Chapter 1

Estado del Arte

1.1 Contexto

El objetivo del reconocimiento de emociones en el habla (a partir de ahora, referido como SER, Speech Emotion Recognition por sus siglas en inglés) es reconocer el trasfondo emocional del mensaje a través de la voz. Esta manifestación sonora posee factores clave para la comunicación humana que ayudan en su interacción sin alterar el contexto del mensaje. Normalmente estos estudios se llevan a cabo en un único lenguaje, lo que en el caso que nos acontece, se traduciría como el reconocimiento de emociones llevado a cabo en la lengua materna; Mientras este ejercicio puede llegar a ser intuitivo, distinguir las mismas emociones en la lengua extranjera supone un reto ya que implicaría importantes matices culturales. Por ejemplo, no sería lo mismo entender qué emociones intenta expresar un italo parlante desde el punto de vista de una persona que entiende el español (ambas lenguas latinas), que comprender las mismas emociones del discurso desde un germano hablante. Así bien, es importante definir qué idioma se está reconociendo y desde cuál, por lo que analizar las raíces lingüísticas y fonéticas de los idiomas a estudiar es esencial.

El proceso de la clasificación emocional en el lenguaje, consiste generalmente, en tres partes: Procesamiento de la señal, extracción de características y clasificación. A continuación se ofrece una revisión más detallada de cada una de estas etapas.

1.2 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 con-

tenida en la señal, mejorar el grado de similitud y/o diferenciación entre las clases.[3] Hasta ahora, por lo general hay dos enfoques principales con respecto al tipo 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 (conocidos como coeficientes cepstrales*) La mayoría de los estudios centrados en SER usan rasgos espectrales como la información extraída del tracto vocal, lo que supone obtener la información derivada del espectro de la señal de voz y se usan para modelar los patrones de entonación y frecuencia del hablante.[4]

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.[7] 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 demuestra ser precisa cuando hay ruido de fondo y podría no ser apropiada para la generalización. En [8] implementa un sistema usando redes neuronales que se centra en el uso de MFCC como extracción de características y añade un filtro de paso alto para reducir el ruido, consiguiendo una precisión de 93.38 de media. En [9] también se utiliza MFCC como método de extracción de características destacando especialmente su efectividad en SER combinada con espectogramas de MEL.

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; 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 Trans-

formada 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 de regresión linear. El espectro es posteriormente pre-acentuado para aproximar la percepción auditiva irregular humana a una variedad de frecuencias,reduciendo la variación de amplitud en la resonancia espectral.

1.3 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 comparación con otros clasificadores tradicionales, el algoritmo K-NN es de los enfoques más simples, Hidden Markov Model(HMM)es a menudo utilizado para lidiar con los cambios temporales en la señal y por último, Gaussian Mixture Model (GMM) el cual es útil para representar las unidades de sonido en características acústicas No obstante,en estudios más recientes, se han propuesto clasificadores basados en aprendizaje profundo los cuales han superado a los enfoques tradicionales resultando ser más eficientes además de tener la capacidad de aprender las características emocionales en el reconocimiento de emociones a través del audio.

RNN Recurrent Neural Network (RNN) son convenientes en tareas en las que los datos son procesados secuencialmente. Lee y Tashev [5] afirman que los sistemas basados en DNN no cubren el *efecto contextual a largo plazo* para interpretar las emociones en el diálogo y resuelven este problema presentando un modelo basado en RNN. Por otro lado, W.Lim [6] estudia el resultado de un sistema híbrido que usa CNN y RNN para clasificar emociones en una secuencia de audio, consiguiendo un 88.01 de precisión. No obstante, los modelos RNN no son suficientes para representar el espectro emocional a lo largo de una conversación debido al problema de desvanecimiento de gradientes*.

