Universidad Internacional de la Rioja (UNIR)

ESIT

Máster Universitario en Inteligencia Artificial

Reconocimiento de Emociones en la Lengua no Aprendida

Trabajo Fin de Máster

presentado por: Luisa Sánchez Avivar

Dirigido por: Ciro Rodríguez León

Ciudad:Ginebra

Fecha: Julio del 2021

Índice de Contenidos

Índice de Ilustraciones

Índice de Tablas

Resumen

En este estudio se llevó a cabo un reconocimiento emocional de la voz multi-lingüístico. Para ello, se implementaron tres modelos distintos entrenados en inglés, y posteriormente fueron evaluados en dos lenguas extranjeras que no formaron parte del entrenamiento (francés y alemán). Las características cepstrales de la escala de Mel se extrajeron a partir de las muestras de audio y fueron usadas en los tres clasificadores con una arquitectura basada en redes convolucionales. El uso de espectrogramas en una arquitectura híbrida de redes convolucionales y LSTM se mostró superior frente a los otros, consiguiendo un 92.06 % de exactitud en una clasificación monolingu?ística. Por otro lado, la clasificación multi-lingüística no arrojó resultados satisfactorios aplicando el mismo método.

Palabras Clave: CNN-LSTM, Reconocimiento de emociones en el habla, características espectrales, Lengua extranjera.

Abstract

This work performs a speech emotional recognition through three languages. For this purpose, three different models have been implemented and trained in english, and subsequently tested in other two languages which never took part in the training (french and german). It is assumed that speech audio signals carry emotional information that can be retrieved and hence MFCC features are extracted since they are recognized as best suited to represent emotions through prosody. Different classifiers based on convolutional neural network (CNN) architecture were used (unidimensional CNN, bidimensional CNN and LSTM-CNN). The results show that CNN-LSTM outperforms over the other options with a 92.06 % of accuracy in a monolinguistic clasification in english, while appliying the same approach in a cross language classification did not deliver satisfactory results.

Keywords: CNN-LSTM, foreign language, MFCC, spectral features, Speech Emotion Recognition.

1. Introducción

La importancia de la comunicación no sólo se basa en qué se dice, sino también en cómo se dice. Desde hace años, el reconocimiento de emociones a través de la voz ha sido motivo de interés para la investigación, sin embargo siempre se ha estudiado sobre un mismo lenguaje debatiendo la habilidad de reconocer y clasificar las emociones oralmente expresadas. Esta habilidad ha sido respaldada por numerosos artículos donde se concluye que es posible distinguir e identificar entre al menos tres emociones básicas (Felicidad, Tristeza, y Enfado) a través de la voz (sin necesidad del procesamiento del lenguaje natural y por lo tanto de un contexto).

Análogamente, el debate del reconocimiento de emociones en un plano intercultural también se ha enfocado a través del estudio de los gestos faciales en conjunto con la expresión vocal, donde se concluye que los factores sociales tienen un gran impacto, ya que la identificación de las emociones es más fácil para los miembros de la misma cultura que para los de otra distinta (Pell, Monetta y col., 2009; Pell, Paulmann y col., 2009). A pesar de ello hay una gran carencia de comparativas con respecto a la voz donde se demuestre una sólida influencia cultural, sin embargo parece claro que las dimensiones socio culturales que engloban nuestras interacciones pueden tener un gran impacto en nuestra comunicación dentro de un marco emocional.

1.1. Motivación

El espectro emocional que una persona esconde en su discurso es un factor esencial de la comunicación humana y ofrece información adicional sin alterar el contenido lingüístico. Las tecnologías orientadas a convertir la voz en texto (*speech to text*) no tienen una forma segura de medir la calidad del diálogo de su interlocutor, impactando en negocios que hacen uso de estos avances (por ejemplo, centros de atención al cliente donde miden su grado de satisfacción). La compañía OTO, dedicada a creación de sistemas inteligentes ("Introducing OTO Systems Inc." 2018) centrados en la decodificación de la voz, reportó que hay al menos 3000 formas de decir "Gracias"(figura 1.1).

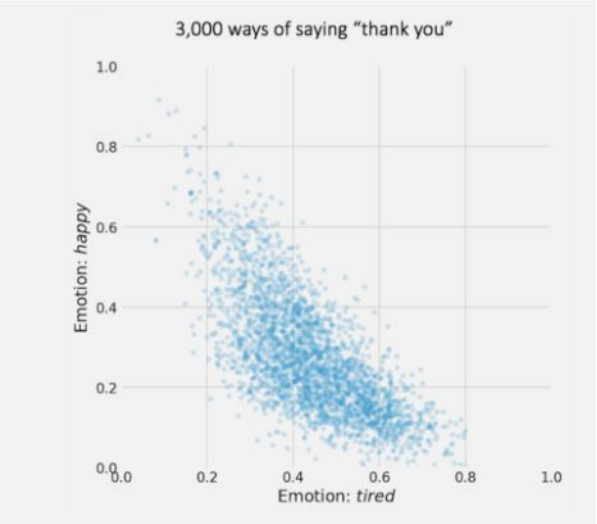


Figura 1.1: Dimensiones acústicas en la forma de decir "Thank you". Fuente: OTO Systems

Es indudable el impacto que ha creado la inteligencia artificial en la forma en la que nos comunicamos con las máquinas a día de hoy (Lisetti, 1998). La importancia de la interacción con las máquinas a través de comandos de voz se ha visto acentuada gracias a la aparición de asistentes inteligentes como Siri en Apple (Effron, 2011) o Alexa en Amazon (Arnold, 2017) , que han explotado las diferentes áreas del análisis de la voz con el objetivo de mejorar la experiencia de usuario. Otras compañías como OTO han desarrollado modelos de análisis de voz capaces de detectar atributos únicos en la voz del interlocutor, lo que es usando por centros de asistencia telefónica para potenciar y mejorar sus sistemas automáticos (OTO-Systems, 2020). En definitiva, desde el primer software de reconocimiento por voz que fue presentado por IBM en 1961 reconociendo dieciséis palabras y dígitos (IBM, 2020), hasta la aparición de Google Home en 2017 (Hautala, 2015), este tipo de asistentes han ido mejorando su alcance y capacidades.

El uso de estos asistentes no sólo se limita al ámbito doméstico, ya que esta tendencia empieza a extenderse hasta en la aplicación de asistentes personalizados para coches automáticos y asistencia de ayuda telefónica en la industria médica. Por ejemplo, el uso de altavoces inteligentes han demostrado su efectividad en contrarrestar los efectos de la soledad, el aislamiento y la depresión en personas de la tercera edad al ser usadas en residencias (Mizak y col., 2017). Novel Effect integró sus sistemas con Alexa (Amazon) para crear una tecnología que ofrece una lectura interactiva con fines educativos incrementando la retención de contenidos y su comprensión (Rowe, 2018). De la misma manera, com-

pañías como Capital One en la industria financiera, o KAYAK en el sector turístico han integrado también sus sistemas con Alexa para dar soporte a sus clientes (Collier, 2016) y (Jet, 2017).

Sin embargo, a pesar de los avances tecnológicos, estos asistentes de voz normalmente carecen de la habilidad de reconocer el estado emocional del usuario, y cerrar esta brecha podría ser un gran avance en las industrias ya mencionadas.

Cabe pensar que en este tipo de tecnología haya un potencial interés para la asistencia sanitaria, o incluso, para la industria automovilística. Visualicemos por ejemplo, un conductor tratando de resolver una incidencia mientras conduce. Esta incidencia puede variar desde buscar una ruta alternativa a un hospital o servicio de emergencia, y el estado emocional en el que se encuentre, puede afectar limitando su habilidad para resolver el problema.

De la misma manera el reconocimiento de emociones puede ocupar un lugar en los asistentes virtuales de cualquier servicio al integrarlo con técnicas del procesamiento del lenguaje natural, permitiendo mayor eficiencia del procesamiento de la conversación al detectar -por ejemplo- irritabilidad o frustración en el usuario. SRI Ventures, una central estadounidense centrada en técnicas de procesamiento de la voz para el desarrollo de aplicaciones principalmente en el ámbito de la salud (Nuance, 2019) y (Mocherman, 2015) desarrolla tecnología para analizar síntomas relacionados con enfermedades respiratorias, así como la aplicación de inteligencia artificial para evaluar por voz los sentimientos del cliente con el fin de mejorar el servicio.

Otro equipo de SRI liderado por Elizabeth Shriberg (Shriberg y col., 2018) decidieron combinar un sistema para entender la expresión oral con el análisis de emociones en la voz, dando como resultado una tecnología capaz de modelar computacionalmente la entonación de la voz del interlocutor derivando el significado sentimental más allá de las palabras usadas.

Estos ejemplos, impulsan la motivación de crear un sistema capaz de crear una respuesta, no sólo coherente en el plano semántico, sino también sensible al estado emocional del usuario.

1.2. Planteamiento del Trabajo

Este trabajo de fin de máster se centra en el uso de técnicas basadas en redes artificiales para la clasificación de emociones a través del discurso en la lengua extranjera. Para

acercarnos a este escenario, se parte del supuesto que dado un modelo entrenado en un lenguaje capaz de reconocer emociones en este, se evalúa en un idioma distinto que nunca ha formado parte del anterior conjunto de datos.

Para ello, el trabajo se divide en dos objetivos principales: el primero será conseguir un modelo lo suficientemente preciso en una sola lengua, y el segundo comparar su comportamiento con otros lenguajes.

Con respecto a esta disciplina, y como ya se verá en el capítulo 2, existen estudios (en el reconocimiento de emociones en una sola lengua) combinando numerosas variables a tener en cuenta, por lo que se probarán distintos enfoques analizando sus fortalezas y debilidades.

Una de las limitaciones a las que se hace frente es el reducido número de muestras en las distintas bases de datos disponibles para entrenar el modelo. Teniendo este punto en cuenta y que el objetivo es incrementar las oportunidades de clasificación en otros conjuntos, este problema se resolverá combinando datos de varias bases de datos distintas, lo que aportará variabilidad a los datos de entrenamiento. Por otro lado, debido a que la mayoría de trabajos encontrados que realizan esta clasificación emocional son en una sola lengua, se parte de un enfoque más ingenuo donde se asume que las señales de audio transportan suficiente información que puede ser extraída.

La evaluación del modelo resultante en otros idiomas no será de manera arbitraria, sino que, dentro de los conjuntos disponibles que se encuentren, seguirá una estrategia atendiendo al grado de proximidad fonética para ver las diferencias.

Con este estudio se pretende entender mejor la relación entre emoción e idioma y arrojar luz a preguntas como ¿Hay emociones que son más fácilmente reconocibles indistintamente del lenguaje? ¿Hay lenguas donde es más fácil reconocer ciertas emociones? ¿Cómo influye la elección de la base de datos?¿Plantean las técnicas más populares un enfoque adecuado?.

1.3. Estructura del trabajo

1. Se introduce de manera general, el contenido de las distintas partes que componen este trabajo. El primer capítulo introduce de manera esquemática el tema principal de este trabajo, justificando su importancia y el impacto en el mundo real. En este mismo capítulo se incluye:
    - La motivación, donde se argumenta la relevancia de este trabajo.

- El planteamiento del trabajo, que propone de manera general cómo solucionar el problema que se va a encarar.

2. Seguidamente, en el Contexto y estado del arte se realizará una revisión de la literatura actual en relación con el tema a tratar. Se analizarán los resultados conseguidos hasta el momento así como las técnicas y métodos más usados en este ámbito, en concreto tendrá la siguiente estructura:

   - Se presentan las características fonéticas del habla y su trasfondo teórico, analizando las posibles limitaciones, o características a tener en cuenta a modo de introducción a las siguientes secciones que son más técnicas.
   - El reconocimiento emocional del discurso propone una forma de modelar el problema haciendo uso de las técnicas que expone.
   - La extracción de características debate sobre qué características encontrar en la voz para reconocer emociones y cómo se pueden extraer.
   - El preprocesado de la señal describe el proceso para convertir la señal de audio a imagen de manera que se aprovechen mejor las ventajas de los clasificadores basados en redes convolucionales.
   - Los algoritmos de clasificación exponen diferentes métodos para categorizar las emociones una vez la señal está procesada.
   - La discusión sobre el estado del arte, revisa otros trabajos donde se han aplicado estas técnicas y cuáles han sido sus resultados.

3. La metodología de trabajo, comprende los objetivos generales y específicos que se han marcado para este trabajo así como el proceso que se seguirá para llevarlo a cabo.

4. El capítulo 4, plantea diferentes experimentos con el objetivo de llegar a una conclusión en la comparativa de esta tesis. Describe todos los componentes que formarán parte de esos experimentos de manera que se puedan reproducir siguiendo los pasos propuestos.

5. El capítulo 5, describe los resultados y el desarrollo de los experimentos planteados en el capítulo anterior, exponiéndolos de manera objetiva.

6. En el capítulo 6 se presenta un análisis de los resultados obtenidos.

7. Finalmente en el capítulo 7 tiene lugar una discusión de la conclusión sobre los resultados, comparándolos con los de otros trabajos, a la vez que se describen posibles líneas futuras

2. Estado del Arte

2.1. Contexto

El reconocimiento de emociones en el habla es una disciplina en inteligencia artificial que trata de reconocer y clasificar emociones a través de la señal de voz. Este campo de estudio se ha hecho cada vez más popular, pero su origen se remonta a 1996, desde que se presentara el primer trabajo defendible sobre el tema Reconociendo emociones en el habla"por F.Daellert (Dellaert y col., 1996). Desde entonces, el reconocimiento de emociones a través de la voz ha sido motivo de interés para la investigación, sin embargo, en su gran mayoría se ha estudiado sobre un mismo lenguaje debatiendo la habilidad de reconocer y clasificar las emociones oralmente expresadas.

El presente capítulo está dividido en dos partes: en la primera se hablará de los conceptos inherentes a las características fonéticas en el discurso, que es el tema en torno al que gira este trabajo. En la segunda parte, cómo se puede orientar ese problema con técnicas de aprendizaje profundo haciendo una revisión de la literatura.

2.1.1. Características fonéticas en el habla

El objetivo del reconocimiento de emociones en el habla 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.

La expresión de las emociones están íntimamente relacionadas con las propiedades fonéticas en el habla donde se observan señales y patrones para marcar contrastes lingüísticos en un idioma (Pell, 2001), por lo tanto, los efectos del lenguaje en la comunicación emocional son evidentes al haber sido observadas y medidas las variaciones en el rango tonal y la frecuencia para expresarlas, cambiando no sólo el tono sino también el patrón lingüístico asociado (Davletcharova y col., 2015). Por otro lado tanto la proporción de consonantes y vocales (que hacen variar la presión de aire que se necesita), como el ratio de sílabas por palabra en cada idioma, caracterizan la expresión oral de las emociones. Existen muchos

factores relacionados con el lenguaje como la morfología o la duración del estímulo que podrían tener un impacto en la decodificación de los matices en la señal vocal, tal y como se explica en (Chen y col., 2017). Existe una clasificación dependiendo de la velocidad silábica en la expresión de dichos idiomas, sin embargo poco se conoce acerca de los efectos en las medidas respiratorias en el habla (X. Huang, A. Acero & Hon, 2001). Esta observación puede llevar a que se pregunte si en lenguajes que son muy distintos fonéticamente, las emociones expresadas mediante la voz puedan ser reconocidas desde el punto de vista del otro idioma.

Normalmente estos estudios se llevan a cabo en un único lenguaje, lo que para este trabajo 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.

2.1.2. Reconocimiento emocional del discurso

Atendiendo a la manera en cómo se modelan las emociones de manera que esta información pueda ser extraída a partir de señales acústicas (Kumar & Iqbal, 2019), se parte de estudios en psicología donde la clasificación de emociones se ha tratado desde dos enfoques principales:

- Las emociones como categorías discretas.
- Las emociones vistas a través de un modelo dimensional.

En el primer punto, a todos los humanos se les atribuye un conjunto de emociones básicas que pueden ser reconocidas interculturalmente. El debate se centra en la definición de dichas emociones, y fue Paul Elkman y su equipo en 1992 (Ekman, 1992) quien estableció que estas eran seis: Enfado, Asco, Miedo , Felicidad, Tristeza y Sorpresa. En el segundo punto, las emociones se definen respecto a una o más dimensiones, donde normalmente las dimensiones que se comprenden tienden a ser la afectividad, la excitación o la intensidad.

Aquí la discusión se centra en encontrar el número de dimensiones que dé lugar a un modelo coherente y pueda incluir las emociones conocidas. El modelo de Plutchik (Plutchik, 2001) sea quizá el más conocido en este enfoque, y propone un modelo tridimensional que organiza las emociones en círculos concéntricos situando la más básicas en el centro como se representa en la figura 2.1.

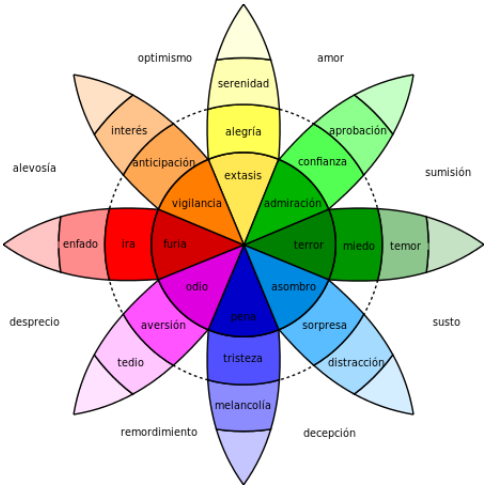


Figura 2.1: Rueda de las Emociones de Plutchik. Fuente: Wikipedia

Sin embargo, por cuestiones de simplicidad, la mayoría de los estudios centrados en la clasificación de emociones se centran en las emociones vistas como categorías discretas.

2.1.2.1. Extracción de Características

La extracción de características es una de las secciones más importantes en el reconocimiento de emociones a través de la voz debido a la ambigüedad de las características y la variedad vocal. La extracción de características es el paso principal en el procesamiento del diálogo, 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 (Hellbernd & Sammler, 2016). Hasta ahora, por lo general hay dos enfoques principales con respecto al tipo de características usadas en el Reconocimiento de Emociones en el Discurso:
Por un lado, 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 este tema 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 (Langari y col., 2020).

