec.unsa.edu.ar/

Trabajo final de Grado

Sistema Automático para el Reconocimiento de Estados Emocionales de Personas Analizando su Voz Aplicado al Sistema de Emergencias 911 de Salta (SE911)

Jacobo León Juárez

Facultad de Ciencias Exactas – Universidad Nacional de Salta
leonjuarez777@gmail.com

Director: Lic. Ismael Orozco (UNSa)

Co-Director: Ing. Emilio Javier Leyes

Carrera: Licenciatura en Análisis de Sistemas

Año: 2022

Objetivo

El objetivo de este trabajo de Tesis de Grado fue desarrollar un sistema de Deep Learning (DL) aplicado a los problemas de Speech Emotion Recognition (SER) y Speaker Diarization (SD), adaptado específicamente para detectar el nivel de estrés del operador del SE911 utilizando el análisis del tono de voz como indicador. El sistema logró un WAR (Weighted Average Recall) de 90.23% y un UAR (Unweighted Average Recall) de 88.88%.

Resumen del trabajo

Este estudio se llevó a cabo siguiendo la Metodología de Investigación Científica Design Science Research Methodology (DSRM), usando un enfoque iterativo y secuencial que constó de 4 iteraciones. Durante las dos primeras iteraciones, se realizó un exhaustivo estudio de factibilidad y se exploraron las arquitecturas de DL necesarias para abordar el desafío del SER. En un principio, se emplearon datasets públicos, como RAVDESS y EMO-DB, que contenían grabaciones de voz en idioma inglés y alemán respectivamente. Posteriormente, en la tercera iteración, se adaptó el prototipo inicial del SER al dataset propio del SE911.

La última iteración fue la más significativa, ya que, se centró en la investigación y desarrollo de arquitecturas más sofisticadas de Computer Vision (CV) aplicadas al SER y SD. Se investigó el uso de técnicas de Audio DL para el análisis frecuencial, y mediante el campo de CV y del Natural Language Processing (NLP), usando arquitecturas como Transformers y ResNet, se lograron obtener resultados más precisos. Además, se abordó la tarea más compleja de implementar el SD en los datos del SE911, enfrentando desafíos debido a la escasez de sistemas preentrenados en diferentes idiomas y los altos requerimientos computacionales involucrados.

Es importante destacar que estos resultados excepcionales se obtuvieron utilizando un dataset de tan solo 176 audios. Esta limitación resalta aún más la eficacia y relevancia del

sistema desarrollado, demostrando su potencial para ser aplicado en entornos de emergencias médicas, seguridad y protección civil.

Referencias

1. C. S. Kanani, et al, “Shallow over deep neural networks: A empirical analysis for human emotion classification using audio data,” in International Conference on Internet of Things and Connected Technologies. Springer, 2020, pp. 134–146.
2. D. H. Rudd, et al, “Leveraged mel spectrograms using harmonic and percussive components in speech emotion recognition,” in Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, 2022, pp. 392–404.
3. Bredin, Hervé, et al. "Pyannote. audio: neural building blocks for speaker diarization." ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
4. C. Luna-Jiménez, et al, “A proposal for multimodal emotion recognition using aural transformers and action units on ravdess dataset,” Applied Sciences, vol. 12, no. 1, p. 327, 2021.