Chapter 1

Metodologia

1.1 Extracción de Características

La extracción de características es una de las secciones más importantes en el reconocimiento de emociones a través de la voz (SER, *Speech Emotion Recognition*) debido a la ambigüedad de las características y la variedad vocal. La extracción de características es el paso principal en el procesamiento del diálogo (speech), y se lleva a cabo para centrarse en la información contenida en la señal, mejorar el grado de similitud y/o diferenciación entre las clases y reducir la dimensionalidad de los cálculos.[**Hellbernd2016**] Hasta ahora, por lo general hay dos categorías de características usadas en SER: Rasgos prosódicos, los cuales extraen información de la prosodia, en concreto, tono, energía y duración, y por otro lado, características del tracto vocal que normalmente indican la distribución de la energía en la frecuencia del rango vocal, a saber, coeficientes cepstrales* La mayoría de los estudios centrados en SER usan rasgos espectrales como la información extraída del tracto vocal. Estos suponen la información derivada del espectro de la señal de voz y se usan para modelar los patrones de entonación y frecuencia del hablante.[**Langari2020**]

Comunmente las técnicas de extracción de características más usadas son MFCC, LPC, LPCC, DWT y PLP. A continuación se ofrece una breve explicación de cada una de estas técnicas analizando sus puntos fuertes y débiles.[**Rashid2018**] El objetivo no es entrar demasiado en detalle, si no dar una guía para entender la importancia de cada unos de los algoritmos en el uso de SER.

Mel Frecuency Cepstral Coefficients (MFCC), se basa en la desintegración de la señal con la ayuda de un filtro de banco*. Es una perfecta representación para el sonido cuando la fuente es estable y consistente. Además puede capturar la información de señales sampleadas* con frecuencias a un máximo de 5 kHz lo que encapsula la mayor parte de energía proveniente del sonido que es generado por humanos, debido a esto, es frecuentemente

usada y sugerida para identificar palabras monosilábicas en un discurso; Sin embargo no reporta ser precisa cuando hay ruido de fondo y podría no ser apropiada para la generalización.

Linear Prediction Coefficients (LPC) se aplica para obtener el coeficiente de predicción lineal equivalente al tracto vocal reduciendo el mínimo error cuadrado entre la señal de audio dada como entrada y la estimada. Normalmente se usa para extraer las propiedades del tracto vocal ya que hace estimaciones bastante precisas de los parámetros en el habla, no obstante, es altamente sensible al ruido de cuantificación y al igual que MFCC podría no ser apropiado para la generalización.

Linear Prediction Cepstral Coefficient (LPCC) calcula una envolvente a LPC y luego hace una conversión a coeficientes cepstrales; Son los coeficientes de la Transformada de Fourier de la magnitud logarítmica del espectro de LPC. Tiene una baja vulnerabilidad al ruido de fondo y mejora el ratio de error en comparación a LPC, pero sigue teniendo una gran sensibilidad al ruido de cuantificación.

Discrete Wavelet Transform (DWT) La Transformada Wavelets descompone la señal en grupos de funciones básicas llamadas wavelets. La Transformada de Wavelet discreta es una extensión de esta donde mejora dicho proceso de descomposición discretizando los parámetros de tiempo y frecuencia. Los parámetros de DWT contienen información de diferentes escalas de frecuencia, lo cual es importante porque supone una mejora en la información que se obtiene del diálogo en la correspondiente banda de frecuencia. A pesar de ello, los coeficientes de Wavelet presentan variaciones indeseadas en los límites ya que las señales de entrada son de una longitud finita.

Perceptual Linear Prediction (PLP) es ligeralmente similar a MFCC pero usa preacentuación sonora igualitaria y reducción de raíz cúbica, en lugar de reducción logarítmica. Combina dos tipos de análisis el espectral y el regresión linear. El espectro es posteriormente preacentuado para aproximar la percepción auditiva irregular humana a una variedad de frecuencias,reduciendo la variación de amplitud en la resonancia espectral

¡tabla comparativa¿

1.2 Clasificación

Convencionalmente, el estudio de SER incluye el uso de diferentes tipos de clasificadores para distinguir entre emociones: Suport Vector Machines (SVNs) los cuales se han usado extensamente para el reconocimiento de emociones y pueden llegar a presentar un buen rendimiento en comparaci'on con otros clasificadores, el algoritmo K-NN es de los enfoques mas simples, Hidden Markov Model(HMM), a menudo utilizado para lidiar con los cambios temporales en la se;al y Gaussian Mixture Model (GMM) el cual es util para

representar las unidades de sonido en caracter'isticas acusticas[**Farooq2020**].

En estudios m'as recientes, se han propuesto clasificadores basados en aprendizaje profundo los cuales han resultado ser m'as eficientes adem'as de tener la capacidad de aprender las caracteristicas emocionales en el reconocimiento de emociones a traves del audio.

CNN Convolutional Neural Networks (CNN) suponen una importante contribucion en SER al utilizar caracteisticas significativas.En [] se propone un sistema basado en CNN capaz de diferenciar entre 6 emociones basicas el cual no necesita preprocesar la se;al y sigue siendo computacionalmente eficiente.

DCNN, RNN Recurren Neural Network (RNN) son convenientes por ANN y LSTM