En (Rashid & Alang, 2018) se ofrece una breve explicación de cada una de las técnicas más comunes analizando sus puntos fuertes y débiles: Así pues, se puede encontrar la Transformada Wavelets Discreta (DWT), que a pesar de mejorar la información que se obtiene del diálogo en la correspondiente banda de frecuencia, presenta variaciones indeseadas en los límites debido a que las señales de entrada son de una longitud finita. También se encuentran trabajos donde se usan Coeficientes de Predicción Lineal (LPC) (Rana & Miglani, 2014) y (Ram y col., 2013) los cuales hacen estimaciones bastante precisas al extraer las propiedades del tracto vocal, pero son altamente sensibles al ruido de cuantificación, por lo que demuestran no ser precisos cuando hay ruido de fondo.

No obstante, a lo largo de los últimos años se ha popularizado el uso de otros métodos reportando mejores resultados. Estos son:

Coeficientes Cepstrales con Predicción Lineal (LPCC) Calcula una envolvente a los Coeficientes de Predicción Lineal (LPC) y luego hace una conversión a coeficientes cepstrales; Esto materializa las características de un canal particular del sonido, teniendo en cuenta que la misma persona con diferentes tonos emocionales tendrá diferentes canales de características, se podrán extraer esos coeficientes para identificar las emociones contenidas (Sandesara y col., 2020). Tiene una baja vulnerabilidad al ruido de fondo y mejora el ratio de error en comparación con LPC, pero sigue teniendo una gran sensibilidad al ruido de cuantificación.

Coeficientes Cepstrales en la escala de Mel (MFCC) Es la representación compacta del espectro de una señal de audio. MFCC 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 una parte, 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 otra, la separación de frecuencias mediante cepstrales (*Cepstrum*, es el resultado de calcular la transformada de Fourier inversa del

espectro de la señal estudiada en escala logarítmica (D. G. Childers, 1977)), 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) (Davis & Mermelstein, 1980) .
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 (Farouk, 2014).

Enumerando los objetivos clave del proceso serían:

1. Divide la señal en segmentos de tiempo cortos. Ya que la frecuencia de la señal cambia en la línea temporal, no tendría sentido aplicar la Transformada de Fourier en toda la señal, puesto que se perdería parte de la frecuencia a lo largo de ese tiempo produciendo distorsiones.

2. Pasar la señal del dominio de tiempo a dominio de frecuencia. Ya que FFT (*Fast Fourier Transform*) asume que la señal de audio es periódica y continua, al fragmentar la señal se garantiza que es periódica, y para la continuidad se aplica la función de ventana de Hann (Smith, 2011). Si no se hiciese este paso se producirían distorsiones en las frecuencias más altas.

3. Aplicación de un filtro de banco para ajustar la señal a la forma en la que los humanos percibimos el sonido y su frecuencia. Concretamente se aplica la escala de Mel, de donde se extraerá la energía en cada banda de frecuencia (Fayek, 2016).

4. Finalmente se aplica la Transformada Discreta del Coseno (comúnmente llamada DCT por sus siglas en inglés *Discrete Cosine Transform*) para generar los coeficientes (Khayam, 2003). Los coeficientes cepstrales de la escala de Mel contienen información sobre los cambios en las diferentes bandas del espectro, así que DCT extraerá esos cambios en las altas y bajas frecuencias de la señal.

	emotion	0	1	2	3	4	5	6	7	8	9	10
753	sad	-299.187337	15.153487	12.559813	8.849728	6.323772	3.134192	-2.395739	-4.958667	-4.889857	-3.971155	-4.065112
1902	disgust	-254.086582	26.924485	-1.261986	3.727236	-4.586822	-1.353001	-4.952505	-8.846091	-6.021211	-0.729188	-5.172279
4350	disgust	-455.742798	86.338509	11.563093	7.967887	4.676503	11.751726	-18.530741	8.897447	-22.486078	0.983860	-10.174376
2485	disgust	-275.853270	32.661419	16.059203	6.270795	-0.364384	-4.201418	-7.061556	-4.365420	-5.474218	-4.132843	-3.606962
482	happy	-347.109253	86.740524	6.508771	22.190424	15.010102	-0.541927	-17.464493	4.609902	-0.187708	-10.953825	-0.551896

Figura 2.2: Coeficientes MFCC. Fuente: propia

Los valores que se muestran en la figura 2.2 son 10 de los coeficientes resultantes. Cuando un valor es positivo, significa que la mayor parte de la energía espectral está concentrada en las regiones de frecuencia baja. Por el contrario, si el valor es negativo, la mayor parte de la energía espectral está concentrada en frecuencias altas.
Se considera que de 12 a 20 coeficientes cepstrales es una cantidad óptima para el análisis de la voz (Poorjam, 2018).

2.1.2.2. Procesamiento de la señal como imagen

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 (Koolagudi & Rao, 2012), y al muestrearlas en un determinado rango de frecuencia, se obtendría algo como lo siguiente:

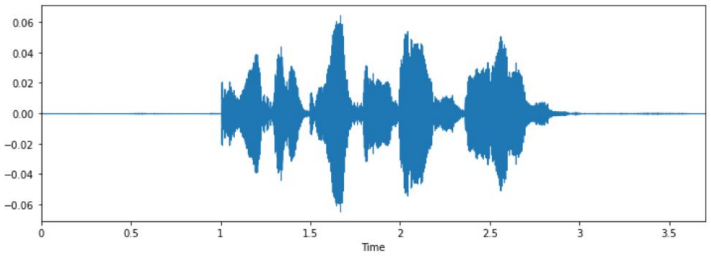


Figura 2.3: Forma de onda de una señal de audio. Fuente: propia

Los que se muestra en la figura 2.3 es una representación digital de la onda, de manera que ahora puede ser interpretada y analizada fácilmente.
La Transformada de Fourier (FFT *Fast Fourier Transform*) responde a cómo extraer características relevantes de esta representación, ya que permite analizar la cantidad de frecuencia contenida en una señal. Ésta transforma la señal de un dominio de tiempo a un dominio de frencuencia (figura 2.4), y el resultado es el espectro ("Fast Fourier Transformation FFT - Basics", 2016).

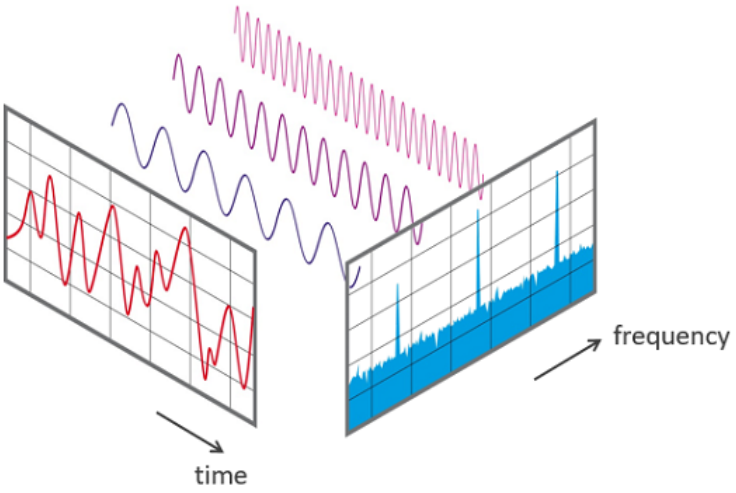
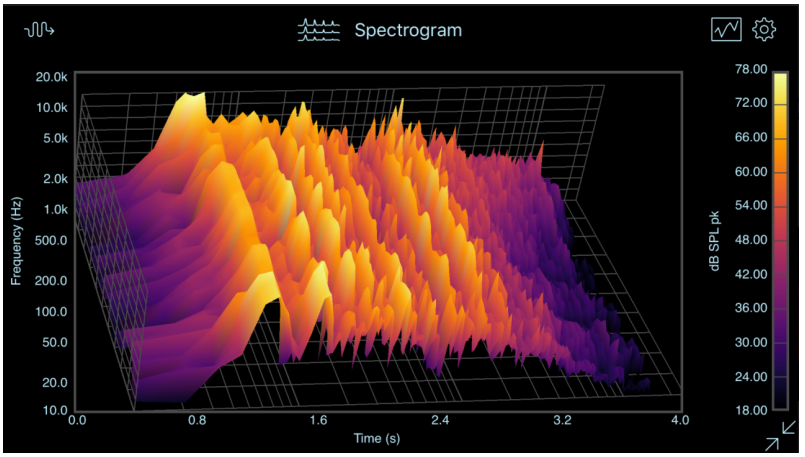


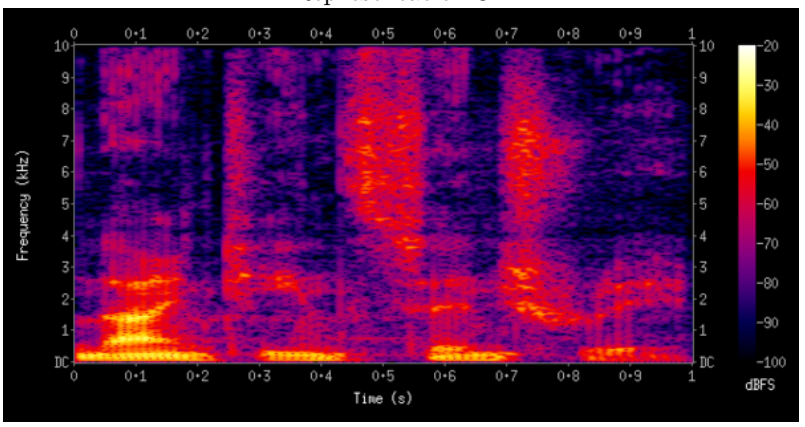
Figura 2.4: Representación de una señal desde dos planos (frecuencia y tiempo). Fuente: Wikipedia

Sin embargo el problema viene cuando en las señales de audio, 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. Tras aplicar el filtro de banco en la escapa Mel mencionado anteriormente, se obtiene finalmente lo que se denomina **espectrograma**.

Este espectrograma (figura 2.5) puede ser entendido como una representación tridimensional de la señal donde sus características (tiempo, frecuencia distribución de energía) pueden ser observadas de manera muy visual. Cuando este espectrograma 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(Kartik, 2020).



Representación 3D



Espectrograma resultante

Figura 2.5: Espectrograma de una muestra aleatoria. Fuente: faberAcoustical

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 (Varshney & Sun, 2013), es decir, nuestra habilidad para distinguir entre frecuencias fluctúa a lo largo del rango en el que somos capaces de percibir. Es por ello, que el rango donde se mueve este espectrograma se adapta a la **escala de Mel**, en la cual los armónicos se observan equidistantes, reduciendo como resultado las variantes acústicas que no son significativas (Stevens & Volkmann, 1940).

Finalmente queda comprender el concepto de *Cepstrum* o coeficientes cepstrales, y para ello hay que entender cómo el sonido (respecto a la articulación de palabras) es producido. Técnicamente, esta producción del sonido en la anatomía se definiría como la combinación de las vibraciones producidas por las cuerdas vocales con las vibraciones producidas por la

resonancia del tracto vocal. Las articulaciones que una persona realiza al hablar controlan la forma del tracto vocal, por lo que la forma de onda de la voz será reprimida o amplificada a diferentes frecuencias por la forma de nuestro tracto vocal (Bao & Huang, 2019).
El papel del Cepstrum es la separación de frecuencias en el algoritmo de MFCC, atendiendo a cómo los sonidos son producidos siguiendo un modelo anatómico, de manera que cuando es computado separa la señal de voz y la resonancia del tracto vocal (Nair, 2018).
En la figura 2.6 se pueden observar el resultado de los pasos que se han discutido hasta ahora.

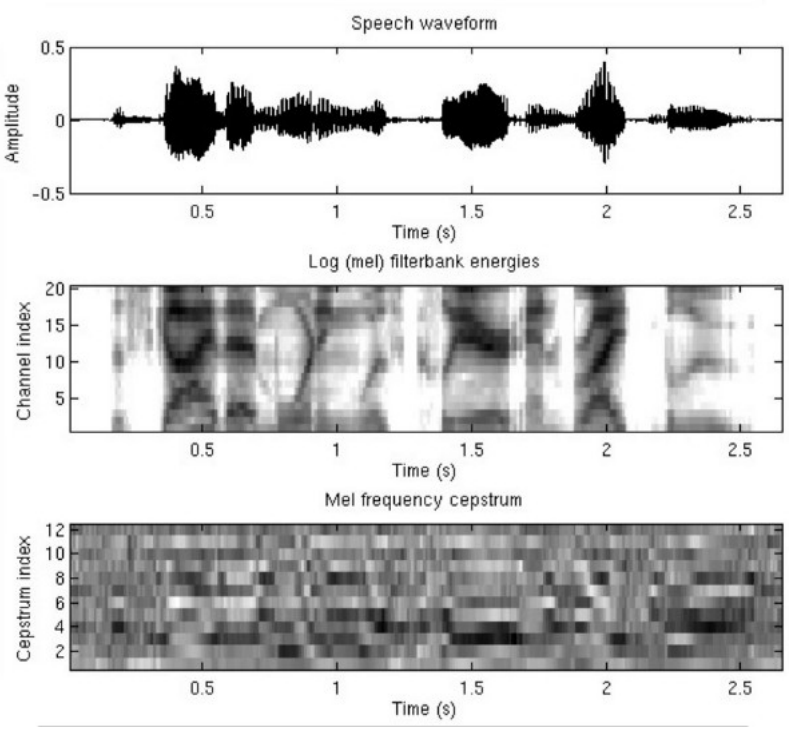


Figura 2.6: Visualización de los coeficientes cepstrales de una muestra aleatoria. Fuente: Columbia University

Dado que se ha 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.

2.1.2.3. Algoritmos de Clasificación

Convencionalmente, el estudio del Reconocimiento de Emociones en el Habla incluye el uso de diferentes tipos de clasificadores entre las que destacan las Máquinas de Vector de Soporte (SVNs *Suport Vector Machines*) ya que se han usado extensamente para el reco-

nocimiento de emociones y pueden llegar a presentar un buen rendimiento en comparación con otros clasificadores tradicionales (Africa y col., 2020) y (Jain y col., 2020).

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(Sandesara y col., 2020).

El aprendizaje profundo es un conjunto de algoritmos de aprendizaje automático que modela abstracciones de alto nivel construyendo conceptos complejos a partir de otros más sencillos mediante el uso de una arquitectura jerárquica. Esta arquitectura jerárquica es lo que se denomina redes neuronales, que son estructuras lógicas cuyo diseño se basa en mayor medida en la organización del sistema nervioso de los mamíferos (Matthew, 2019). Las neuronas artificiales que la componen son unidades de proceso especializadas en detectar determinadas características de aquello que es percibido (Nielsen, 2015).

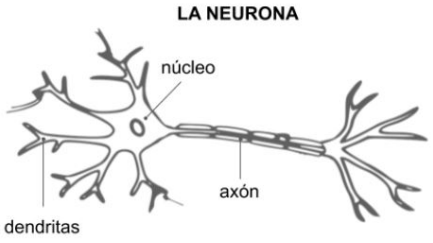


Figura 2.7: Representación de una neurona real. Fuente: CeBe

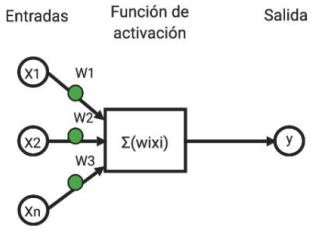


Figura 2.8: Representación de una neurona artificial. Fuente: futureLab

La imagen 2.8 muestra el funcionamiento de estas neuronas artificiales claramente inspirado en el diseño de la figura 2.7. Al nodo le llegan unas entradas x_i que tienen asignadas unos pesos w_i y unos sesgos b_i (*biases*), que son procesadas por la función de activación. El proceso de aprendizaje se consigue por transferencia de información de unas capas a otras gracias al algoritmo de retropropagación (*backpropagation*) cuyo objetivo es encontrar la distribución adecuada a cada una de las variables de entrada (Goodfellow y col., 2016).

Redes Neuronales Recurrentes

Las Redes Neuronales Recurrentes (RNN) son convenientes en tareas en las que los datos son procesados secuencialmente. Esto puede ser especialmente una ventaja ya que las distintas entradas de la señal no se tratan de manera independiente (Lim y col., 2017). En

(Lee & Tashev, 2015) se propone un sistema basado en RNN aprovechando esta particularidad donde cada nodo toma en cuenta la información recogida en los anteriores; Esto hace que cubra un espectro más amplio de información creando una especie de contexto.

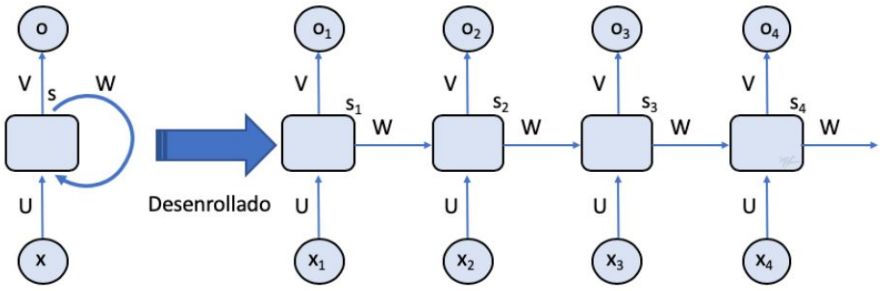


Figura 2.9: Estructura de RNN. Fuente: Cimat

La figura 2.9 muestra la estructura básica de una Red Neuronal Recurrente donde se puede observar como la red actualiza el peso de las entradas a través del algoritmo Descenso de Gradientes. En algunos casos, estos gradientes se irán haciendo más pequeños a medida que la red avanza, evitando así que los pesos cambien su valor y por lo tanto, la red siga aprendiendo (lo que se conoce como desvanecimiento de gradientes).

Redes LSTM

Como se puede ver, los trabajos anteriormente mencionados que implementan este tipo de arquitectura RNN son relativamente antiguos (2017, y 2015 respectivamente) y en estudios más recientes , a destacar (Wang y col., 2020) y (Atmaja & Akagi, 2019), los retos que presenta la clasificación de emociones en el habla, son comúnmente abordados a través de una red de Memoria a Largo Corto-Plazo (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.

Las redes LSTM son un tipo de red recurrente que fueron diseñadas para resolver el problema de la dependencia a largo plazo del que sufre RNN.

