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 habitualmente, en tres partes: Procesamiento de la señal, que aplica filtros de audio a la señal original; Extracción de características, el cual es un punto esencial en esta modalidad ya que necesita enfatizar el contenido emocional sin depender del contexto lingüístico; Y por último, la clasificación, que será la encargada de mapear el conjunto de características extraídas anteriormente con las emociones etiquetadas que hayamos definido. Seguidamente se ofrece una revisión más detallada 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 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 y 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:
Los rasgos prosódicos, los cuales extraen información de la prosodia, en concreto, tono, energía y duración, y por otro lado, las 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 la 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 y DWT. A continuación se ofrece una breve explicación de cada una de estas técnicas analizando sus puntos fuertes y débiles.[8] El objetivo no es entrar demasiado en detalle, si no dar una guía para entender la importancia de cada unos de los algoritmos para la extracción de características en el uso de SER.

Los **Coeficientes Cepstrales en la escala de Mel** (MFCC, *Mel Frecuency Cepstral Coefficients*), se basa en la desintegración de la señal para tener como resultado un resumen de las características que la forman. La obtención de este conjunto de valores numéricos se basa por un lado, en el rango de frecuencias de Mel, el cual consiste en una adaptación de frecuencias de la señal a aquellas más fácilmente percibidas por el oído humano, y por lo otro lado, la separación de frecuencias mediante *Cepstrum* que divide la señal en dos bandas de frecuencias, baja (correspondientes a los fonemas producidos por el tracto vocal) y alta (correspondientes a la excitación de las cuerdas vocales). Debido a esto, encapsula la mayor parte de energía proveniente del sonido que es generado por humanos, por lo que es frecuentemente usada y sugerida para identificar palabras monosilábicas en un discurso.

Los objetivos clave, serían:

- Eliminar la excitación del tracto vocal.
- Hacer independientes a las características extraídas.
- Ajuste a como los humanos percibimos el ruido y la frecuencia del

sonido.

- Capturar la dinámica fonética, que definirá el contexto.

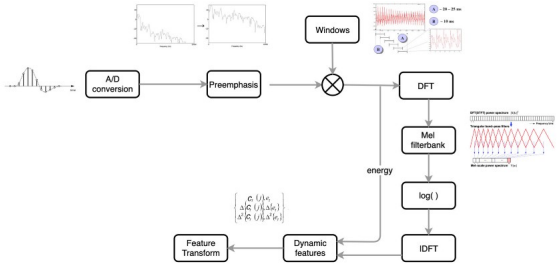

Figure 1.1: Proceso en la aplicación de MFCC en una señal.

MFCC constituye una perfecta representación para el sonido cuando la fuente es estable y consistente. En [9] 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 [12] también se utiliza MFCC como método de extracción de características destacando especialmente su efectividad en SER combinada con espectogramas de Mel.

Los **Coeficientes de Predicción Lineal** (LPC, *Linear Prediction Coefficients*) se aplican 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, por lo que demuestra no ser precisa cuando hay ruido de fondo y podría, y al igual que MFCC, no ser apropiada para la generalización.

Los **Coeficientes Cepstrales con Predicción Lineal** (LPCC, *Linear Prediction Cepstral Coefficient*) 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.

La **Transformada Wavelets** (DWT, *Discrete Wavelet Transform*) 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.

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 y en redes neuronales densas (DNN, *Deep Neural Network*) 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.

1.3.1 RNN

Las Redes Recurrentes Neuronales (RNN, *Recurrent Neural Network*) son convenientes en tareas en las que los datos son procesados secuencialmente. Lee y Tashev [5] afirman que los sistemas basados en simples redes neuronales densas no cubren el efecto contextual a largo plazo para interpretar las emociones en el diálogo (esto es, la necesidad de un contexto emocional previo en la prosodia para mejorar la clasificación en el tramo que se está actualmente analizando) 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*.

1.3.2 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. Esto a su vez, resuelve el problema de desvanecimiento de gradientes que presenta RNN. En el trabajo citado anteriormente [12], 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.

1.3.3 CNN

La tendencia de los modelos 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 las redes neuronales convolucionales (CNN, *Convolutional Neural Networks*) suponen una importante contribución en SER debido al uso de características significativas, y su uso en recientes estudios se ha incrementado a lo largo de los años. En [2] describe un método que utiliza una arquitectura basada en redes convolucionales sin selección de características para distinguir únicamente entre tres emociones en alemán (usando Berlin Database Emotional Speech, la cual contiene 800 muestras de audio 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, donde obtiene finalmente un 81.63% de precisión.

1.4 Valoración y Propuesta

En esta sección se valorarán las conclusiones que se extraigan del análisis previo sobre los distintas etapas correspondientes al desarrollo de un modelo para la clasificación de emociones en el habla. Los retos que plantea SER se han abordado anteriormente desde distintos enfoques, pero en su mayoría, para un único lenguaje. Las dificultades que presenta el reconocimiento de emociones en la lengua extranjera se deben principalmente a [...]

