Reconocimiento de Estados Emocionales de Personas Analizando su Voz: Experiencia entre I+D y Transferencia Tecnológica

Jacobo León Juárez[1], Carlos Ismael Orozco[1], and Emilio Javier Leyes[2]

[1] Departamento de Informática. FCE. Universidad Nacional de Salta, Argentina
[2] Sistema de Emergencia 911. Salta, Argentina.
leonjuarez777@gmail.com; ciorozco.unsa@gmail.com

Abstract. Los sistemas de reconocimiento de emociones del habla (SER) tienen como objetivo identificar el estado emocional de una persona analizando únicamente su voz, es decir, el sistema deberá seleccionar una clase (alegre, enojado, triste, miedo, sorpresa, etc.) aquella que sea más probable para el audio de entrada. SER es un tema de interés en el área de procesamiento digital de audio debido a las potenciales aplicaciones que se pueden desarrollar, por ejemplo: sistemas que interactúan con el humano en base a las emociones percibidas, asistentes en terapias psicológicas, detección de mentiras en interrogatorios, entre otros. En este trabajo se resume nuestra experiencia inicial en el desarrollo de un sistema SER producto de la vinculación entre el DI-UNSa y el SE-911, ambas instituciones de la provincia de Salta.

Keywords: reconocimiento de emociones · voz humana · redes neuronales.

1 Caracterización General del Proyecto

1.1 Instituciones y Empresas Participantes

DI-UNSa: Departamento de Informática, Facultad de Ciencias Exactas de la Universidad Nacional de Salta fue creado el 4 de setiembre de 2002. Actualmente con un plantel docente de más de 40 profesionales dictan 2 carreras de grado. En cuanto a investigación, nuestros docentes trabajan en áreas tales como Procesamiento de imágenes, Visión por computadora, Redes neuronales, Data y Web Mining, Optimización Combinatoria, TICS, Redes, Ingeniería de Software, Cloud Computing, entre otras.
SE-911: El Sistema de Emergencias 911 de Salta, es un servicio que tiene como finalidad brindar respuestas de forma inmediata a la comunidad en el área de Salud, Bienes, Servicios y Ambiente; coordinando las acciones en primera instancia con Policía, Bomberos, Emergencias Médicas (S.A.M.E.C.) y Urgencias Psicológicas (Servicios Psicológicos), y en segunda instancia coordinando acciones con los diferentes organismos y entidades Municipales, Provinciales, Nacionales y ONG.

1.2 Descripción del proyecto

La principal motivación de la vinculación entre el DI-UNSa y el SE-911 se centra en desarrollar un sistema para el reconocimiento de estados emocionales de personas analizando su voz para detectar denuncias falsas y generar una alerta en casos donde la víctima trate de encubrir la denuncia, por ejemplo, en casos de violencia de género. Dicha vinculación forma parte de un convenio firmado entre ambas instituciones. Para el desarrollo del prototipo se estima un tiempo de 1 año con los siguientes lineamientos: (1) Colaboración en I+D. (2) Inclusión de alumnos avanzados del DI-UNSa para la realización de pasantías cortas en el SE-911. (3) Desarrollo de tesis de grado para optar al título de Licenciado en Análisis de Sistemas en el DI-UNSa.

El grupo de investigación está formado por 3 integrantes del DI-UNSa: Jacobo Juárez (estudiante avanzado de LAS). Lic. Ismael Orozco (docente-investigador) y el Dr. Cristian Martínez (docente-investigador y asistente externo del proyecto). Del SE-911, el Ing. Emilio Leyes es coordinador del área de TICs.

2 Detalles de Ejecución del Proyecto

La Tabla 1 resume las principales actividades planificadas en el proyecto.

Actividades	**Pers.**		
	J	O	L
1.1: Estudio de metodología de desarrollo de software	X	X	
2.1: Búsqueda de literatura	X	X	
2.2: Repaso de técnicas de aprendizaje automático y métricas de evaluación	X	X	
2.3: Definición del marco de trabajo	X	X	X
3.1: Elaboración del estado del arte	X	X	
4.1: Elaboración de un prototipo experimental	X	X	
4.2: Evaluación y mejoras	X	X	X
5.1: Construcción de un dataset del SE-911			X
5.2: Adaptación del modelo	X	X	
5.3: Evaluación final	X	X	X

Table 1. Planificación de actividades. La segunda columna muestra la asignación de los integrantes del proyecto, donde J:Juárez. O:Orozco y L:Leyes.

3 Resultados del Proyecto

Actividad 1: Para la etapa de análisis y diseño se estudió Design Science Research Methodology (DSRM) propuesta por Peffers et al. [4]. DSRM es una metodología de desarrollo de software orientado a la investigación científica, la cual consiste en seis fases: (1) identificación del problema y motivación,

(2) definición de los objetivos para la solución, (3) diseño y desarrollo, (4) demostración, (5) evaluación y (6) comunicación. La Fig. 1 muestra el flujo de trabajo de la metodología.