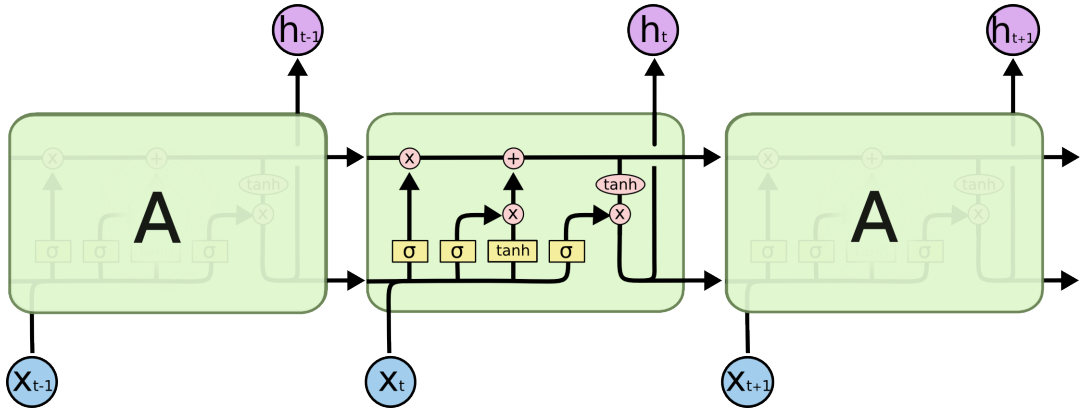


Figura 2.10: Estructura de una red LSTM. Fuente: Colah

Como muestra la figura 2.10, estas presentan una estructura en cadena al igual que las redes recurrentes, pero el módulo de repetición en lugar de tener una única capa de red neuronal, tiene cuatro que interactúan.

Redes Neuronales Convolucionales

Uno de los mayores avances de los últimos años en el campo de la inteligencia artificial son las redes convolucionales, debido a la alta precisión que proporcionan en el procesamiento de imágenes.

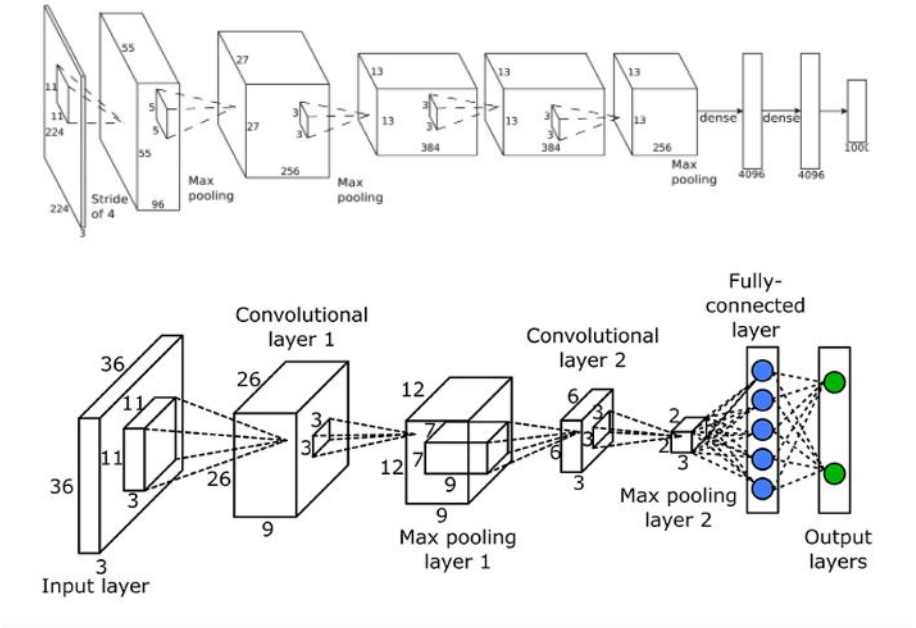


Figura 2.11: Estructura de una red CNN. Fuente: Brilliant

La tendencia de los modelos basados en redes neuronales densas en este ámbito es aprender características específicas desde varios métodos usados en el reconocimiento de emociones a través de la percepción acústica. En el uso de las Redes Neuronales Convolucionales (CNN) la idea principal es tomar ventaja de las propiedades de la señal, como la conectividad local, los pesos compartidos y el uso de varias capas (Lim y col., 2017). Estas suponen una importante contribución en la Clasificación Emocional de la Voz debido al uso de características significativas, y su uso en recientes estudios se ha incrementado a lo largo de los años donde destacan los trabajos de (Abdul Qayyum y col., 2019) y (Anvarjon y col., 2020).

2.1.2.4. Bases de datos

Aquí se recogen las bases de datos sobre el reconocimiento de emociones más usadas en los estudios comentados anteriormente.

- SAVEE (por sus siglas en inglés, *Surrey Audio-Visual Expressed Emotion*) es un conjunto de datos aplicado al reconocimiento de emociones que consiste en grabaciones de 480 frases en total en inglés británico ejecutadas por cuatro actores profesionales masculinos modulando siete emociones distintas (Enfado, Asco, Tristeza, Alegría, Miedo, Sorpresa y Neutral). El estudio (Jackson & ul haq, 2011) explora y detalla esta base de datos que data del 2011, y es de acceso público para propósitos de investigación.

- IEMOCAP (por sus siglas en inglés, *Interactive Emotional Dyadic Motion Capture*) es una base de datos multimodal privada utilizada para el reconocimiento y análisis de emociones. Consiste en doce horas de contenido audiovisual donde diez actores (cinco hombres y cinco mujeres) mantienen diálogos en inglés previamente transcritos en los que se interpretan cinco emociones distintas (Enfado, Tristeza, Alegría, Frustración y Neutral).

- RAVDESS (por sus siglas en inglés, *Ryerson Audio-Visual Database of Emotional Speech and Song*) es un popular conjunto dinámico multimodal (contiene varios formatos) donde veinticuatro actores profesionales vocalizan frases en inglés norteamericano modulando ocho emociones (Enfado, Calma, Asco, Tristeza, Alegría, Miedo, Sorpresa y Neutral). Tiene un total de 7356 grabaciones entre audio (hablado), au-

dio(canciones) y vídeo. Cuenta con una completa documentacion (Livingstone & Russo, 2018) y es de acceso público.

- TESS (por sus siglas en inglés, *Toronto Emotional Speech Set data*) es un conjunto de datos compuesto por 2800 archivos de audio donde dos actrices de 26 y 64 años cuya lengua materna es el inglés americano, articulan 200 frases cada una y modulándolas en siete emociones (Enfado, Asco, Tristeza, Alegría, Miedo, Sorpresa, y Neutral). Este proyecto fue realizado por la universidad de Toronto (Pichora-Fuller & Dupuis, 2020) y es de acceso público.

- EMO-DB es una base de datos alemana cuyos detalles se recogen en (Burkhardt y col., 2005), data del 2005 y es de acceso público. La conforma una colección de 800 grabaciones interpretadas por diez actores (cinco hombres y cinco mujeres) matizando seis emociones (Enfado, Asco, Tristeza, Alegría, Miedo y Neutral) y son llevadas a cabo en una cámara anecoica (capaz de absorber las ondas sonoras o electromagnéticas sin reflejarlas).

- CaFE (por sus siglas en inglés *Canadian French Emotional Speech Dataset*) es una base de datos canadiense en idioma francés donde seis hombres y seis mujeres, pronuncian un total de seis frases interpretando siete emociones (Enfado, Asco, Tristeza, Alegría, Miedo, Sorpresa y Neutral). Es de acceso libre y fue introducida en (Gournay y col., 2018)

2.2. Estado del Arte

Aunque la mayoría de estudios se basan en análisis que evalúan la precisión del clasificador sobre la propia lengua, el auge de las técnicas basadas en aprendizaje profundo ha permitido aumentar la capacidad de clasificación en los modelos de reconocimiento de emociones. La creación de un mapa de características a partir de diferentes representaciones de la onda sonora con respecto a su frecuencia y tiempo, permiten a las redes neuronales distinguir más atributos para hacer una predicción más exacta. A partir de este paradigma, el intento de clasificar emociones a través de la voz ha sido objeto de estudio aplicando diferentes técnicas.

En (Atmaja & Akagi, 2019) proponen sistema basado en una arquitectura LSTM bidireccional (BLSTM) que aplican a un subconjunto de la base de datos IEMOCAP para

distinguir entre cuatro emociones(Enfado, excitación, Tristeza y Neutral). Discuten la incapacidad del un modelo BLSTM para detectar características relevantes, y palían el problema añadiendo un modelo de atención. Teniendo ese modelo como punto de partida (BLSTM + modelo de atención), experimentan escogiendo diferentes valores de duración del silencio para medir la eficacia que tendría el modelo si este se eliminara. Además utilizan un complejo sistema de extracción de características entre las que destacan MFCC. Los resultados muestran un máximo del 70.34 % de precisión en los datos sin necesidad de eliminar el silencio de la señal de audio. Un año después, J.Wang (Wang y col., 2020) propone un modelo dual LSTM donde cada uterancia se procesa con características MFCC y espectrogramas de MEL simultáneamente. El modelo es entrenado y evaluado en el conjunto de datos IEMOCAP llegando a un 72.7 % de exactitud.

Las redes LSTM cubren en mencionado antes, efecto de contexto, resolviendo el desvanecimiento de gradientes que que se puede encontrar en las redes recurrentes, pero de manera aislada no son las que mejores resultados ofrecen y es por ello, que en otros trabajos se han combinado con CNN para aumentar su rendimiento. Por ejemplo en (Lim y col., 2017) 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. Al mismo tiempo, W.Lim y su equipo estudian el resultado de un sistema híbrido que usa una red convolucional distribuida en el tiempo (una combinación de CNN y LSTM) para clasificar emociones en una secuencia de audio, consiguiendo un 88.01 % de precisión. De nuevo, para aprovechar las ventajas que ofrecen las redes convolucionales, la señal es convertida a imagen (espectrograma), que es entrenada y probada con el corpus EMO-DB (alemán) distinguiendo entre siete emociones.

Por otro lado, en (Harar y col., 2017) describe un método que utiliza una arquitectura basada en Redes Convolucionales sin selección de características para distinguir únicamente entre tres emociones (Enfado, Neutral, y Tristeza) a través de la voz. Su objetivo es predecir el estado emocional de una persona en una grabación corta de audio donde la mencionada arquitectura consiste en seis capas convolucionales y tres densas. Como conjunto de datos utilizan la Base de Datos de Berlín de Discurso Emocional (EMO-DB), un corpus alemán que contiene un total de 800 frases (diez frases distintas re-interpretadas en siete emociones por cinco mujeres y cinco hombres), del cual extraen un subconjunto de 271 grabaciones etiquetadas. De las señales de audio con las que el sistema es alimentado se han eliminado los segmentos de silencio después de ser estandarizadas consiguiendo una

exactitud de 96.97 %.

Con el fin de eliminar el preprocesado de la señal, en (Abdul Qayyum y col., 2019) presenta un modelo de redes convolucionales para una clasificación de emociones en el idioma inglés. Este utiliza la base de datos SAVEE, la cual contiene 480 muestras que distinguen entre seis emociones, interpretadas por hombres y mujeres angloparlantes, donde obtiene finalmente un 81.63 % de precisión. Este trabajo llega a sus resultados mediante la comparación de tres enfoques (MVR, SVN, y RNN) en el que a cada uno le aplican tres métodos de extracción de características distintos, con el sistema propuesto basado en CNN sin ningún procesado de la señal, siendo este último el que consigue una mejor capacidad de predicción.

En estudios más recientes, (Anvarjon y col., 2020) aborda el problema del Reconocimiento de Emociones en el Habla con una red CNN computacionalmente eficiente que es alimentada con los espectrogramas de la señal; es decir, se consigue una representación en 2D de la señal de audio aprovechando mejor las ventajas de una red de este tipo.

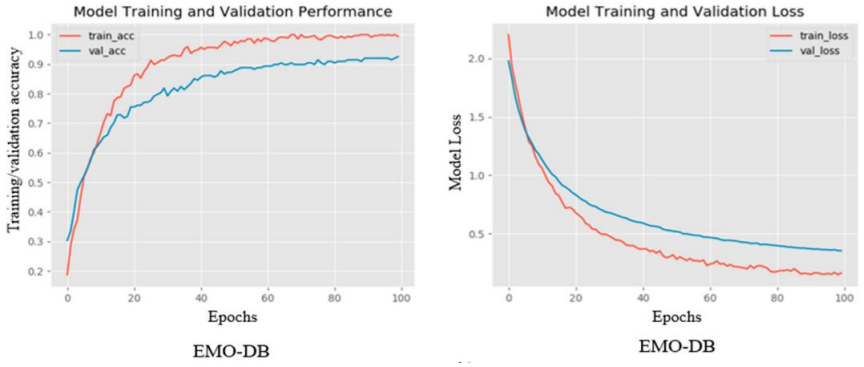


Figura 2.12: Resultados de T.Anvarjon sobre la base de datos EMO-DB. Fuente: (Anvarjon y col., 2020)

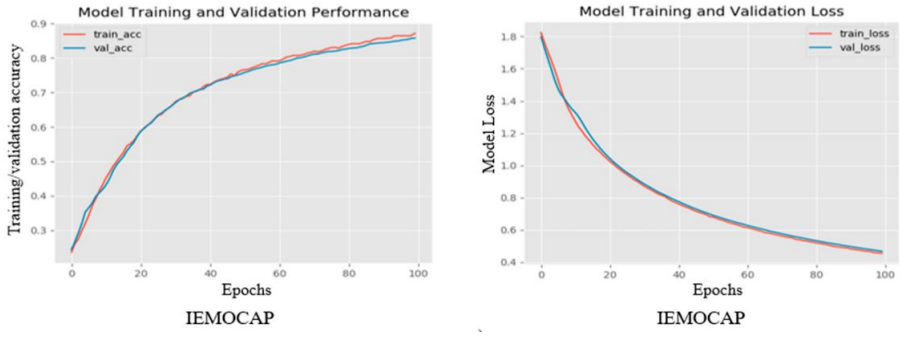


Figura 2.13: Resultados de T.Anvarjon sobre la base de datos IEMOCAP. Fuente: (Anvarjon y col., 2020)

El sistema es probado en con dos datasets distintos independientemente, IEMOCAP (en inglés, y eliminando 'frustración') y EMO-DB (alemán) consiguiendo un 77.01 % y 92.02 % de precisión respectivamente. Las figuras 2.12 y 2.13 muestran el rendimiento del modelo en los conjuntos de datos correspondientes comprobando el sólido resultado.

Siguiendo por el enfoque de CNN, en (Mustaqeem & Kwon, 2020) exploran una arquitectura basada en CNN compuesta por siete capas bidimensionales. Paralelamente la red es alimentada con espectrogramas y características extraídas a través de MFCC, y lleva a cabo una clasificación de 5 emociones evaluando el resultado en RAVDESS donde consiguen un 81.0 % de precisión e IEMOCAP con un 84.00 %

Finalmente, es obligatorio hablar del trabajo de (Tamulevicius y col., 2020), en una línea más cercana al objetivo de este proyecto, lleva a cabo un estudio del reconocimiento emocional empleando el cruce de seis lenguas (lituano, inglés, serbio, español, alemán y polaco). Aunque sus resultados no son realmente señalables y acaban diseñando un clasificador que es entrenado con todas las lenguas para distinguir las emociones indistintamente, dicha clasificación se lleva a cabo mediante el uso de una red neuronal convolucional bidimensional de tres capas, e insisten en la importancia del uso de características en dos dimensiones, ya que proveen información temporal además de las características acústicas de las emociones. En el estudio exploran varias de estas características, siendo el uso de cocleogramas las que consiguen una mayor exactitud.

Bajo estas líneas, la tabla resume los estudios previamente comentados en la sección 2.2.

Trabajo	Año	Método	Datos usados	Acierto
J.Wang	2020	LSTM dual + MFCC	IEMOCAP	72.7 %
Atmaja	2019	BLSTM + Att	IEMOCAP (4 emociones)	73.34 %
Anvarjon	2020	CNN 2D + espectrogramas Mel	IEMOCAP (4 emociones)	77.01 %
Mustaqeem	2020	CNN 2D + MFCC	RAVDESS (5 emociones)	81.01 %
Abdul	2019	CNN	SAVEE	81.63 %
Mustaqeem	2020	CNN 2D + MFCC	IEMOCAP (5 emociones)	84.00 %
W.Lim	2017	LSTM + CNN	EMO-DB	88.01 %
Anvarjon	2020	CNN 2D + espectrogramas Mel	EMOD-DB	92.01 %
Harar	2017	CNN	EMOD-DB (3 emociones)	96.97 %
Tamulevicius	2020	CNN 2D + cocleogramas	Lithuanian	97.00 %

Tabla 2.1: Tabla comparativa y resumida de los trabajos mencionados

En la tabla 2.1 se muestra un resumen de los estudios relacionados con este campo y sus correspondientes resultados (ordenados de manera ascendente por porcentaje de acierto), así como los métodos y datos que se han usado para ello. La columna de Método, comprende la arquitectura completa, es decir, el sistema de clasificación usado y el método de extracción de características si lo hubiera. En la tercera columna se describen los datos usados (la base de datos) y el número de clases entre las que el sistema ha tenido que diferenciar. Cuando estas no se especifican, quiere decir que se ha usado el conjunto de datos al completo. En el último trabajo, (Tamulevicius y col., 2020) se lleva a cabo la comparación de distintos datos, pero es *Lithuanian* la base de datos lituana en la que se basan para crear el modelo, cuya arquitectura es la que se especifica en la tabla.

2.3. Conclusiones parciales

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. Se observa que los retos que plantea SER se han abordado anteriormente desde distintos enfoques, pero en su mayoría, desde el punto de vista de un único lenguaje. Las dificultades que presenta esta tarea en la lengua extranjera se deben

principalmente a las posibles variaciones de aire para expresar la mismas emociones. Este mismo problema se plantea en una modalidad diferente pero bastante relacionada como es la transcripción de la voz a texto (ASR, Reconocimiento Automático del Discurso), sin embargo estos estudios requieren un análisis más profundo de la fonética propia de cada lenguaje.

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 (Pell & Skorup, 2008) 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.

Cabe destacar que la combinación de métodos, véase algoritmos de clasificación, filtros para preprocesar la señal, y métodos de extracción de características, así como distintos conjuntos de datos, es realmente diversa, por lo que visualizar una dirección clara para determinar qué línea es la mejor se diluye.

No obstante hay observaciones, que pueden llevar a una conclusión general; Por ejemplo, y de manera intuitiva, cuanto mayor es el número de clases (emociones en nuestro caso), peor será la capacidad de clasificación de la red, y por eso en algunos trabajos se extrae un subconjunto reduciendo las opciones entre las que clasificar.

En la extracción de características el uso de MFCC ha sido amplia y tradicionalmente escogido al reportar resultados más elevados en comparación con otros métodos. En (Langari y col., 2020) denota que los métodos de extracción de características son MFCC y LPCC porque las variaciones en la frecuencia del tono están significativamente relacionados con la expresión humana de emociones. Los parámetros que se computan en MFCC como el número de filtros o la escala de frecuencia, son a menudo escogidos de manera experimental y dependen en gran medida del conjunto de datos con el que se pruebe y el clasificador que implemente el sistema.

