



# Proyecto **DEEP FEELING**



Identificar emociones reentrenando un modelo de  
Deep Learning

# ÍNDICE

01. Introducción

02. Estado del arte

03. Objetivo General

04. Objetivos Específicos

05. Hipótesis

06. Metodología preliminar



# INTRODUCCIÓN



El creciente desarrollo de modelos de Deep Learning se ha abierto espacio diversas áreas investigación, donde uno de ellos es la identificación de emociones humanas. Este campo resulta ser desafiante debido a que las emociones pueden estar influenciadas por experiencias previas y percepciones subjetivas.

Durante la exploración de modelos relacionados con este tema se llegó a la constante que todos ellos han sido entrenados con datos de personas angloparlantes, esto plantea el desafío de **aplicarlos en contextos culturales y lingüísticos distintos**, como el de los hispanohablantes.

En este trabajo de investigación, se utilizará el modelo **wav2vec 2.0** para identificar **6 emociones (ansiedad, felicidad, enojo, neutral, aburrimiento y tristeza)**, medidas en dimensiones de **valencia (positivo/negativo)** y **excitación (alta/baja intensidad)**. Para reentrenar el modelo, se empleará la base de datos **EMOVOME**, que destaca por contener grabaciones reales de personas en situaciones auténticas, a diferencia de otras bases de datos que utilizan audios actuados.

El modelo anteriormente mencionado ya fue testeado con los datos de EMOVOME y resultó tener una baja precisión de identificación por sólo haber sido entrenado con datos angloparlantes por lo que se buscará obtener alguna mejora respecto a estos resultados.



# ESTADO DEL ARTE

**El reconocimiento de emociones de la voz (REV) se basa en el análisis de señales de voz para identificar patrones de las emociones usando modelos emocionales discretos y continuos. Para identificar estas emociones se usan bases de datos obtenidas con emociones actuadas, inducidas o naturales. En el idioma español la mayoría de las bases de datos son con emociones actuadas y eso conlleva al primer desafío de este proyecto.**