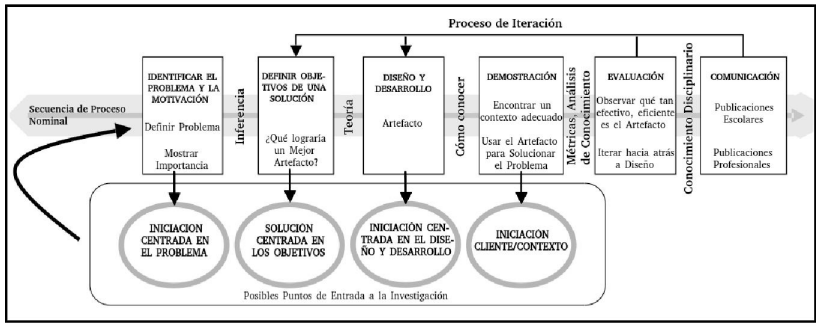

Fig. 1. Fases de la Metodología DSRM

Actividad 2: Como resultados de esta actividad se completaron las fases 1 y 2 de la metodología DSRM. También se definió el marco de trabajo colaborativo usado entre los integrantes del equipo, como por ejemplo: (1) Carpetas compartidas para recopilación de material bibliográfico. (2) implementaciones en entorno colaborativo, particularmente Colab es una herramienta de Google que permite ejecutar código Python en la nube con uso de GPUs y TPUs.

Actividad 3: Estudio de SER. Se elaboró un reporte con los conceptos relevantes del problema junto al estado del arte incluyendo los enfoques más reciente propuestos en la bibliografía. También se familiarizó con conceptos claves del área como: (1) Espectrograma: análisis de cómo fluctúan las vibraciones del sonido que se propagan a lo largo del espacio-tiempo. (2) La transformada de Fourier (DTF) es una representación matemática de la señal de audio. Es una función que recibe una señal en el dominio del tiempo como entrada y descompone dicha señal en diferentes frecuencias, entre otros.

Actividad 4: Para el sistema SER se propuso la implementación de una arquitectura de red neuronal LSTM [2] ya que están formuladas de manera que recordar información por largos períodos de tiempo sea su comportamiento natural. Para la etapa de extracción de características empleamos: (1) Escala de Meel [5] es una transformación no lineal de la escala de frecuencias. Se basa en que los sonidos que estén a una misma distancia (es decir, sean equidistantes), sean audibles para dos observadores humanos, puesto que son equidistantes. Esta escala se suele utilizar en aplicaciones de tratamiento de voz, junto con otros coeficientes, como los mfccs. (2) Los coeficientes MFCC [1] ayudan a extraer las características de la señal de audio más relevantes. Para ello, dividen la señal en tramos (a cada tramo se le aplica una transformada de Fourier discreta o (DFT) para obtener el espectro de la señal. Luego se aplican unos filtros de la mel-scale,

una transformación de logarítmica es aplicada a dicho resultado, finalmente a la transformación anterior se le aplica una transformación de cosenos discreta o (DCT) es aplicada a la transformación anterior.

Para medir el rendimiento, utilizamos la base de datos propuesta por Livingstone et al. [3]. RAVDESS consiste en 24 actores profesionales que vocalizan el acento norteamericano neutral. Los audios están clasificados en las siguientes emociones: neutro, calma, felicidad, tristeza, enojo, temor, disgusto y sorpresa.

Actividad 5: Actualmente, mientras mejoramos la precisión del sistema, el SE-911 se esta encargando de recopilar y categorizar audios para armar una base de datos robusta que sirva para la adaptación del prototipo que se esta desarrollando en la actividad anterior.

4 Balance y Conclusiones

La vinculación entre los integrantes del DI-UNSa y SE-911 se está desarrollando de una manera muy positiva y enriquecedora. La comunicación entre ambas partes es fluida, interactuando a medida que se obtienen avances en el proyecto, o cuando es necesario la toma de decisiones sobre diferentes alternativas. Puntos positivos que se quieren destacar son:

- Primeros resultados prometedores, por ende, se seguirá avanzando sobre el refinamiento.
- Experiencia en el desarrollo de un sistema para un organismo real (SE-911).
- Firmas de convenios de colaboración para la asesoría en el desarrollo/mejora de algunos problemas que el SE-911 planea resolver (y así, motivar la investigación de alumnos avanzados en la carrera).
- Tesis de grado en curso de Juárez, donde Orozco y Leyes son sus directores.

References

1. Todor Ganchev, Nikos Fakotakis, and Kokkinakis George. Comparative evaluation of various mfcc implementations on the speaker verification task. *Proceedings of the SPECOM*, 1, 01 2005.
2. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
3. Steven R. Livingstone and Frank A. Russo. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):e0196391, May 2018.
4. Ken Peffers, Tuure Tuunanen, Charles E Gengler, Matti Rossi, Wendy Hui, Ville Virtanen, and Johanna Bragge. The design science research process: a model for producing and presenting information systems research. In *Proceedings of the first international conference on design science research in information systems and technology (DESRIST 2006)*, pages 83–106. sn, 2006.
5. 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–I/532 Vol. 1, 2005.