En cuanto a los métodos usados como clasificadores, se observa también que CNN (con diferentes modificaciones dependiendo del estudio) es la opción más sólida entre los trabajos más recientes, debido principalmente a que reduce la señal de audio a sus características más relevantes, y la combinación de probabilidades resultantes identifica conjuntos de atri-

butos que determinan una clasificación. El uso de espectrogramas, aprovecha las bondades que las redes convolucionales ofrecen.

3. Objetivos y metodología de trabajo

3.1. Objetivo General

1. Hacer un estudio comparativo del reconocimiento de emociones por voz, a través de lenguajes no aprendidos (lenguajes que no hayan formado parte del entrenamiento), una vez se haya conseguido un modelo capaz de clasificar en una lengua conocida con un porcentaje de acierto superior al 81 %.

3.2. Objetivos específicos

Para conseguir el alcance establecido, es necesario que los siguientes puntos sean satisfechos:

- Hacer un estudio del estado del arte sobre diferentes métodos, técnicas, y conjunto de datos utilizados en el reconocimiento de emociones a través de la voz. Aquí también se explorará si se dispone de la documentación necesaria, cómo cada uno de esos métodos pudiera estar relacionado con la lengua que usa para aplicarlo y su fonética.
- Conseguir al menos tres datasets pertenecientes a tres idiomas diferentes donde uno de ellos será usado como referencia, y además, deberán cumplir las siguientes condiciones: Uno de los conjuntos de datos restantes deberá tener raíces fonéticas distintas al corpus de referencia, y el otro tener raíces fonéticas similares.
- Diseñar una solución en la que el conjunto de datos de referencia tenga un porcentaje de acierto superior al 81 % en la clasificación de emociones. Esta referencia ha sido marcada por los resultados reportados en la revisión del estado del arte.
- Aplicar el modelo diseñado en el paso anterior a los otros conjuntos de datos.
- Evaluar la tasa de acierto obtenida en cada uno de eso conjuntos y comparar los resultados obtenidos.

3.3. Metodología de trabajo

Para este proyecto se plantea una metodología de desarrollo iterativa (ver figura 3.1), en las que tras una fase inicial, el proyecto entra en un bucle donde el trabajo pasa por una serie de etapas que se repiten durante la vida del proyecto.
Al contrario que en desarrollos de software más tradicionales donde podrían verse flujos de trabajo basados en metodologías ágiles o en cascada, se ha considerado que este modelo se adapta mejor a las necesidades de este proyecto, debido principalmente, al grado de incertidumbre que presenta un proyecto basado en Inteligencia Artificial en comparación con la ingeniería del software estándar. Por ejemplo, una metodología ágil asume que pequeños cambios funcionales hace posible el alcance de los objetivos a bajo coste y alta predictibilidad. Esto no se corresponde con este tipo de trabajo por las siguientes características:

- Es difícil conocer los costes y riesgos de la mayoría de los requisitos. Por ejemplo el estudio del conjunto de datos es algo que afecta directamente a la elaboración de los distintos modelos, y por lo tanto el crecimiento funcional es indeterminado.
- Los cambios o modificaciones no pueden ser aplicados por diseño, requieren experimentos, además esas modificaciones son realmente complicadas de atomizar, por lo que el coste es impredecible.
- Si bien es difícil tener una conclusión final por las estrictas fechas de entrega, un modelo iterativo permite obtener datos presentables a lo largo del proyecto.

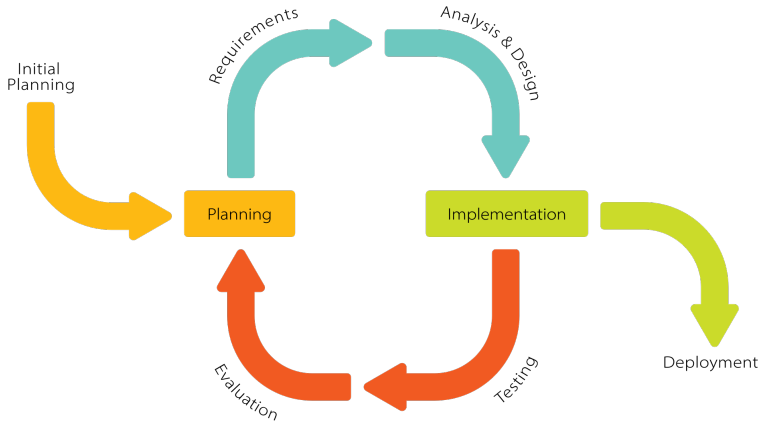


Figura 3.1: Proceso iterativo propuesto Fuente: Wikipedia

Dentro de la metodología propuesta, en cada iteración se diseñan unas modificaciones y capacidades funcionales que son añadidas en función de la etapa anterior.
Este tipo de metodología es normalmente adoptada en desarrollo de producto (Larman & Basili, 2003), pero para esta comparación se ha modificado mediante la extracción de unos pasos iniciales del bucle que la caracteriza, adaptándola mejor a nuestras necesidades.

1. Fase inicial:

   a) Revisión de la literatura sobre el reconocimiento de emociones en el habla, así como los métodos usados y los resultados obtenidos. Este paso permite una mayor comprensión del problema, y su alcance.

   b) Análisis y recolección de posibles conjuntos de datos en diferentes idiomas, aptos para los experimentos que se quieren realizar.

2. Elaboración: se llevan a cabo los experimentos y la implementación de los componentes mediante unos pasos iterativos, a saber:

   a) Identificación y redacción de una serie de pruebas iniciales con los diferentes métodos y técnicas descritas en la literatura, aplicados según el análisis de las bases de datos.

   b) Implementación en Python de las pruebas diseñadas con las técnicas y arquitecturas identificadas.

   c) Ajuste de los parámetros así como el balance de los datos con el fin de conseguir un mejor resultado.

   d) Evaluación: Se evalúan los resultados obtenidos de la implementación antes de decidir la iteración por finalizada. En caso de que los resultados no sean los esperados hay dos posibilidades: si los errores obtenidos no distan demasiado de las especificaciones parciales, se pueden realizar reajustes en el modelo. Si por el contrario, dichos resultados están demasiado lejos del objetivo, se creará una segunda versión con una estrategia diferente, añadiendo más iteraciones al proceso de elaboración.

3. Evaluación y comparación de los resultados.

4. Planteamiento de la comparativa

En este capítulo se identificará el problema en concreto a tratar, a la vez que el diseño de los experimentos para acometerlo. Para ello se exponen los datos utilizados junto con las técnicas de procesamiento y el diseño de las redes neuronales que se usan en este trabajo. El objetivo de esta comparativa es contrastar los resultados obtenidos tras aplicar el mismo sistema de reconocimiento de emociones en la voz entrenado con un lenguaje de referencia, con los otros dos lenguajes escogidos. Mediante esta comparativa se pretende responder a la pregunta si es posible reconocer emociones en un idioma que en principio se desconoce.

4.1. Conjunto de Datos

Los datos en un proyecto de inteligencia artificial son clave de cara a la obtención de un resultado coherente en nuestro trabajo. Este estudio pretende analizar si es posible clasificar emociones en la lengua extranjera, y para encontrar una respuesta se seguirá la estrategia que se presenta a continuación respecto a los datos. Es importante mencionar, que por cuestiones de coherencia entre las bases de datos (garantizar que todas las bases de datos con las que se va a trabajar comparten exactamente las mismas emociones) se extraerán de los conjuntos originales seis emociones para clasificar en este trabajo: Enfado, Asco, Tristeza, Miedo, Felicidad y Neutral. En las figuras correspondientes a cada conjunto se ha marcado en un color más claro la emoción que se suprime.

4.1.1. Idioma de referencia: Inglés

El idioma de referencia será el que aprenda el modelo, y desde el cual se intenten reconocer emociones en otras lenguas, es decir, el idioma que se use en el entrenamiento. En este caso se propone el inglés.
Las bases de datos a las que se ha tenido acceso son limitadas y no presentan un gran número de muestras, por lo que en previsión de un rendimiento pobre en el modelo, se decidió escoger varias del mismo idioma de cara a un entrenamiento más completo.

RAVDESS Como ya se hizo ver en la sección 2.1.2.4, RAVDESS (por sus siglas en inglés *Ryerson Audio-Visual Database of Emotional Speech and Song*), contiene 7356 archivos en

total, entre los cuales se pueden encontrar tres modalidades: sólo audio (en 16 bit, 48 kHz y en formato wav), audio-video (720p H.264, AAC 48kHz, en formato mp4) y sólo video sin sonido. Esta base de datos contiene veinticuatro actores profesionales vocalizando dos frases en inglés norte americano (*Kids are talking by the door* y *Dogs are sitting by the door*).

Cada uno de estos archivos están nombrados de manera única mediante siete números a modo de descripción de las características del audio. Éste respeta la siguiente convención:

- Modalidad (01 Audio y vídeo, 02 Sólo vídeo, 03 Sólo audio).
- Canal vocal (01 discurso normal, 02 canción).
- Emoción que representa.
- Intensidad Emocional Si es normal o fuerte. La voz Neutral no contempla la intensidad fuerte.
- Repetición (si es la primera repetición 01, si es la segunda 02).
- Actor que ejecuta la acción.

Así por ejemplo, el archivo 03-01-03-01-01-01-01.wav dirá que es un archivo de sólo audio (03), donde se vocaliza una frase de manera hablada (01) y con tono alegre (03). La intensidad es normal (01), corresponde a la primera repetición (01) y el actor que la ejecuta es el n.01.

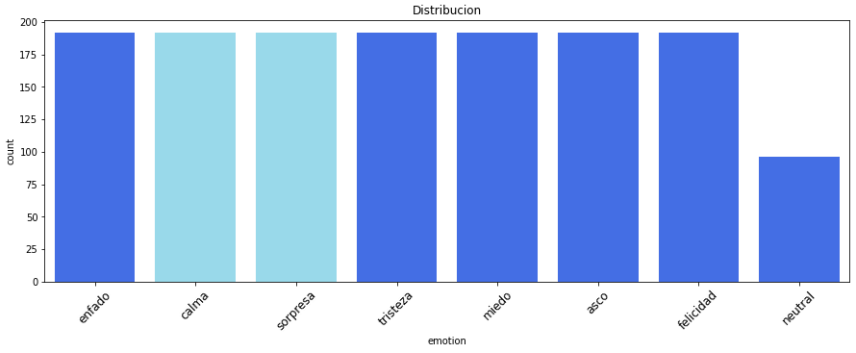


Figura 4.1: Distribución de las emociones en RAVDESS en la modalidad sólo audio (1440 archivos)

A pesar de que este dataset hay 7356 muestras para este proyecto únicamente se usarán aquellas que presentan una modalidad de sólo audio, lo que deja un total de 1056 muestras (tras eliminar Sorpresa y Calma), sin embargo como se puede apreciar en la figura 4.1 las emociones están bien distribuidas.

SAVEE (por sus siglas en inglés, *Surrey Audio-Visual Expressed Emotion*), cuenta con 480 archivos de audio en formato wav muestreados a 44.1 kHz, donde cuatro actores anglosajones (inglés británico) de 27 a 31 años, modulan siete emociones con frases específicas a cada una. Cada archivo ha sido etiquetado de manera que el primer carácter (o caracteres, antes de un dígito) corresponde a la emoción que representa. Así las letras 'a', 'd', 'f', 'h', 'n', 'sa' y 'su', corresponden a Enfado (*angry*), Asco (*disgust*), Miedo (*fear*), Felicidad (*happiness*), Neutral (*neutral*), Tristeza (*sadness*), Sorpresa (*surprise*). El número que le sigue a continuación se refiere a la frase pronunciada. Por ejemplo, el archivo d03.wav hace referencia a la tercera frase de la emoción Asco.

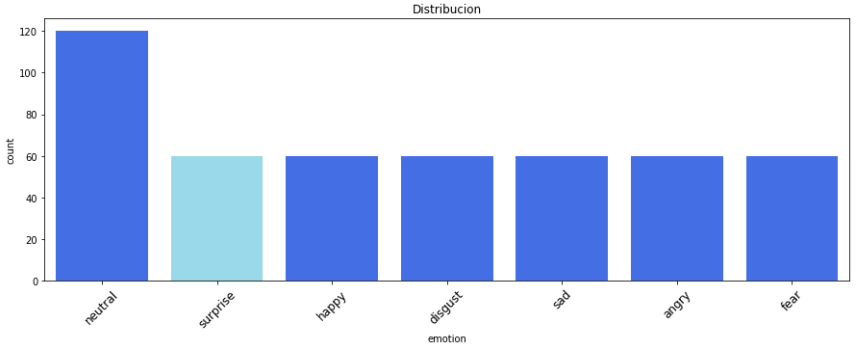


Figura 4.2: Distribución de las emociones en SAVEE

A pesar de que sólo presenta 480 audios, el hecho de que se use una frase distinta para cada emoción, lo convierte en un conjunto de datos muy completo para el entrenamiento. La figura 4.2 muestra la distribución de las emociones, observando que están bien balanceadas.

TESS Este dataset lo conforman 2800 archivos de audio en formato wav donde dos actrices angloparlantes de 26 y 64 años vocalizan 200 palabras insertadas en la frase *Say the word___* donde se interpretan siete emociones a Enfado, Asco, Miedo, Felicidad, Neutral, Tristeza y Sorpresa. Los archivos están organizados en carpetas atendiendo a la actriz y a la emoción que representa, y nombrados mediante 3 cadenas de caracteres separadas por un guión bajo donde la primera indica la actriz, la segunda la palabra que se pronuncia, y el tercero la emoción.

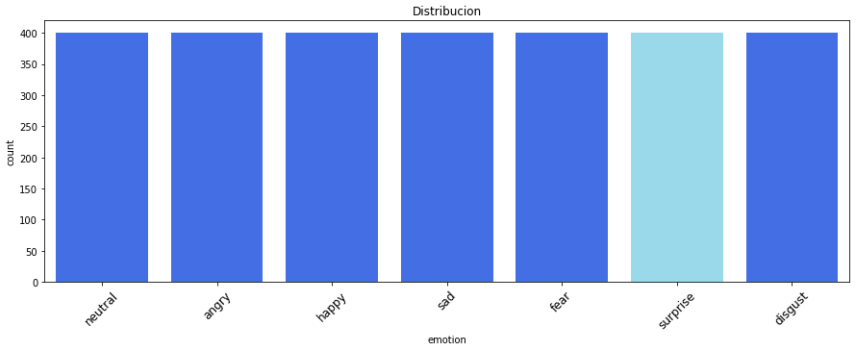


Figura 4.3: Distribución de las emociones en TESS

En la figura 4.3 se puede apreciar la distribución de las emociones en la base de datos TESS.

4.1.2. Idioma con raíces fonéticas similares: Alemán

Este conjunto de datos pertenecerá a un idioma con unas raíces similares al idioma de referencia, de manera que se espera al menos haya un porcentaje mayor de reconocimiento que en el segundo conjunto de datos de validación. Para este caso se propone el alemán ya que el idioma de referencia (inglés) es una lengua germánica occidental (Fennell, 2001).

Para este caso el conjunto de datos propuesto es la Base de Datos del Discurso Emocional de Berlín (EMODB, por sus siglas en inglés *Berlin Database of Emotional Speech*). Este corpus contiene 800 grabaciones interpretadas por diez actores (cinco hombres y cinco mujeres) modulando siete emociones en el idioma alemán. Cada archivo tiene una frecuencia de muestreo de 16 kHz con una resolución de 16 bits, y una duración de 3 segundos de media. Como en el anterior, se utiliza una nomenclatura para nombrar a los archivos que satisface lo siguiente:

- Las dos primeras posiciones determinan el actor que las interpreta.
- De la posición 3 a las 5 se define el texto que se pronuncia
- La posición 6 indica la emoción.
- Versión del audio en caso de que la hubiese (codificado con letras).

Como ejemplo, el archivo *03a01Fa.wav* indica que el actor 03 (hombre de 31 años) cita el texto a01 (*Der Lappen liegt auf dem Eisschrank*, en alemán .[El] mantel está colgando del frigo"), con la emoción F (Felicidad), y es la versión *a* (la primera).

La documentación del corpus también nos ofrece información sobre el género y edad de los actores, lo cual se ha determinado irrelevante, y las distintas frases que pueden aparecer en los archivos. Las emociones que clasifica son Enfado (W), Aburrimiento (L), Asco (E), Miedo o Ansiedad (A), Felicidad (F), Tristeza (T) y Neutral (N) codificadas en el nombre del archivo por su inicial en alemán (especificada entre paréntesis).

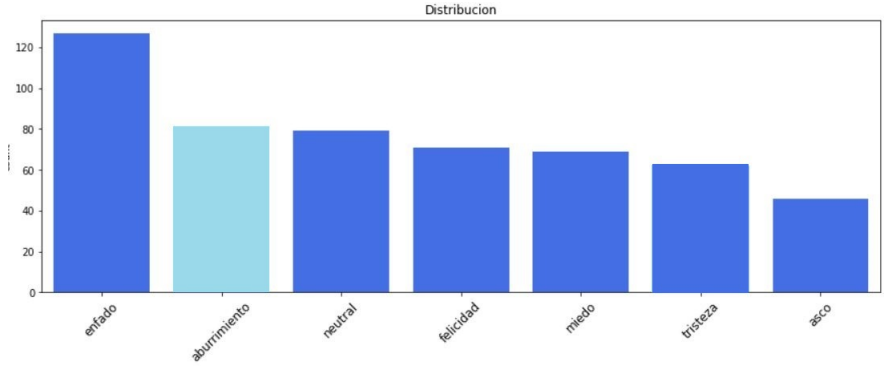


Figura 4.4: Distribución de las emociones en EMO-DB

En la figura 4.4 se puede apreciar la distribución de las clases en EMO-DB que como puede verse, están muy desajustadas.

4.1.3. Idioma con raíces fonéticas distintas: Francés

Este conjunto de datos pertenecerá a un idioma con unas raíces más distantes al idioma de referencia.