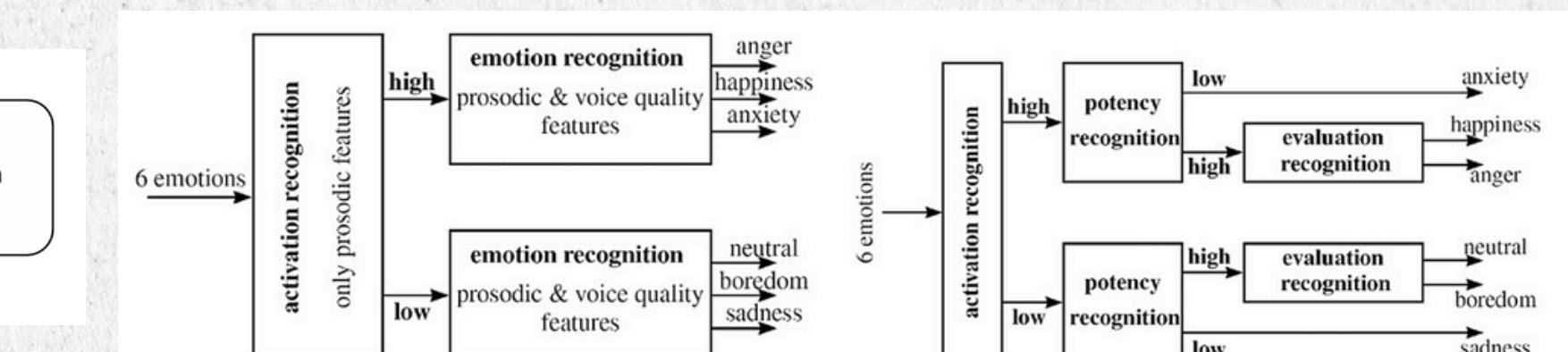
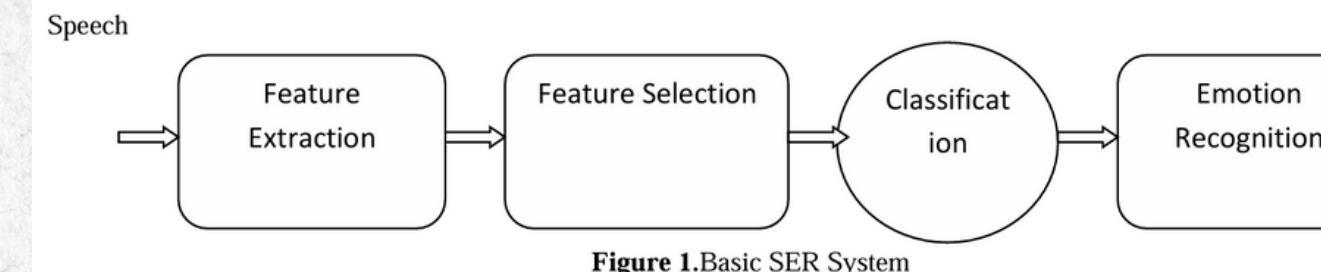
NO.	DATABASE	LANGUAGE	TYPE	SIZE	EMOTIONS
1.	Berlin Emotional Database [EMO-DB] [18]	German	Acted	7 emotions, 10 utterances, 10 speakers (5 male and 5 female)	Neutral, anger, sadness, fear, boredom, happiness, disgust, boredom.
2.	Surrey Audio-Visual Expressed Emotion (SAVEE)[19]	English	Acted	7 emotions, 4 speakers (male), 120 utterances	Surprise, anger, fear, disgust, sadness, neutral, happiness.
3.	RECOLA Speech Database [20]	French	Natural	7 hours of speech, 46 speakers (27 females, 19 males)	5 social behaviors (engagement, performance, agreement, rapport, dominance); valence and arousal.
4.	SAMAINE Database [21]	English Greek Hebrew	Natural	959 conversation, 150 speakers.	power, valence, expectation, activation, overall emotional intensity.
5.	eINTERFACE05 Audio-Visual Emotion Database [22]	English	Elicited	1116 video sequences, 8 females, 34 males, a total of 42 speakers, from 14 different countries.	Surprise, disgust, happiness, fear, anger, sadness.
6.	Interactive Emotional Motion Capture (USC-IEMOCAP)[23]	English	Elicited	Five sessions where each session includes the conversation between two people (one male and one female) and its corresponding labelled speech text	Anger, happiness, sadness, frustration, neutral
7.	FAU Aibo Emotion Corpus [24]	German	Natural	51 children talking to robot dog Aibo, 9 hours of speech	Bored, joyful, helpless, touchy, anger, reprimanding, emphatic, surprised, neutral, motherese, rest.
8.	BAUM-1 Speech Database [25]	Turkish	Acted and Natural	1222 spontaneous video clip, 288 acted, 31 speakers (13female, 18 male)	Anger, surprise, sadness, disgust, contempt, fear, concentration, bothered, being thoughtful, unsure, happiness, boredom, interest.
9.	Oriya Emotion Speech Dataset [26]	Odia/Oriya	Elicited	35 speakers (12 female and 23 male) recoded the text fragments of Oriya drama scripts.	Astonish, sadness, fear, anger, happiness, neutral.
10.	Persian Emotion Speech Dataset [27]	Persian	Simulated	33 native speakers (15 females and 18 males) recorded 748 utterances from Persian Drama Radio Emotional Corpus (PDREC)	Happiness, anger, sadness, surprise, fear, boredom, neutral, disgust.
11.	Assamese Emotion Speech Dataset [28]	Assamese	Simulated	30 students and faculty members (3 males and 3 females per language) recorded 140 utterances of 5 native languages of Assam.	Happiness, surprise, sadness, anger, fear, disgust, neutral.
12.	Chinese Emotion Speech Dataset [29]	Chinese	Simulated	A professional actress of a Reader's Digest Collection recorded 3649 phases and 1500 utterances	Happiness, anger, fear, anger, neutral
13.	Situation Analysis in a Fictional and Emotional corpus (SAFE) [30]	English	Elicited	4724 segments of speech were recorded by students. 400 sequences of audio/visual taken from 30 movies of 7 hours duration.	Positive, negative, neutral
14.	Multilingual Database [31]	Japanese, English, German	Natural and Simulated	Four emotional databases, 1. LEGO emotion database (English) 2. EMO-DB (GERMAN database). 3. UUDB (The Utsunomiya University Spoken Dialogue Database for paralinguistic information studies), and 4. SAVEE (Surrey Audio-Visual Expressed Emotion) corpus in (English).	1. Angry, slightly angry, very angry, neutral, friendly, and nonspeech (critical noisy recordings or just silence) 2. Neutral, anger, fear, joy, sadness, boredom, or disgust 3. Happy-exciting, angry, anxious, sad-bored, relaxed, serene. 4. Anger, disgust, fear, happiness, sadness, surprise, and neutral.
15.	Multilingual Database [32]	Indian English Malayalam and Tamil	Simulated	10 speakers Emotionally biased utterances	Angry, sad, happy.