Los trabajos de Pell se han centrado durante años en el análisis de la prosodia a través de los idiomas. A pesar de su antigüedad y que no entra demasiado en detalles técnicos, merece la pena mencionar que en [7] lleva a cabo un estudio comparativo entre la detección emocional de la prosodia en la lengua materna y la extranjera, concluyendo que el proceso para entender

las emociones vocales en una lengua no aprendida, implica una mayor exposición a esta para familiarizarse con señales prosódicas correspondientes a significados subyacentes.

En la extracción de características el uso de MFCC ha sido ampliamente escogido al reportar resultados más elevados en comparación con otros métodos, mientras que en la sección donde se discuten los diversons métodos de clasificación usados en SER, vemos que CNN tiene mayor preferencia en la literatura, sin despreciar el uso de modelos basados en una combinación entre CNN y LSTM.
El uso de CNN se ha vuelto cada vez más popular, especialmente en el procesamiento de imágenes, debido a que la principal ventaja que ofrecen en comparación con otras arquitecturas, es la habilidad de detectar características relevantes sin supervisión, a la vez que son computacionalmente eficientes. Ahora bien, el problema en el que se desarrolla en este trabajo gira en torno al procesamiento del audio, entonces, ¿cómo se pueden aprovechar las ventajas que provee CNN en este ámbito?

No olvidemos que el tipo de dato con el que se trabajará principalmente serán señales, concretamente, de audio. En las señales de audio hay una cierta presión de aire que varía con respecto al tiempo, y al muestrearlas en un determinado rango de frecuencia, obtendríamos algo como lo siguiente:

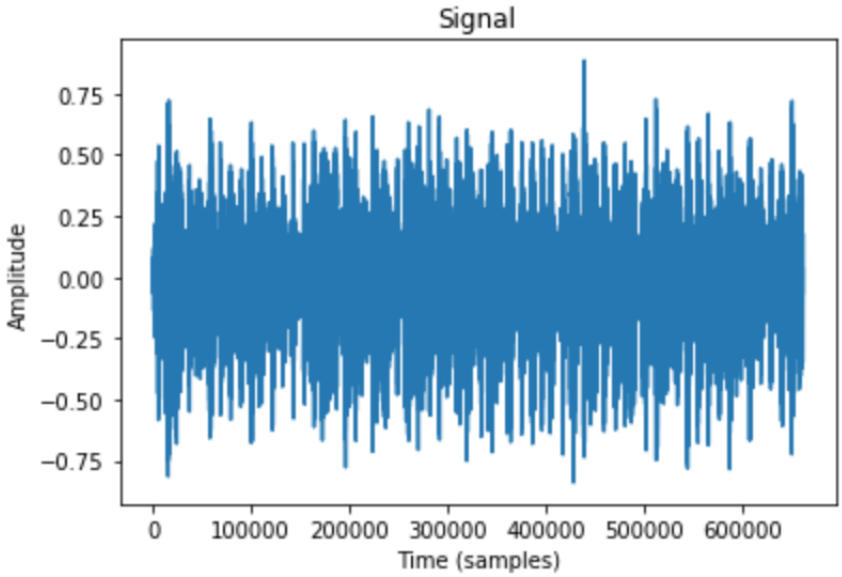

Figure 1.2: waveform de una señal de audio

Esto que vemos es una representación digital de la onda, también llamada *waveform*, de manera que ahora puede ser interpretada y analizada

fácilmente.
El siguiente punto que deberíamos atender, sería preguntarnos cómo extraer características relevantes de esta representación, y para ello se hace obvio pensar en la Transformada de Fourier (FFT *Fast Fourier Transform*), la cual nos permite analizar la cantidad de frecuencia contenida en una señal. La Transformada de Fourier transforma la señal de un dominio de tiempo a un dominio de frencuencia, y el resultado de esta transformación se denomina espectro.
Sin embargo el problema viene cuando en las señales de audio (que son el tipo de señales con las que queremos trabajar), la cantidad de frecuencia varía en el tiempo, por lo que FFT es insuficiente al no poder representar en el espectro resultante esta variación de la señal en el tiempo. La Transformada de Fourier en tiempo reducido, (*short-time Fast Fourier Transform*) resuelve este problema calculando la FFT en segmentos (ventanas de tiempo) superpuestos de la señal. Lo que finalmente obtenemos se denomina **espectograma**.