Con este conjunto se espera que haya una diferencia notable con respecto al idioma con raíces fonéticas similares, reportando un menor porcentaje de reconocimiento. Para este caso se propone el idioma francés (que es una lengua romance), utilizando para ello el conjunto de datos en francés canadiense para el reconocimiento de emociones, CaFE, el cual contiene seis frases diferentes pronunciadas por seis hombres y seis mujeres en seis emociones básicas además de Neutral, grabadas cada una en dos intensidades distintas. Este conjunto fue grabado a alta resolución en formato aiff con una frecuencia de muestreo de 192 kHz y 24 bits por muestra. Los archivos están organizados en una jerarquía de carpetas, de manera que no es necesario escanear el archivo para saber el genero del actor, la intensidad, o la emoción representada en el audio, sino que basta irse a la carpeta correspondiente. Las emociones que se incluyen aquí son Enfado (*colere*), Asco (*degout*), Alegría (*joie*), Tristeza(*tristesse*), Miedo (*peur*), Sorpresa (*surprise*) y Neutral (*neutre*). En la figura 4.5 se ve la distribución original de las mismas.

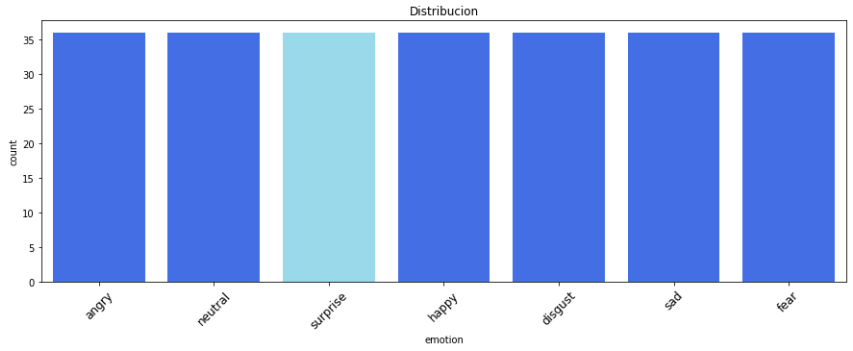


Figura 4.5: Distribución de las emociones en CaFE

4.2. Extracción de características

Teniendo en cuenta el previo estudio de la literatura en el capítulo 2 se concluye que los métodos más prometedores, y que por lo tanto merecen la pena aplicar a este estudio comparativo serían los siguientes:

4.2.0.1. Coeficientes Cepstrales en la escala de Mel

Como ya se ha mencionado, MFCC es uno de los mejores algoritmos para capturar características de la señal de audio por su similitud a cómo el sistema auditivo humano procesa el sonido y las frecuencias, de la misma manera, su efectividad se ha visto reportada y discutida a lo largo de otros estudios. La librería usada para la manipulación de audio Librosa ofrece la posibilidad de extraer características MFCC de un archivo de audio. En cuanto a la configuración, se extraerán 13 características MFCC usando el rango de muestreo del propio archivo de audio (Bao & Huang, 2019).

4.2.1. Espectrogramas

Como se vio en el capítulo 2, y siguiendo los pasos de los trabajos (Anvarjon y col., 2020) y (Mustaqeem & Kwon, 2020), el uso de espectrogramas hace referencia a la conversión de la señal a imagen, y el objetivo de esta técnica es aprovechar las fortalezas de las redes convolucionales en las imágenes aplicándolas a un problema de señal de audio. En concreto, para este trabajo se hará uso de los espectrogramas de las características MFCC de la señal, cuyo proceso constará de dos partes:

1. Generación de espectrogramas como imagen.

2. Lectura y procesado de las imágenes que alimentarán la red.

Esto espectrogramas se generarán con el paquete Librosa, especificando en los correspondientes parámetros la extracción 13 características MFCC; una vez generadas, las imágenes se guardan en disco recortando el *padding* 0.05 pulgadas y en formato jpg. Finalmente, las imágenes generadas son leídas con la ayuda de OpenCV, donde se transforma su canal de color a RGB y son re-dimensionadas con un tamaño de 40 x 30 píxeles. Las figuras generadas tendrán un aspecto como el que se muestra en la figura 4.6

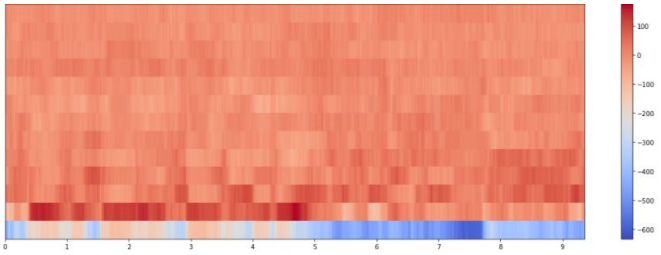


Figura 4.6: Espectrograma MFCC de una onda de audio. Fuente propia

4.3. Configuración

En esta sub-sección se muestra los recursos a los que se ha accedido para el desarrollo del estudio, así como su correspondiente configuración.

Google Colab Para la exploración de los datos así como el desarrollo, y entrenamiento de los modelos se ha hecho uso de la plataforma gratuita desarrollada por Google, Google Colab. Esta plataforma ofrece 12GB de RAM y 107.77GB de espacio en disco, que será más que suficiente dado el tamaño que nuestro dataset.

Librosa 0.8.1 Librosa es un paquete que ofrece diversas funcionalidades para el análisis de audio y música, cuya información más en detalle se puede encontrar en (McFee y col., 2015). Esta librería ha sido esencial para la extracción de características MFCC así como algunas técnicas de aumento de datos.

OpenCV 3.4.2 OpenCV es una librería de código abierto desarrollada por INTEL en 1999, cuyo principal objetivo es la provisión de funciones y recursos para visión computacional, cubriendo áreas como la reconstrucción 3D, detección de movimiento o reconocimiento

de objetos, etc. Si bien en este proyecto no se necesitarán sus recursos más avanzados, será útil en la lectura de espectrogramas como imagen.

Tensorflow 2.0 Es una plataforma de código abierto originalmente creada por Google que provee un conjunto de librerías y recursos para el desarrollo de modelos con aprendizaje automático. Tensorflow, que además ofrece soporte de Keras, se ha usado tanto para la estandarización de los datos, como para la compilación y entrenamiento del modelo.

Python 3 Python es un lenguaje interpretado de alto nivel. Todo el código para este proyecto ha sido desarrollado en Python 3.

4.4. Pre-procesado de los datos

Como se ha visto en la sección 4.1 no hay una abundante disposición de datos, esto podría convertirse en un problema y perjudicar el rendimiento del modelo en el entrenamiento. Será necesario, antes del entrenamiento, un previo procesado de los datos.

4.4.1. Normalización, estandarización y balanceado de los datos

Los siguientes puntos se aplicarán a todas las pruebas:

- El proceso de estandarización consistirá en el cociente entre la media artimética de los valores de los datos de entrenamiento, y la desviación normal de los de test. En esta técnica los valores son centrados con respecto a su media con una desviación estándar, consiguiendo una mejora en la estabilidad numérica del modelo. Ya que, a lo largo de las pruebas se usarán combinaciones de aumentos de datos e incluso mezclas entre distintos datasets, es recomendable aplicar este paso.

- La categorización cambia el formato de los datos para su uso en el modelo con keras. En nuestro caso se utilizará la codificación *One Hot* que representa los enteros en secuencias de bits. Esto se hace con el fin de evitar que los números enteros no confundan al modelo asignándole una peso proporcional al índice asociado a una clase (categoría).

- La división de los datos en entrenamiento (70 %), y validación(20 %) y test(10 %), se hará con el algoritmo *StratifiedShuffleSplit* de la librería de Python *sklearn* que

además se encarga de barajar de manera aleatoria los datos previamente y asignar cantidades equitativas a las clases cuando se divide en entrenamiento y test.

4.4.2. División de los datos por género

En un primer análisis exploratorio de los datos, se han estudiado las diferencias entre la voz masculina y la voz femenina en las emociones, observando lo siguiente:

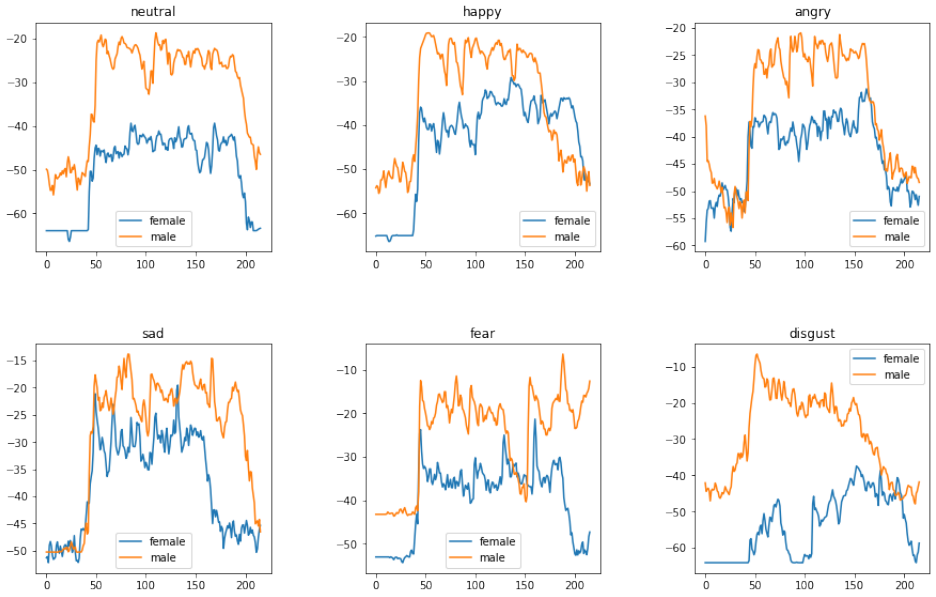


Figura 4.7: Comparativa de los extractos de voz por género en RAVDESS. Fuente propia

En la figura 4.7 se puede observar la comparación de la voz masculina (naranja) y la voz femenina (azul) por cada una de las emociones en el idioma inglés. Ya que es muy distinto, puede ser recomendable dividir el conjunto de datos atendiendo a esta característica.

4.4.3. Técnicas de aumento de datos

Dado el bajo número de muestras en los distintos conjuntos de datos a los que se pudo acceder, se ha visto conveniente explorar distintas técnicas de aumento de datos. El aumento de datos es una técnica por la cual se aumenta el número de muestras en un conjunto mediante la creación de nuevas muestras sintéticas con pequeñas modificaciones a cada uno de los archivos. Esta aumento de los datos se puede traducir por una reducción del overfitting (sobreajuste), ya que el modelo se mantendría invariable mejorando así su capacidad de generalización. Esta técnica es ampliamente conocida cuando se procesan imágenes, siendo esas modificaciones rotaciones, transposiciones etc.

Se sabe que el sonido tiene las siguientes características: tono, duración, timbre e intensidad(Ayadi y col., 2011), por lo que se deben modificar levemente los datos alrededor de esas características de manera que sólo difieran en pequeños factores de la muestra original.

4.4.3.1. Ruido Blanco

Añadir ruido blanco a la pista de audio, implica la inyección de valores aleatorios distribuidos de manera irregular con una media de 0 y una desviación de estándar de 1. Para implementar este método se usará el paquete numpy.

4.4.3.2. Desplazamiento del sonido

Desplaza el sonido hacia la izquierda o la derecha una cantidad aleatoria de segundos. Por ejemplo, si el sonido ha sido desplazado hacia delante (izquierda) x segundos, los x primeros segundos se marcan con 0. Si por el contrario han sido desplazados hacia detrás (derecha), los últimos x segundos se marcarán con 0.

4.4.3.3. Modulación del tono

Se refiere al proceso de cambiar el tono a un sonido sin variar su velocidad. Para implementar este método se usará la librería Librosa que ofrece un método específico.

4.5. Arquitectura

Como se ha podido ver en la revisión de la literatura del capítulo 2, las redes convolucionales esta una tendencia muy adoptada en los últimos trabajos en esta área de estudio.

4.5.1. Arquitectura principal

4.5.1.1. Modelo CNN-LSTM

- 3 capas convolucionales unidimensionales con 64 filtros de 3 x 3 y activación Relu, seguidas de una capa Max Pooling con tamaño 2 para la pool.
- Una capa Flatten con un Dropout del 25 %
- 2 capas LSTM unidimensionales con 50 y 20 unidades respectivamente y un Dropout del 50 %. Sólo se permitirá a la primera capa LSTM devolver el estado oculto de salida

por cada entrada de tiempo. Ya que a este nivel se cambia a redes unidimensionales, se deberá redimensionar la entrada a 1 x 960.

- Capa de salida densa de 7 nodos y función Softmax.

La figura 4.8 muestra un esquema gráfico de los puntos mencionados.

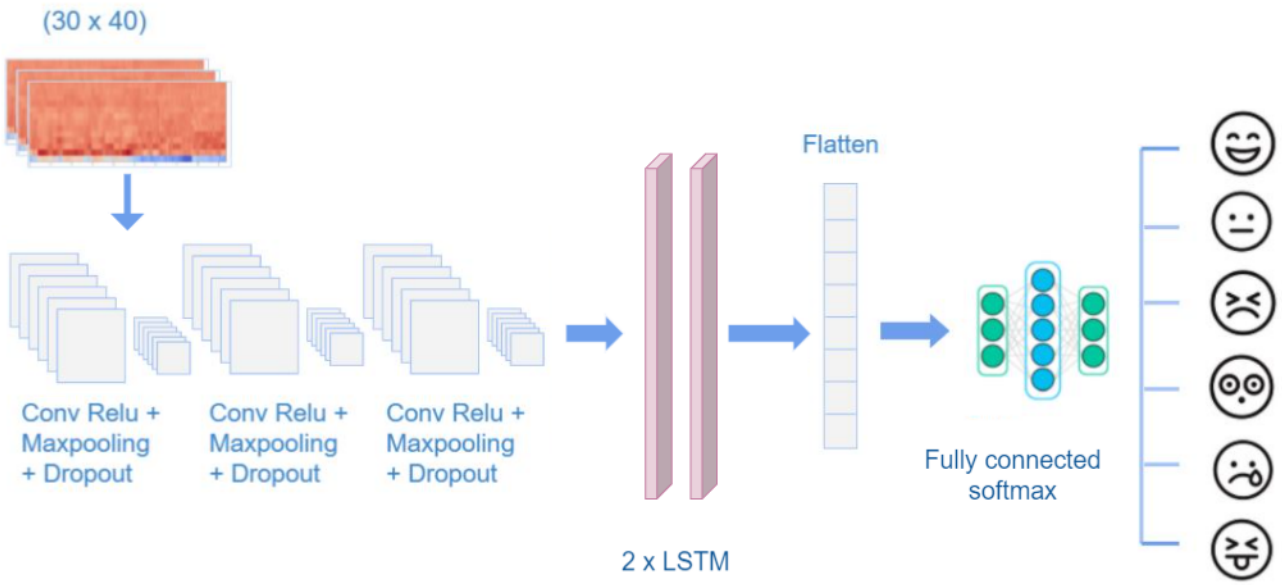


Figura 4.8: Arquitectura propuesta. Fuente propia

La estrategia de entrenamiento que se siguió fue un optimizador Adam con los parámetros por defecto que ofrece Keras y la entropía cruzada categórica (*categorical crossentropy*) como función de pérdida.

4.5.2. Otras arquitecturas

A pesar de que la arquitectura citada en la sección anterior, será con la que se lleven a cabo los principales experimentos y con los resultados más interesantes, se considera que la mención de otras arquitecturas que han formado parte de los experimentos de este trabajo, ayudará al lector a entender la progresión de los mismos, por lo que se ha visto conveniente incluirlas.

4.5.2.1. CNN 1D simplificado

- 3 capas convolucionales unidimensionales con activación Relu. El número de filtros es de 128 y y el tamaño del kernel de 5 con BatchNormalization y MaxPooling de

tamaño 8.

- 1 capa convolucional unidimensional con activación Relu. El número de filtros es de 128 y y el tamaño del kernel de 5 con BatchNormalization y Dropout del 10 %.
- 1 capa densa de 128 nodos.
- Por último esta arquitectura cierra con una capa densa con 7 nodos (número de clases) con una función de activación Softmax.

4.5.2.2. Modelo CNN 1D

En un principio se optó por una línea de trabajo inspirada en el estudio de (Abdul Qayyum y col., 2019) ya que combina buenos resultados y un sistema sencillo. Pero tras varias pruebas esta arquitectura se ha refinado hasta definirse, por ahora, lo siguiente:

- 2 capas convolucionales unidimensionales con activación Relu. El número de filtros es de 128 y y el tamaño del kernel de 5. En las dos capas convolucionales se usa regularización de tipo L2 para aplicar una penalización a las capas del kernel con un valor de 0.01 y corregir así el overfitting.
- La primera capa convolucional está seguida de una capa Dropout del 0.5 y una capa Max Pooling con un tamaño 8.
- A la segunda capa convolucional se sigue otra capa de Dropout con un valor del 25 % y una capa Flatten.
- Por último esta arquitectura cierra con una capa densa con 7 nodos (número de clases) con una función de activación Softmax.

En cuanto al entrenamiento del modelo se usará un optimizador RMSprop con una tasa de aprendizaje de 0.00005, valor de ρ de 0.9 y ϵ a 'None', por dar mejores resultados frente a Adam del caso inicial desde el que se partió, mientras que la función de pérdida utilizada para este propósito será entropía cruzada categórica (*categorical crossentropy*).
Se añaden además, para intentar afinar el modelo los callbacks ReduceLROnPlateau, que reduce la tasa de aprendizaje cuando el modelo ha dejado de mejorar y EarlyStopping, que detiene el entrenamiento si se ha llegado una meseta, es decir, si durante un determinado número de épocas, el modelo ha dejado de mejorar. En ReduceLROnPlateau se monitorizará la *val loss* con el fin de minimizarla, y como configuración se empleará un factor

de reducción de la tasa de aprendizaje de 0.9, una paciencia de 20 épocas y una tasa de aprendizaje mínimo de 0.000001. Con respecto a EarlyStopping, la variable monitorizada será 'val accuracy' con el fin de maximizarla y una paciencia de 20 épocas.
Este entrenamiento se llevará a cabo durante 1000 épocas cpn nun batch de tamaño 16.

4.5.2.3. Modelo CNN 2D

Esta arquitectura consiste en:

- 3 capas convolucionales bidimensionales con 32 filtros y un tamaño del kernel de 4 x 10. Como función de activación se usa Relu y pading establecido a 'same'.
- A las todas las capas convolucionales, les sigue una capa Max Pooling con tamaño para la *pool* de 3, que posteriormente se aplica un Dropout del 20 %
- Una capa Flatten, seguida de la capa densa de salida con 7 nodos y activación Softmax.

La estrategia de entrenamiento que se siguió fue un optimizador Adam con los parámetros por defecto que ofrece Keras y la entropía cruzada categórica (*categorical crossentropy*) como función de pérdida.