LISTA DE BASES DE DATOS PROMINENTES EN REV. "A COMPREHENSIVE REVIEW OF SPEECH EMOTION RECOGNITION SYSTEMS.

# ESTADO DEL ARTE

El proceso de REV se puede hacer mediante métodos de machine learning convencionales, Que depende de 3 elementos fundamentales, el preprocesamiento de los datos, la extracción y clasificación de características y la clasificación de emociones basadas en esas características.

Los métodos de clasificación más comunes en esta área son las máquinas de soporte de vectores (SVM), los modelos de mezcla gaussiana (GMM) y los modelos de cadenas de Márkov ocultas (HMM), aunque en el último tiempo se han preferido modelos de deep learning , ya que estos no requieren de la extracción manual de características



# ESTADO DEL ARTE

Los sistemas REV consisten de varias partes, modelamiento basado en lenguaje, modelamiento acústico, extracción y categorización de características, entre otros, lo que significa usar estrategias de clasificación complejas. Para evitar deficiencias de estas estrategias se prefieren modelos deep learning, ya que estos detectan automáticamente estructuras y características complejas, mostrando un mejor rendimiento que con métodos convencionales. Entre las técnicas usadas para estos sistemas están los autoencoders(AE), las redes de creencia profunda(DBN), las redes convolucionales(CNN), las recurrentes(RNN) y los transformers, así como modelos híbridos.

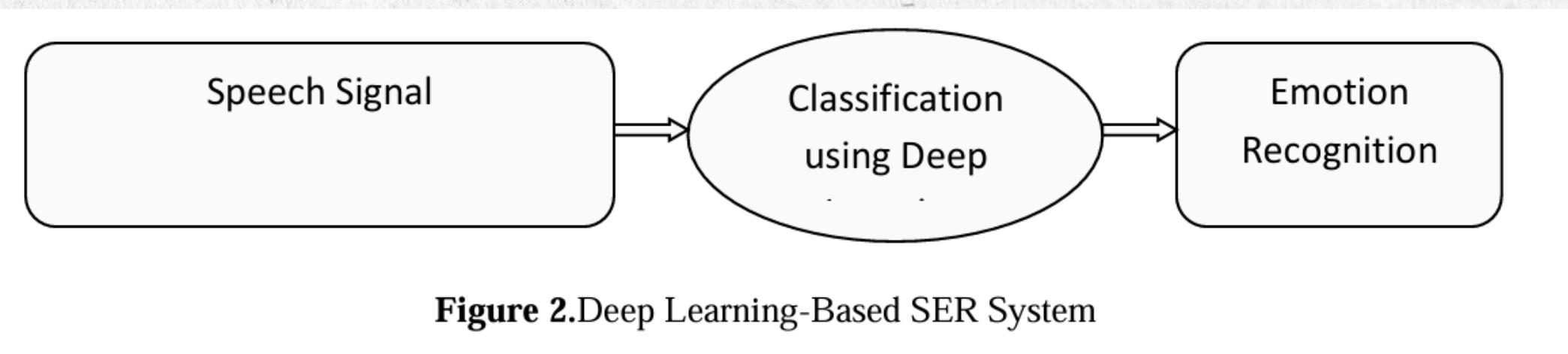
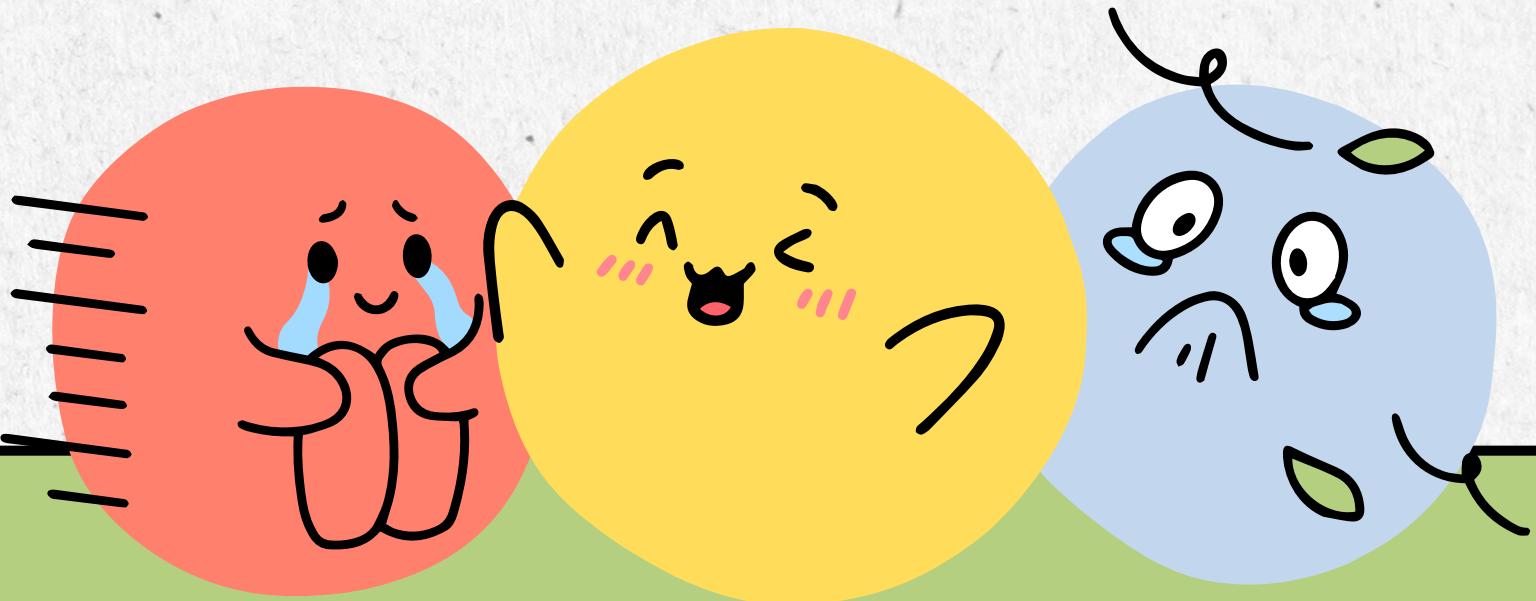