LSTM Los retos que presenta la clasificación de emociones en el habla, son comúnmente abordados a través de una red LSTM la cual es capaz de retener información de entradas anteriores en el tiempo y tener en cuenta

dependencias temporales largas, ya que cada nodo es una célula de memoria. En el trabajo citado anteriormente [9], Wang propone un modelo dual LSTM para procesar dos espectogramas mel simultáneamente, consiguiendo una precisión del 73.3. Sin embargo este tipo de modelos no suelen implementarse como enfoque único, si no combinados con otro clasificador en una arquitectura más compleja. Por ejemplo en [6] se lleva a cabo una comparación de tres arquitecturas (CNN, LSTM y CNN distribuida en el tiempo) donde LSTM (utilizada de manera aislada) es la que puntúa más bajo.

CNN La tendencia de los modelos Deep Neural Network (DNN) en este ámbito, es aprender características específicas desde varios métodos usados en el reconocimiento de emociones a través de la percepción acústica, en especial Convolutional Neural Networks (CNN) suponen una importante contribución en SER debido al uso de características significativas. En [2] describe un método que utiliza una arquitectura basada en redes convolucionales sin seleción de características para distinguir únicamente entre tres emociones en alemán (usando Berlin Database Emotional Speech, la cual contiene 800 muestras de audio (frases) etiquetadas) consiguiendo una exactitud de 96.97. En [1] se presenta un modelo de redes convolucionales que no necesita de un preprocesado de la señal para una clasificación de emociones en el idioma inglés. Este utiliza la base de datos SAVEE, la cual contiene 480 muestras que distinguen entre 6 emociones, interpretadas por hombres y mujeres angloparlantes, y obtiene finalmente un 81.63 de precisión

Bibliography

[1] Alif Bin Abdul Qayyum, Asiful Arefeen, and Celia Shahnaz. "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition". In: *2019 IEEE International Conference on Signal Processing, Information, Communication and Systems, SPICSCON 2019* November (2019), pp. 122–125. DOI: 10.1109/SPICSCON48833.2019.9065172.

[2] Pavol Harar, Radim Burget, and Malay Kishore Dutta. "Speech emotion recognition with deep learning". In: *2017 4th International Conference on Signal Processing and Integrated Networks, SPIN 2017* February 2017 (2017), pp. 137–140. DOI: 10.1109/SPIN.2017.8049931.

[3] Nele Hellbernd and Daniela Sammler. "Prosody conveys speaker's intentions: Acoustic cues for speech act perception". In: *Journal of Memory and Language* 88 (2016), pp. 70–86. ISSN: 0749596X. DOI: 10.1016/j.jml.2016.01.001.

[4] Shadi Langari, Hossein Marvi, and Morteza Zahedi. "Efficient speech emotion recognition using modified feature extraction". In: *Informatics in Medicine Unlocked* 20 (2020), p. 100424. ISSN: 23529148. DOI: 10.1016/j.imu.2020.100424. URL: https://doi.org/10.1016/j.imu.2020.100424.

[5] Jinkyu Lee and Ivan Tashev. "High-level feature representation using recurrent neural network for speech emotion recognition". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2015-January (2015), pp. 1537–1540. ISSN: 19909772.

[6] Wootaek Lim, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and Recurrent Neural Networks". In: *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016* (2017). DOI: 10.1109/APSIPA.2016.7820699.

[7] Sabur Ajibola Alim Rashid and Nahrul Khair Alang. "Some Commonly Used Speech Feature Extraction Algorithms". In: tourism (2018), p. 13. URL: https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics.

[8] Vaibhav Kumar Sarkania and Vinod Kumar Bhalla. "Emotion Recognition through Speech Using Neural Network". In: *Android Internals* 3.6 (2015), pp. 143–147.

[9] Jianyou Wang et al. "Speech emotion recognition with dual-sequence LSTM architecture". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 2020-May (2020), pp. 6474–6478. ISSN: 15206149. DOI: 10.1109/ICASSP40776.2020.9054629. arXiv: 1910.08874.