4.6. Criterios de éxito

El objetivo de esta sección es definir las métricas que se usarán para comparar los distintos modelos en los experimentos parciales, así como los resultados obtenidos al aplicar dichos modelos a los datos mencionados en la sección 4.1.
Las dos principales métricas que se usarán para decidir cómo de buena es la predicción del modelo serán:

- **Exactitud o *Accuracy*** Establece una comparación entre los resultados predichos y los obtenidos determinando cómo de preciso es el algoritmo cuando se trata de identificar las clases

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

- **F1 score** Siendo el *recall* la fracción de elementos relevantes que son recuperados (el cociente de las predicciones positivas y el número de clases positivas en el conjunto de test), la medida de F1 Score conviene el balance entre la precisión y el *recall*.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

En algunos casos, la *accuracy* puede ser engañosa debido a la paradoja de la exactitud donde puede existir un sesgo por a una distribución desbalanceada de las clases (Valverde-Albacete y col., 2013). Esto hace que pueda ser más inteligente elegir un modelo con menor exactitud pero con mayor poder predictivo. Para ver ese poder predictivo es, por lo tanto, aconsejable elegir más de una métrica de evaluación. Para ello se contará con F1 Score que es la media armónica entre el *recall* y la precisión.

Estas métricas se aplicarán a todas las clases en cada uno de los experimentos con el objetivo de determinar cómo de bueno es el modelo a la hora de reconocerlas. Como se ha visto en la presentación de las bases de datos, la clases tienden a estar muy balanceadas con la excepción de EMO-DB en alemán. Por ello, para tener una representación más igualitaria se aplicarán técnicas de aumento de datos a este conjunto y se balanceará el número de sus clases.

Accuracy y F1 se tomarán por cada una de las clases, a la vez que se observa el rendimiento del modelo por cada época para comprobar si reporta un ajuste correcto.

4.7. Diseño de los experimentos

En la presente sección se expondrán los diseños de los experimentos para cumplir los objetivos establecidos en el capítulo 3. Por un lado hay que conseguir un modelo que sea capaz de clasificar satisfactoriamente en el idioma con el que se entrena, para que finalmente se evalúe este modelo en otras lenguas distintas permitiendo hacer una comparación. Por

ello, se divide la presente en dos bloques, con la intención de que al lector le resulte más fácil seguir la narrativa de este proyecto. Dentro de estos bloques se muestran los experimentos planteados, que a su vez están formados por varias pruebas con una configuración similar.

4.7.1. Búsqueda del mejor modelo

Estos experimentos giran en torno a la búsqueda del modelo en el que, más tarde, se validarán los lenguajes extranjeros. Como resumen se ofrece al final, la tabla 4.1 que resume las características más diferenciadoras de cada uno de estas pruebas.

4.7.1.1. Experimento 1

Prueba 1.A RAVDESS subdividido por género femenino.

Prueba 1.B RAVDESS subdividido por género masculino.

Prueba 1.C RAVDESS sin subdivisión.

4.7.1.2. Experimento 2

Prueba 2.A RAVDESS sin subdivisión con aumento de datos: Ruido Blanco.

Prueba 2.B RAVDESS sin subdivisión con aumento de datos: Desplazamiento.

Prueba 2.C RAVDESS sin subdivisión con aumento de datos: Modulación.

Prueba 2.D RAVDESS sin subdivisión con aumento de datos: Ruido Blanco, Desplazamiento y Modulación.

4.7.1.3. Experimento 3

Prueba 3.A Ensamblado con RAVDESS y TESS.

Prueba 3.B Ensamblado con SAVEE y TESS.

Prueba 3.C Ensamblado con RAVDESS, SAVEE y TESS.

4.7.1.4. Experimento 4

Prueba 4.A Ensamblado con RAVDESS y TESS con aumento de datos.

Prueba 4.B Ensamblado con SAVEE y TESS con aumento de datos.

Prueba 4.C Ensamblado con RAVDESS, SAVEE y TESS con aumento de datos.

4.7.1.5. Experimento 5

Prueba 5.A Ensamblado con SAVEE y TESS con modelo CNN usando espectrogramas.

4.7.1.6. Experimento 6

Prueba 6.A Ensamblado con SAVEE y TESS con modelo CNN-LSTM usando espectrogramas.

Prueba	**Datos Entr.** [1]	**Método** [2]	**Aumento** [3]	**Vol. Datos** [4]
1.A	RAVDESS (subdivisión femenina)	Modelo CNN 1D simple		528 muestras
1.B	RAVDESS (subdivisión masculina)	Modelo CNN 1D simple		528 muestras
1.C	RAVDESS	Modelo CNN 1D simple		1056 muestras
2.A	RAVDESS	Modelo CNN 1D	Ruido Blanco	2112 muestras
2.B	RAVDESS	Modelo CNN 1D	Desplazamiento	2112 muestras
2.C	RAVDESS	Modelo CNN 1D	Modulación	2112 muestras
2.D	RAVDESS	Modelo CNN 1D	Ruido Blanco, Desplazamiento y Modulación	4224 muestras
3.A	RAVDESS y TESS	Modelo CNN 1D		3056 muestras
3.B	TESS y SAVEE	Modelo CNN 1D		2824 muestras
3.C	RAVDESS, SAVEE y TESS	Modelo CNN 1D		3476 muestras
4.A	RAVDESS y TESS	Modelo CNN 1D	Modulación	6112 muestras
4.B	SAVEE y TESS	Modelo CNN 1D	Modulación	5648 muestras
4.C	RAVDESS, SAVEE y TESS	Modelo CNN 1D	Modulación	10112 muestras
5	SAVEE y TESS	Modelo CNN 2D con espectrogramas MFCC		2824 muestras
6	SAVEE y TESS	Modelo CNN-LSTM con espectrogramas MFCC		2824 muestras

Tabla 4.1: Resumen de las pruebas para la obtención de un modelo óptimo en la propia lengua (inglés)

[1] Se refiere a las bases de datos que se usarán tanto para el entrenamiento, validación y test.
[2] Arquitectura usada en la prueba.
[3] Técnica de aumento de datos que se usará.

4.7.2. Pruebas con lenguajes extranjeros

En estas pruebas se evaluarán los mejores modelos de la sección anterior, con lenguajes que no han sido vistos en su entrenamiento. Al final de este bloque se presenta la tabla 4.2 que resume los puntos mas característicos de las pruebas.

4.7.2.1. Experimento 7

Prueba 7.A Evaluación del dataset EMO-DB con el tercer mejor modelo.

Prueba 7.B Evaluación del dataset EMO-DB con el segundo mejor modelo.

Prueba 7.C Evaluación del dataset EMO-DB con el primer mejor modelo.

4.7.2.2. Experimento 8

Prueba 8.A Evaluación del dataset CaFE con el tercer mejor modelo.

Prueba 8.B Evaluación del dataset CaFE con el segundo mejor modelo.

Prueba 8.C Evaluación del dataset CaFE con el primer mejor modelo.

Prueba	**Modelo**	**Datos Testt**	**Aumento** [5]	**Vol. Datos**
7.A	Tercer mejor modelo	EMO-DB	Modulación	552
7.B	Segundo mejor modelo	EMO-DB	Modulación	552
7.C	Primer mejor modelo	EMO-DB	Modulación	552
8.A	Tercer mejor modelo	CaFE	Modulación	552
8.B	Segundo mejor modelo	CaFE	Modulación	552
8.C	Primer mejor modelo	CaFE	Modulación	552

Tabla 4.2: Resumen de las pruebas para la obtención de un modelo óptimo en la propia lengua (inglés)

[4]Volumen de datos con los que se trabajará.

[5]Este aumento se sólo se aplica a los datos de test con el fin de balancearlos.

5. Desarrollo de la comparativa

A continuación se presentan el resultado de los experimentos diseñados en el capítulo anterior. Acorde con la estructura que se le dio entonces, este capítulo se divide en dos partes: La primera sección presenta los datos de los experimentos que buscan un modelo óptimo que funcione en la lengua con la que es entrenado. Los primeros dos experimentos pueden parecer irrelevantes desde la perspectiva de una conclusión final, pero ayuda a entender el punto de partida y su progresión. La segunda sección resulta de la evaluación de los tres mejores modelos de la anterior con los idiomas extranjeros.

5.1. Búsqueda del mejor modelo

5.1.1. Experimento 1

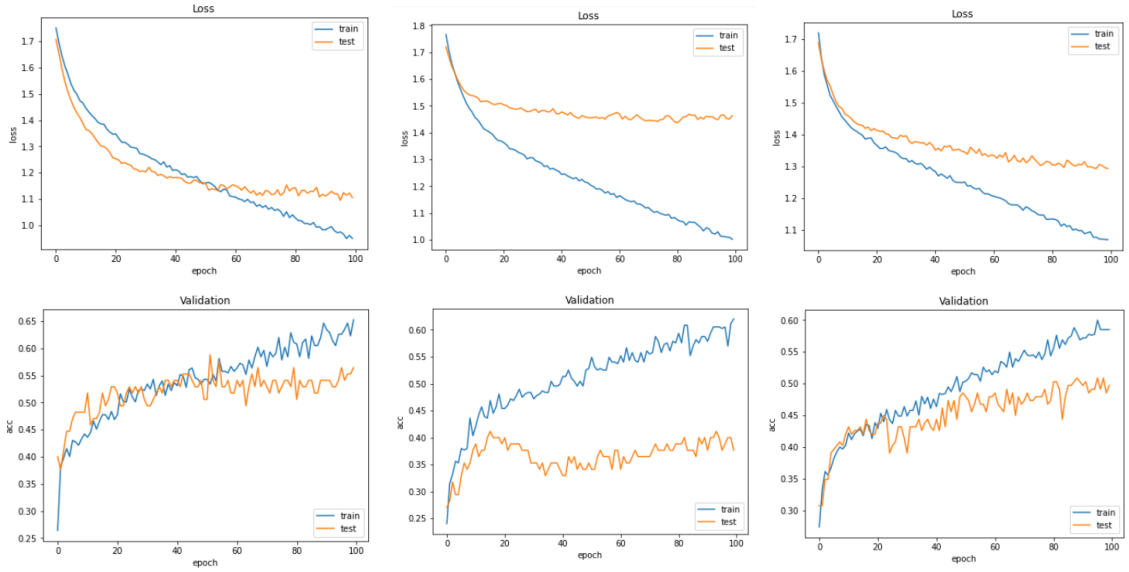
A continuación se exponen los resultados del experimento 1, que está formado por las pruebas 1.A, 1.B y 1.C.

Procesado de los datos La distribución de las clases en los conjuntos de datos usados para este experimento queda como se muestra en la tabla 5.1

Femenino		Masculino		RAVDESS sin dividir	
enfado	96	enfado	96	enfado	192
asco	96	asco	96	asco	192
miedo	96	miedo	96	miedo	192
felicidad	96	felicidad	96	felicidad	192
tristeza	96	tristeza	96	tristeza	192
neutral	48	neutral	48	neutral	96

Tabla 5.1: Distribución de los datos de las pruebas del experimento 1.

Entrenamiento El rendimiento de los distintos modelos se puede visualizar a través de la gráfica de loss y accuracy en la figura 5.1



Femenino Masculino RAVDESS sin dividir

Figura 5.1: Rendimiento en el resultado de las pruebas del experimento 1.

Evaluación En la tabla 5.2 se muestran las métricas de éxito por cada una de las clases.

Emoción	**RAVDES Fem.**		**RAVDES Masc.**		**RAVDES sin dividir**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
enfado	0.24	0.38	0.23	0.35	0.21	0.33
asco	1.00	0.42	0.75	0.44	0.62	0.33
miedo	0.80	0.55	0.60	0.53	0.76	0.59
felicidad	0.79	0.65	0.47	0.43	0.77	0.52
tristeza	0.17	0.08	0.00	0.00	0.42	0.14
neutral	1.00	0.18	0.00	0.00	0.20	0.06
accuracy	**56.47 %**		**38.65 %**		**50 %**	

Tabla 5.2: Resultado de la evaluación de las pruebas en el experimento 1.

5.1.2. Experimento 2

En este experimento se muestran los resultados de las pruebas 2.A, 2.B, 2.C y 2.D donde se exploran las diferentes técnicas de aumento de datos y cómo estas afectan al rendimiento del modelo.

Procesado de los datos La distribución de las clases en los conjuntos de datos usados para este experimento queda como se muestra en la tabla 5.3.

Ruido Blanco		Desplazamiento		Modulación		Los tres métodos	
enfado	384	enfado	384	enfado	384	enfado	384
asco	384	asco	384	asco	384	asco	384
miedo	384	miedo	384	miedo	384	miedo	384
felicidad	384	felicidad	384	felicidad	384	felicidad	384
tristeza	384	tristeza	384	tristeza	384	tristeza	384
neutral	192	neutral	192	neutral	192	neutral	192

Tabla 5.3: Distribución de los datos en las pruebas del experimento 2.

Entrenamiento El rendimiento de los distintos modelos se puede visualizar a través de la gráfica de loss y accuracy en la figura 5.2.

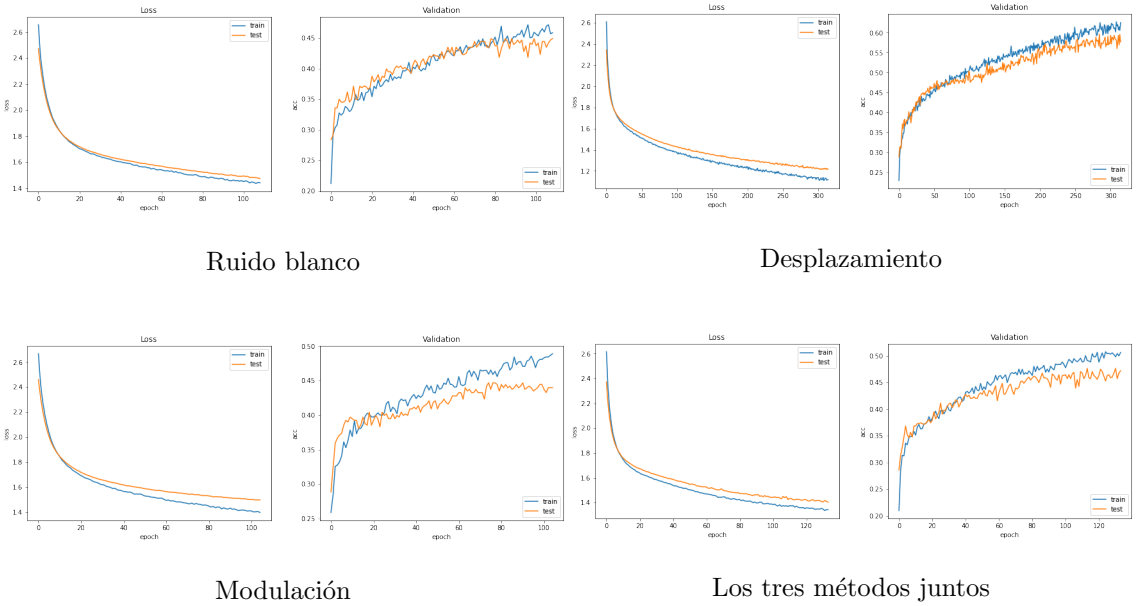


Figura 5.2: Rendimiento en las pruebas del experimento 2

Evaluación En la tabla 5.4 se muestran las métricas de éxito por cada una de las clases.

Emoción	**Ruido Blanco**		**Desplazamiento**		**Modulación**		**Tres métodos**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
enfado	0.19	0.32	0.22	0.36	1.00	0.59	0.20	0.33
asco	1.00	0.21	0.89	0.35	1.00	0.62	0.88	0.26
miedo	0.91	0.23	0.79	0.64	1.00	0.76	0.87	0.44
felicidad	0.75	0.07	0.95	0.39	0.98	0.70	0.73	0.21
tristeza	0.00	0.00	0.75	0.14	1.00	0.76	0.00	0.00
neutral	0.00	0.00	0.00	0.00	0.25	0.40	0.00	0.00
accuracy	**44.91 %**		**57.68 %**		**43.97 %**		**47.45 %**	

Tabla 5.4: Resultado de la evaluación de las pruebas del experimento 2.

5.1.3. Experimento 3

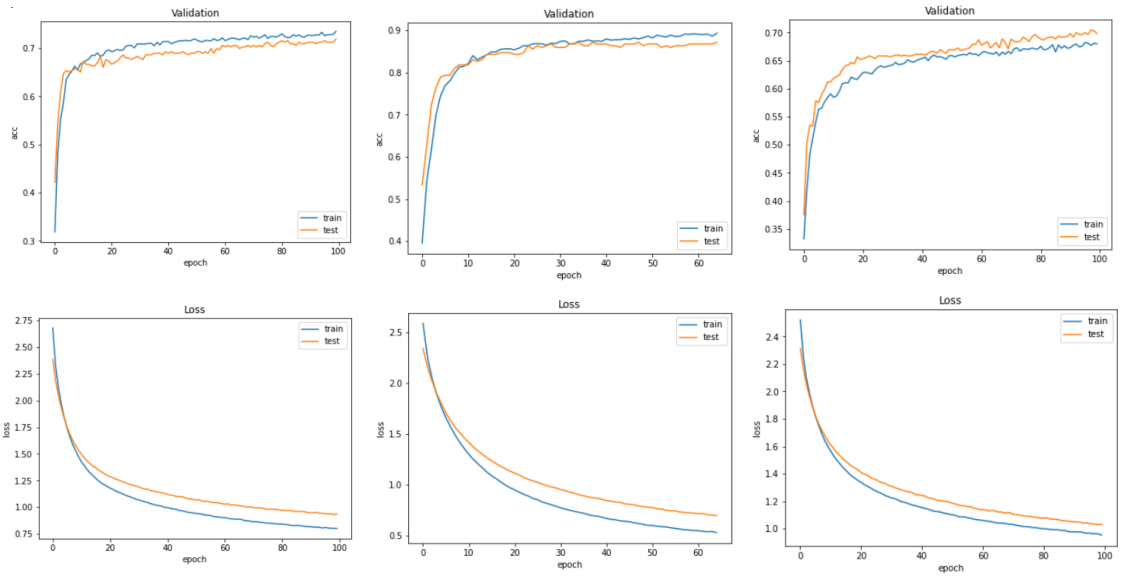
En este experimento se muestran los resultados de las pruebas 3.A (RAVDESS y TESS), 3.B (TESSy SAVEE), y 3.C (RAVDESS, TESS y SAVEE) donde se explora el comportamiento del modelo CNN 1D con diferentes combinaciones de bases de datos para aumentar el volumen de muestras.

Procesado de los datos La distribución de las clases en los conjuntos de datos usados para este experimento queda como se muestra en la tabla 5.5.