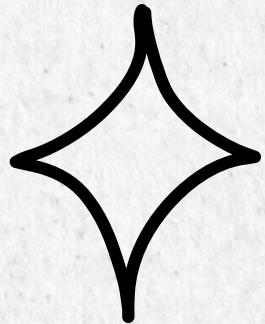
Figure 2. Deep Learning-Based SER System



# OBJETIVO GENERAL

Identificar 6 emociones medidas en valencia y excitación de personas hispanohablante mediante un modelo de Deep Learning ya entrenado con personas angloparlantes.

# OBJETIVOS ESPECÍFICOS



①

Conseguir una base de datos de personas hispanohablantes para el ajuste fino (reentrenamiento) del modelo.

②

Implementar un modelo de acuerdo a la base de datos previamente encontrada para su posterior ajuste fino.

③

Reentrenar el modelo con la base de datos para mejorar la precisión en personas hispanohablantes.

④

Evaluar los resultados del modelo reentrenado para medir cuantitativamente el cambio en la precisión de la identificación de las emociones.

U



# HIPÓTESIS

El modelo de Deep Learning entrenado con datos de personas angloparlantes puede ser ajustado mediante reentrenamiento con datos de personas hispanohablantes para identificar con mayor precisión 6 emociones (ansiedad, felicidad, enojo, neutral, aburrimiento y tristeza) medidas en valencia y excitación respecto al modelo sólo entrenado con personas angloparlantes y una precisión similar a la que obtiene cuando el modelo se testea en personas de habla inglesa.

## 1.- Selección de modelo

Como propuesta inicial tenemos el modelo wav2vec2 preentrenado autosupervisado.



## 3.- Entrenamiento

Se propone utilizar pytorch , implementación basada en Transformer.

# MÉTODOLOGÍA PRELIMINAR

## 2.- Ajuste de modelos

Se propone aplicar average pooling a la ultima capa de transformers y entrenamos el resultado a través de una capa oculta y una capa final de salida



## 4.- Resultados

Finalmente, realizar un análisis de los resultados junto a métricas de rendimiento.



# MUCHAS GRACIAS

*Por su atención*

---



### referencia 1

Sarmah, K., Gogoi, S., Das, H. C., Patir, B., & Sarma, M. J. (2024). A State-of-arts Review of Deep Learning Techniques for Speech Emotion Recognition. *Journal of Electrical Systems*, 20(7s), 1638-1652.

### referencia 2

Gómez-Zaragozá, L., Valls, Ó., del Amor, R., Castro-Bleda, M. J., Naranjo, V., Raya, M. A., & Marín-Morales, J. (2024). Speech emotion recognition from voice messages recorded in the wild. arXiv preprint arXiv:2403.02167.

### referencia 3

T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

### referencia 4

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9), 10745-10759.