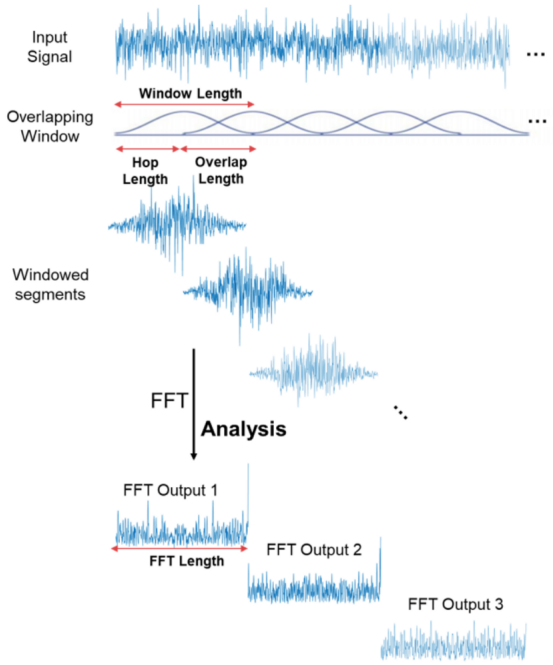

Figure 1.3: Proceso del uso de FFT en una señal. Fuente: Mathworks

Este espectograma puede ser entendido como una representación tridimensional de la señal donde sus características (tiempo, frecuencia y amplitud de la distribución de energía) pueden ser observadas de manera muy visual. Cuando este espectograma se computa, el eje X representa el tiempo, el eje Y representa la frecuencia, que es convertida a una escala logarítmica,

y la gama de colores que se utiliza es para simbolizar la variación de energía expresada (medida decibelios), donde los tonos más oscuros indican unos valores de energía más altos, y viceversa.

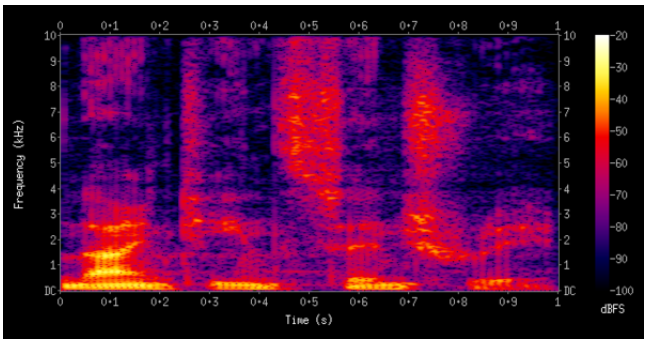

Figure 1.4: Espectograma de una muestra aleatoria. Fuente: Wiki

La respuesta a por qué las frecuencias son convertidas a una escala logarítmica es sencillamente porque los humanos no percibimos las frecuencias en una escala lineal [11], es decir, nuestra habilidad para distinguir entre frecuencias fluctúa a lo largo del rango en el que somos capaces de percibir[10]. Es por ello que el rango donde se mueve nuestro espectograma, se adapta a lo que se llama **escala de Mel**, en la cual los armónicos se observan equidistantes, reduciendo como resultado las variantes acústicas que no son significativas.

Finalmente nos queda entender el concepto de *Cepstrum* o coeficientes cepstrales, y para ello debemos entender como el sonido (en el caso de la articulación de palabras) es producido. En el modelo de la figura se representa el discurso como la combinación de las vibraciones producidas por las cuerdas vocales con las del tracto vocal, y nuestras articulaciones controlan la forma del tracto vocal. La forma de onda de la voz será reprimida o amplificada a diferentes frecuencias por la forma de nuestro tracto vocal.
El papel de Cepstrum es la separación de frecuencias atendiendo a como los sonidos son producidos siguiendo un modelo anatómico, en el algoritmo de MFCC, de manera que cuando es computado separa la señal de voz y la resonancia del tracto vocal. En 1.1 observamos el paso IDFT, donde tras ajustar la señal a la escala de Mel, se calcula una variante de FFT (*Inverse Discrete Fourier Transform*) y se obtienen los **coeficientes de MFCC**.

Dado que hemos convertido una señal de audio en una imagen, ahora se podrá proceder a usar un modelo basado en redes convolucionales, aprovechando sus ventajas en el campo donde mejor rendimiento reporta: el procesamiento de imágenes.

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] Marc D. Pell and Vera Skorup. "Implicit processing of emotional prosody in a foreign versus native language". In: *Speech Communication* 50.6 (2008), pp. 519–530. ISSN: 01676393. DOI: 10.1016/j.specom.2008.03.006.

[8] 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.

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

[10] S. S. Stevens and J. Volkmann. "The Relation of Pitch to Frequency: A Revised Scale". In: *The American Journal of Psychology* 53.3 (1940), pp. 329–353. ISSN: 00029556. URL: http://www.jstor.org/stable/1417526.

[11] Lav Varshney and John Sun. "Why do we perceive logarithmically?" In: *Significance* 10 (Feb. 2013). DOI: 10.1111/j.1740-9713.2013.00636.x.

[12] 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.