RAVDESS y TESS		TESS y SAVEE		RAVDESS, TESS y SAVEE	
enfado	592	enfado	460	enfado	652
asco	592	asco	460	asco	652
miedo	392	miedo	260	miedo	452
felicidad	592	felicidad	460	felicidad	652
tristeza	392	tristeza	260	tristeza	452
neutral	496	neutral	520	neutral	652

Tabla 5.5: Distribución de los datos en las pruebas del experimento 3.

Entrenamiento El rendimiento de los distintos modelos se puede visualizar a través de la gráfica de loss y accuracy en la figura 5.3.



RAVDESS y TESS SAVEE y TESS RAVDESS, SAVEE y TESS

Figura 5.3: Resultado del rendimiento de los modelos del experimento 3.

Evaluación En la tabla 5.6 se muestran las métricas de éxito por cada una de las clases.

Emoción	**RAVDES y TESS**		**TESS y SAVEE**		**RAVDES, TESS y SAVEE**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
enfado	0.38	0.55	0.51	0.66	0.32	0.48
asco	0.97	0.75	1.00	0.85	0.97	0.62
miedo	1.00	0.53	1.00	0.79	1.00	0.52
felicidad	1.00	0.74	0.97	0.89	1.00	0.80
tristeza	1.00	0.80	1.00	0.87	1.00	0.39
neutral	1.00	0.89	1.00	0.79	1.00	0.81
accuracy	**71.89 %**		**88.22 %**		**64.08 %**	

Tabla 5.6: Resultados de las pruebas del experimento 3.

5.1.4. Experimento 4

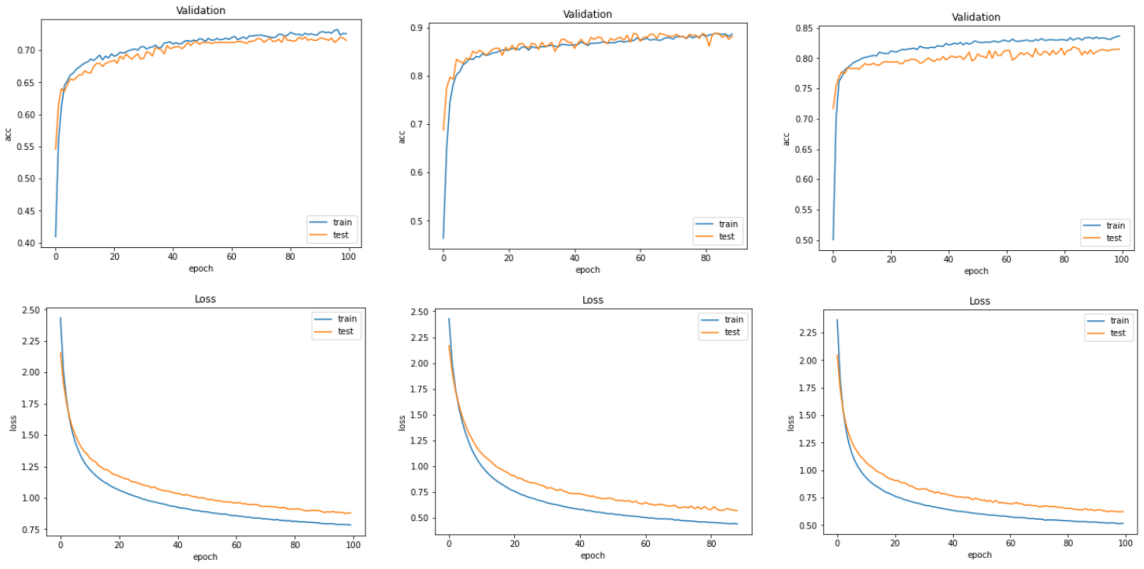
En este experimento se muestran los resultados de las pruebas 4.A, 4.B y 4.C donde se examina el comportamiento de la arquitectura y configuración del experimento 3 aplicando aumento de datos.

Procesado de los datos La distribución de las clases en los conjuntos de datos usados para este experimento queda como se muestra en la tabla 5.7.

RAVDESS y TESS		TESS y SAVEE		RAVDESS, TESS y SAVEE	
enfado	1184	enfado	920	enfado	1984
asco	1184	asco	920	asco	1984
miedo	784	miedo	520	miedo	1184
felicidad	1184	felicidad	920	felicidad	1984
tristeza	784	tristeza	520	tristeza	1184
neutral	992	neutral	1040	neutral	1792

Tabla 5.7: Distribución de los datos en las pruebas del experimento 4.

Entrenamiento El rendimiento de los distintos modelos se puede visualizar a través de la gráfica de loss y accuracy en la figura 5.4.



RAVDESS y TESS SAVEE y TESS RAVDESS, SAVEE y TESS

Figura 5.4: Resultado de los modelos en las pruebas del experimento 4.

Evaluación En la tabla 5.8 se muestra el accuracy y F1 por cada una de las clases en los tres modelos.

Emoción	**RAVDES y TESS**		**TESS y SAVEE**		**RAVDES, TESS y SAVEE**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
enfado	0.37	0.53	0.48	0.65	0.50	0.66
asco	0.97	0.77	0.92	0.83	0.96	0.84
miedo	1.00	0.51	1.00	0.86	0.96	0.71
felicidad	0.97	0.72	0.94	0.91	0.94	0.84
tristeza	1.00	0.61	1.00	0.79	1.00	0.82
neutral	1.00	0.91	1.00	0.85	1.00	0.95
accuracy	**71.56 %**		**88.19 %**		**81.81 %**	

Tabla 5.8: Resultado de los datos en las pruebas del experimento 4.

5.1.5. Experimento 5

A continuación se muestran los resultados del experimento 5, donde se trabajó con un modelo de redes convolucionales alimentado por espectrogramas.

Procesado de los datos La distribución de las clases en el conjunto de datos usado (SAVEE y TESS) para este experimento queda como se muestra en la tabla 5.9.

TESS y SAVEE	
enfado	460
asco	460
miedo	460
felicidad	460
tristeza	460
neutral	520

Tabla 5.9: Distribución de los datos en las pruebas del experimento 5.

Entrenamiento El rendimiento del correspondiente modelo se puede visualizar a través de la gráfica de loss y accuracy en la figura 5.5.

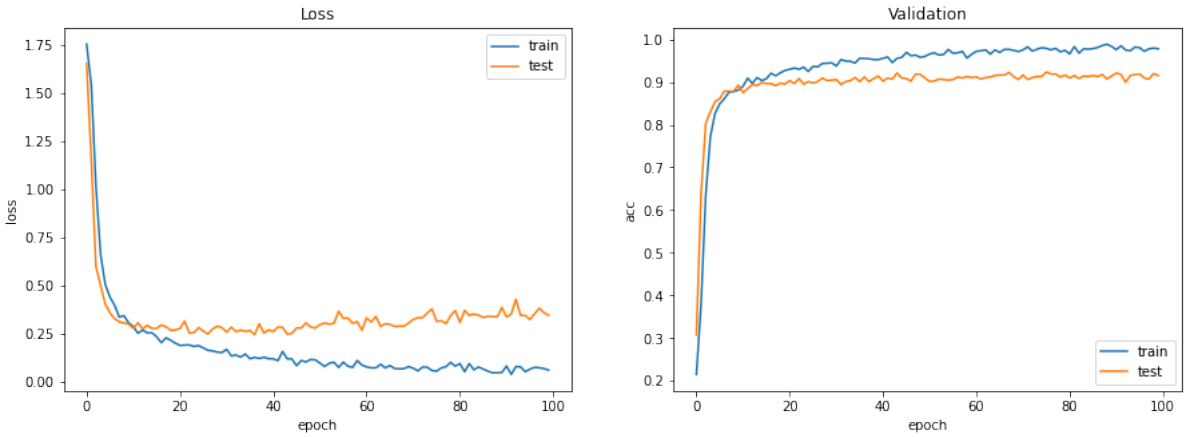


Figura 5.5: Rendimiento del modelo CNN 2D usando los datos de SAVEE y TESS

Evaluación En la tabla 5.10 se muestran las métricas de éxito por cada una de las clases.

Emoción	**CNN 2D**	
	Accuracy	F1
enfado	0.86	0.89
asco	0.90	0.91
miedo	0.94	0.93
felicidad	0.94	0.94
tristeza	0.96	0.92
neutral	0.88	0.89
accuracy	**91.50 %**	

Tabla 5.10: Resultados del experimento 5 usando una arquitectura CNN 2D.

5.1.6. Experimento 6

Se exponen los resultados del experimento 6 donde se trabajó con un modelo basado en una arquitectura de redes convolucionales y LSTM, alimentado por espectrogramas.

Procesado de los datos y Evaluación La distribución de las clases en el conjunto usado para este experimento (TESS y SAVEE) queda como se muestra en la tabla 5.11, mientras que en la tabla 5.12 se muestran las métricas de éxito por cada una de las clases.

TESS y SAVEE	
enfado	460
asco	460
miedo	460
felicidad	460
tristeza	460
neutral	520

Tabla 5.11: Distribución de los datos en las pruebas del experimento 6.

Emoción	**CNN-LSTM**	
	Accuracy	F1
enfado	0.96	0.94
asco	0.67	0.80
miedo	0.99	0.93
felicidad	1.00	0.95
tristeza	0.99	0.91
neutral	0.96	0.91
accuracy	**92.06 %**	

Tabla 5.12: Resultados del experimento 6 usando un arquitectura CNN-LSTM.

Entrenamiento El rendimiento del correspondiente modelo se puede visualizar a través de la gráfica de loss y accuracy en la figura 5.6.

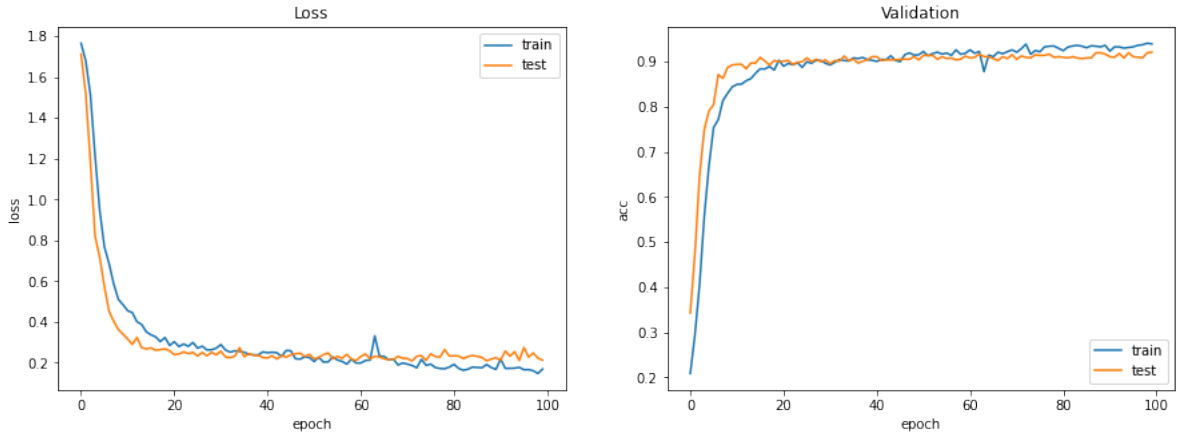


Figura 5.6: Rendimiento del modelo CNN-LSTM usando los datos de SAVEE y TESS

5.2. Pruebas con lenguajes extranjeros

5.2.1. Experimento 7

Se exponen los resultados del experimento 7 que evaluó los tres mejores modelos resultantes del bloque anterior en el idioma alemán, usando como conjunto de test la base de datos EMO-DB.

Preprocesado de datos La distribución de las clases en el conjunto de test usado para este experimento (EMO-DB) queda como se muestra en la tabla 5.13.

CNN 1D		CNN 2D		CNN-LSTM	
enfado	92	enfado	92	enfado	92
asco	92	asco	92	asco	92
miedo	92	miedo	92	miedo	92
felicidad	92	felicidad	92	felicidad	92
tristeza	92	tristeza	92	tristeza	92
neutral	92	neutral	92	neutral	92

Tabla 5.13: Distribución resultante de las clases en la base de datos EMO-DB.

Evaluación En la tabla 5.14 se muestran las métricas de éxito por cada una de las clases, en los tres modelos.

Emoción	**CNN 1D**		**CNN 2D**		**CNN-LSTM**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
enfado	0.16	0.26	0.17	0.18	0.18	0.21
asco	0.25	0.20	0.06	0.04	0.10	0.12
miedo	0.35	0.21	0.30	0.40	0.22	0.19
felicidad	0.56	0.36	0.21	0.25	0.24	0.26
tristeza	0.36	0.08	0.26	0.10	0.38	0.22
neutral	0.26	0.17	0.21	0.17	0.16	0.16
accuracy	**32 %**		**20 %**		**22 %**	

Tabla 5.14: Resultados de evaluar los mejores modelos en el idioma alemán.

5.2.2. Experimento 8

Se exponen los resultados del experimento 8 que evaluó los tres mejores modelos resultantes del bloque anterior en el idioma francés, usando como conjunto de test la base de datos CaFE.

Preprocesado de datos La distribución de las clases en el conjunto de test usado para este experimento (CaFE) queda como se muestra en la tabla 5.15.

CNN 1D		CNN 2D		CNN-LSTM	
enfado	92	enfado	92	enfado	92
asco	92	asco	92	asco	92
miedo	92	miedo	92	miedo	92
felicidad	92	felicidad	92	felicidad	92
tristeza	92	tristeza	92	tristeza	92
neutral	92	neutral	92	neutral	92

Tabla 5.15: Distribución resultante de las clases en la base de datos CaFE.

Evaluación En la tabla 5.16 se muestran las métricas de éxito por cada una de las clases, en los tres modelos.

Emoción	**CNN 1D**		**CNN 2D**		**CNN-LSTM**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
enfado	0.14	0.22	0.20	0.30	0.20	0.26
asco	0.26	0.15	0.17	0.13	0.21	0.22
miedo	0.00	0.00	0.13	0.12	0.11	0.10
felicidad	0.23	0.18	0.17	0.15	0.18	0.25
tristeza	0.35	0.11	0.14	0.05	0.21	0.11
neutral	0.10	0.04	0.18	0.10	0.35	0.11
accuracy	**18 %**		**18 %**		**21 %**	

Tabla 5.16: Resultados de evaluar los mejores modelos en el idioma francés.

6. Discusión y análisis de los resultados

El objetivo de los dos primeros experimentos fue probar el comportamiento del método elegido para el desarrollo inicial de las pruebas con el conjunto de datos RAVDESS. Aunque la observación de los datos llevó a la conclusión de que una separación por género tendría sentido, los datos no fueron consistentes, dando las dos divisiones, porcentajes de exactitud muy diferentes además de un enorme desajuste. Respecto al segundo experimento, formado por las pruebas 2.A, 2B, 2.C y 2.D trataron de comparar el resultado aplicando diferentes técnicas de aumento de datos al conjunto de datos RAVDESS, que si bien los frutos de esta prueba no arrojaron unos resultados satisfactorios, condujeron a una sólida reducción del overfitting. No se pasa por alto, el hecho de que el uso de ruido blanco llega a ser contraproducente, debido principalmente a la incapacidad de reconocer la Tristeza y la Neutralidad, y por mostrar una enorme diferencia entre las métricas accuracy y F1. Aquí se comprueba la sensibilidad de las características MFCC al ruido de fondo. De la misma manera cabe mencionar que la Neutralidad en esta prueba tuvo menor representación, lo que claramente impactó en el resultado.
Comparando estos resultados con una configuración similar (clasificación de emociones en RAVDESS con redes convolucionales unidimensionales) en la revisión del estado del arte, se sabe que (Abdul Qayyum y col., 2019) consiguen un 81.63 % con una arquitectura de siete capas convolucionales unidimensionales usando una base de datos similar en inglés con siete emociones (SAVEE). Por otro lado, (Mustaqeem & Kwon, 2020) implementa una red CNN bidimensional de siete capas usando espectrogramas MFCC para llegar al 81.01 % de accuracy en RAVDESS.

Siguiendo por el experimento 3, el objetivo fue probar un enfoque distinto para aumentar el número de elementos con los que la red entrenará teniendo en cuenta que el modelo resultante se probaría finalmente con un idioma que no ha sido visto en el entrenamiento. Se propuso un ensamblado de conjuntos de datos en el mismo idioma, aportando una mayor diversidad a los datos, y más capacidad de aprendizaje a la red.
Paradójicamente, el que peor resultados logró fue la combinación con mayor número de características RAVEE-TESS-SAVEE, mientras que TESS y SAVEE con 2824 muestras, consiguieron un 87.19 % frente al 64.08 % de la combinación de los tres conjuntos de datos

con 3476 muestras.

El conjunto de datos que se había usado en los experimentos previos (RAVDESS) únicamente contempla 2 tipos de frases, lo que apunta a que el bajo rendimiento de esas pruebas se deba a la poca diversidad de los datos.

Cabe resaltar que el desempeño del modelo con redes convolucionales unidimensionales compite con los resultados reportados por (Abdul Qayyum y col., 2019), (Mustaqeem & Kwon, 2020), y (Anvarjon y col., 2020), los cuales con arquitecturas más complejas consiguen menor exactitud.

Analizando los resultados hasta ahora, da la sensación que la técnica de aumento de datos no consigue aportar demasiado, por lo que para corroborar esta hipótesis, el experimento 4 repitió esta técnica en la configuración del experimento anterior. Se pudo ver que los resultados no arrojaron demasiadas diferencias en comparación con su versión no aumentada, salvo en el combinación de los tres conjuntos (RAVDESS, TESS y SAVEE) donde se tuvo que casi triplicar el tamaño para que hubiera una mejoría del 17.73 % en el accuracy:

- RAVDESS y TESS pasaron de 3056 muestras con un 71.89 % de precisión, a 6112 instancias con 71.56 % en su versión aumentada.
- TESS y SAVEE pasaron de sólo 2824 muestras con un 88.22 % de precisión, a 5648 instancias con un 88.19 %.
- RAVDESS, TESS y SAVEE, sin embargo, pasaron de un 64.08 % con 3476 muestras a un 81.81 % con 10112 instancias.

Los experimentos 5 y 6 exploraron el uso de espectrogramas MFCC como estrategia. Este enfoque consiguió los mejores resultados: 92.06 % en CNN-LSTM en primer lugar y CNN con un 91.50 % en una clasificación de seis emociones en inglés.

Emoción	**CNN 1D**		**CNN 2D**		**CNN-LSTM**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
enfado	0.51	0.66	0.86	0.89	**0.96**	**0.94**
asco	0.92	0.83	0.90	0.91	**0.67**	**0.80**
miedo	1.00	0.86	0.94	0.93	**0.99**	**0.93**
felicidad	0.94	0.91	0.94	0.94	**1.00**	**0.95**
tristeza	1.00	0.79	0.96	0.92	**0.99**	**0.91**
neutral	1.00	0.85	0.88	0.89	**0.96**	**0.91**
accuracy	**88.22 %**		**90.50 %**		**92.06 %**	

Tabla 6.1: Comparación de los tres mejores modelos resultantes en el idioma inglés.

La tabla 6.1 compara los resultados de los tres mejores modelos de los experimentos realizados. La efectividad del uso de espectrogramas en un clasificador mono-lingüístico también se refleja en cómo de estables son los valores de accuracy y F1, ya que presentan menor variación entre ellos.

En la tabla 6.2 se expone una comparación de los principales trabajos que se han llevado a cabo en **el idioma inglés** con el mejor modelo propuesto (CNN-LSTM).

Emoción	**Abdul 2020** SAVEE		**Mustaqeem 2020** IEMOCAP		**Anvarjon 2020** IEMOCAP		**Modelo propuesto** SAVEE y TESS	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
enfado	0.85	0.79	0.87	0.80	0.68	0.76	**0.96**	**0.94**
asco	0.82	0.82	-	-	-	-	**0.67**	**0.80**
miedo	0.85	0.82	-	-	-	-	**0.99**	**0.93**
felicidad	0.83	0.85	0.97	0.91	0.85	0.73	**1.00**	**0.95**
tristeza	0.83	0.86	0.82	0.84	0.74	0.75	**0.99**	**0.91**
sorpresa	0.84	0.86	-	-	-	-	-	-
neutral	0.83	0.85	0.77	0.83	0.81	0.80	**0.96**	**0.91**
accuracy	**84.01 %**		**84 %**		**77.01 %**		**92.06 %**	

Tabla 6.2: Comparación de los trabajos presentados en la revisión de la literatura en el idioma inglés.

Un aspecto que llama la atención, es que en todas las pruebas que se hicieron en un modelo basado únicamente en redes convolucionales entrenado y evaluado en inglés, el Enfado fue la emoción más complicada de distinguir. Esto sugiere un patrón para este lenguaje, lo que encaja con los resultados encontrados en la revisión del estado del arte.
Se han extraído los espectrogramas de Mel de seis emociones pertenecientes a SAVEE (figura 6.1) y se comparan con las observaciones hechas en (Huang y col., 2013).

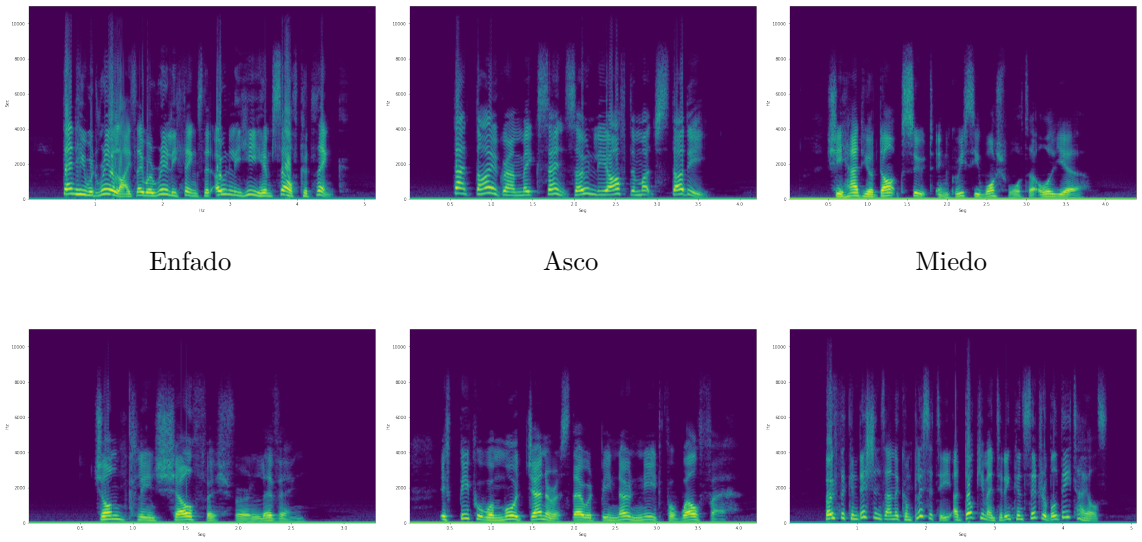


Figura 6.1: Espectrogramas de Mel de las emociones pertenecientes a SAVEE. Fuente Propia

La figura 6.1 muestra la representación de la banda del espectro de frecuencia de las seis emociones trabajadas en este estudio procedentes de la misma base de datos (SAVEE). Estos espectrogramas visualizan el cambio de frecuencia de una señal no estacionaria ("Spectrogram SciPy v1.7.0 Manual", 2008).

Posteriormente la figura 6.2 muestra las variaciones en la intensidad y frecuencia de algunas de las emociones que forman la base de datos TESS.

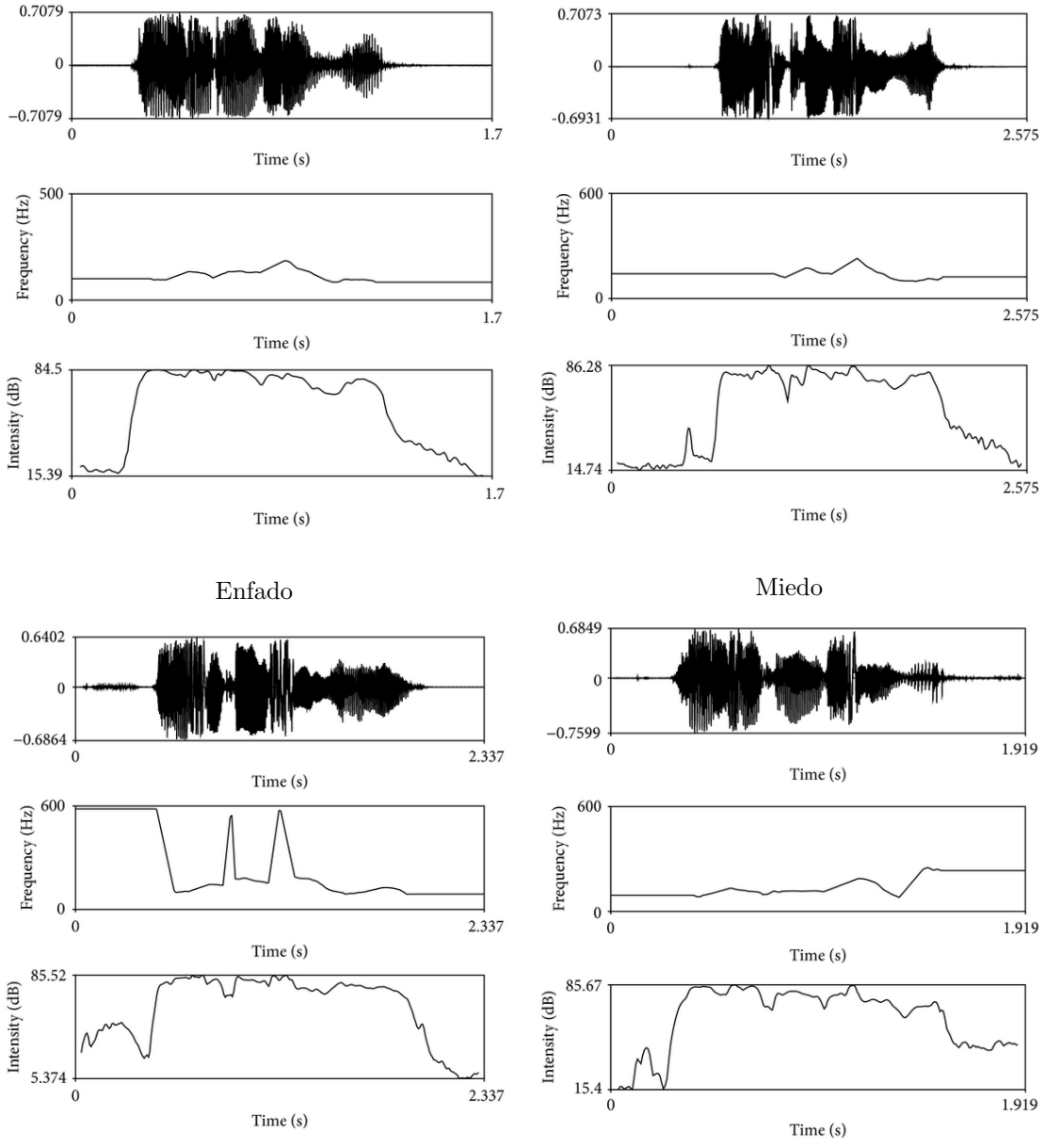


Figura 6.2: Onda acústica, frecuencia e intensidad de cuatro emociones en TESS. Fuente: (Huang y col., 2013)

Respecto a las figuras sobre estas líneas se resalta lo siguiente:

- El Miedo y la Felicidad tienen una duración más corta, mientras que el Enfado, la Tristeza y la Neutralidad se expanden más en el tiempo (en la figura 6.1, menos zonas oscuras en el eje X).
- El Enfado y el Miedo tienen mayor amplitud, que se puede ver en el en la represen-

tación de la onda acústica de cada emoción en la figura 6.2.

- La emoción del Enfado mantiene más constante la frecuencia en la línea de tiempo y su variación en la intensidad es más suave en comparación con las otras emociones (figura 6.2).

Estas observaciones parecen indicar que el factor más distinguible respecto al Enfado es cómo sus características varían en el tiempo, lo que tiene sentido con el hecho de que este patrón no se continúe en los resultados del experimento 6 con el modelo CNN-LSTM, ya que este sistema permite aprender la relevancia temporal de cada secuencia del habla.
Un enfoque basado puramente en redes convolucionales aprende a diferenciar patrones a través del espacio. Sin embargo la señal acústica es continua en el domino de tiempo, de forma que las características emocionales percibidas en cada segmento no pueden ser tratadas de manera aislada. LSTM resuelve ese problema aumentando la información entre esas ventanas de tiempo, lo cual ayuda a reflejar una continuidad temporal de las características.

La tabla 6.3 compara los tres mejores trabajos (con cualquier lenguaje) expuestos en la revisión del estado del arte con el mejor modelo propuesto (CNN-LSTM).

Trabajo	**Método**	**Datos usados**	**Acierto**
Anvarjon, 2020	CNN 2D + espect. Mel	EMOD-DB	92.01 %
Harar, 2017	CNN	EMOD-DB (3 emociones)	96.97 %
Tamulevicius, 2020	CNN 2D + cocleogramas	Lithuanian	97.00 %
Modelo propuesto	CNN-LSTM + espect. MFCC	TESS+SAVEE	**92.06 %**

Tabla 6.3: Comparación de los tres mejores trabajos presentados en la revisión de la literatura con el modelo propuesto.

Finalmente, tal y como se predijo para los experimentos 7 y 8, se obtuvo un porcentaje de accuracy un poco mayor en la evaluación con alemán (lengua con raíces fonéticas más próximas al idioma que se usó en el entrenamiento), que en la evaluación con francés (lengua con raíces fonéticas más lejanas al idioma que se usó en el entrenamiento). En el experimento 7 se evaluó el rendimiento de los tres mejores modelos en el idioma alemán utilizando como test la base de datos EMO-DB. En un primer momento se pudo observar en el accuracy el mismo patrón sobre la dificultad para reconocer la emoción del Enfado,

pero comprobando la métrica de F1 en el modelo CNN 1D, las emociones Neutral y Tristeza bajan de un 26 % a un 17 % y de un 36 % a un 8 % respectivamente. La exactitud general de las tres pruebas, tampoco mantiene la misma lógica que los modelos evaluados en inglés, ya que aquí fue el modelo basado en redes convolucionales unidimensionales, el que mejor porcentaje de exactitud consiguió.

No se pudieron percibir las mismas observaciones en el experimento 8, que evaluaba el idioma francés usando como conjunto de test la base de datos CaFE en los tres mejores modelos entrenados y evaluados en inglés. Por un lado, no fue posible encontrar un patrón sobre la emoción más difícil (o más fácil) de reconocer y por otro el modelo CNN-LSTM tuvo mayor porcentaje de exactitud con un 21 %.

En la tabla 6.4 se comparan los resultados de este trabajo con el de (Tamulevicius y col., 2020) tras haber evaluado un modelo entrenado en una lengua con otras distintas (con espectrogramas y con cocleogramas).

Idioma Test	**Tamulevicius 2020** espectrogramas		**Tamulevicius 2020** cocleogramas		**Modelo propuesto**	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Serbio	0.37	0.18	0.41	0.36	-	-
Polaco	0.21	0.17	0.20	0.19	-	-
Alemán	0.49	0.27	0.42	0.40	**0.19**	**0.19**
Español	0.3	0.19	0.35	0.20	-	-
Francés	-	-	-	-	**0.20**	**0.18**

Tabla 6.4: Comparación de los modelos evaluados en lenguas extranjeras.

Tamulevicius evalúa en las lenguas serbio, polaco, alemán y español un modelo que ha aprendido lituano. El lituano, usado como idioma de referencia es una lengua con raíces bálticas, las cuales mantienen cierto grado de similitud fonética con las lenguas eslavas (Kortlandt, 1983), y sin embargo el polaco (lengua eslava), no es donde consigue un porcentaje de acierto más elevado. Atendiendo a los datos que se muestran en la tabla 6.4, no parece que se pueda establecer una relación de proximidad fonética entre distintos idiomas para la clasificación de emociones en la lengua extranjera siguiendo un enfoque basado en redes convolucionales.

Por último, se desea analizar el hecho de que el modelo no fuera capaz de abstraer la infor-

mación extraída de las características para clasificar emociones en otras lenguas. Haciendo una revisión de lo aprendido en este trabajo, se hacen algunas observaciones sobre por qué, usando espectrogramas que han resultado ser tan exitosos para la clasificación en una sola lengua, no se comportan de la misma manera en otros idiomas:

- Los objetos visuales y los sonidos no se agrupan en una imagen de la misma manera. Cuando en una imagen hay un píxel de un determinado color se puede asumir que pertenece a un determinado objeto. Por el contrario, los cambios instantáneos en las características del sonido como el tono o la intensidad, no se separan en capas distinguibles en un espectrograma. Esto quiere decir que no se puede asumir que una determinada frecuencia representada en un espectrograma pertenezca a un determinado tipo de sonido (por ejemplo, podría estar producida por la interacción de varias ondas). En definitiva, esto hace que separar sonidos simultáneos en espectrogramas tal y como lo se hace con objetos opacos en imágenes, sea especialmente difícil (Wyse, 2017). En la figura 6.3 se muestran tres emociones pertenecientes a las bases de datos que se han usado en los experimentos. Como puede verse, es difícil establecer un patrón común entre los idiomas para la misma emoción.

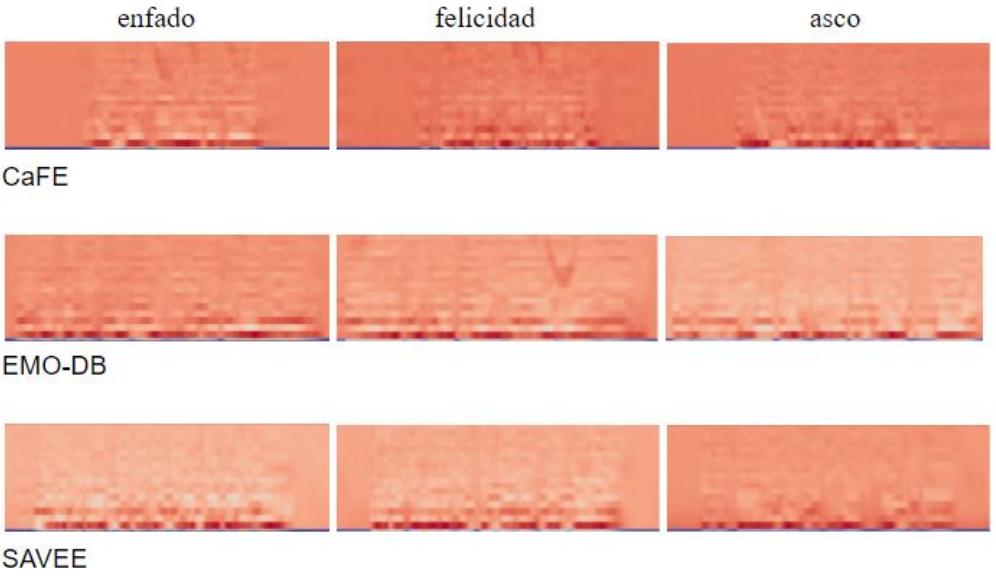


Figura 6.3: Espectrogramas de tres emociones en tres bases de datos distintas usadas en este trabajo. Fuente: Propia.

- Los ejes X e Y de una imagen no tienen el mismo significado que en un espectrograma. Por ejemplo, un perro en una foto, seguiría siendo un perro independientemente

si se mueve horizontal o verticalmente en la imagen; Es decir, la información representada **no cambia su significado**. Sin embargo esto no ocurre de la misma manera en un espectrograma donde estas dos dimensiones representan unidades distintas: frecuencia y tiempo (Verma & Smith, 2018).

- Desde el punto de vista humano, la forma en la que se procesan esas señales no es comparable. A la hora de localizar un objeto de manera visual en el entorno, se escanea lo que hay alrededor varias veces, ya que los objetos que lo conforman son estáticos. Por otro lado, un sonido toma la forma física de la presión manifestada en la onda, y desde el punto de vista de quien la percibe, esa determinada onda con un determinado estado sólo existe en un momento específico. Dicho de otro modo, la información que contiene una onda de audio está dispuesta de manera secuencial (Jonathan Hui, 2019).

7. Conclusiones y Trabajo Futuro

Para finalizar esta tesis de máster, se reúnen a continuación las conclusiones en base a los objetivos establecidos en el capítulo 3.

Respecto al primer objetivo específico, el cual consistía en una revisión en profundidad de la literatura actual, se abordó en el capítulo 2. Se llegó a la conclusión de que no existía una línea definida sobre qué métodos, técnicas y datos eran los más adecuados, para abordar este problema. Añadir además que, si bien el número de artículos sobre el reconocimiento de emociones en una lengua es abundante, no existe un oferta tan extensa para abordar el problema del reconocimiento de emociones en la lengua extranjera. Debido a esto, se decidió apostar por las técnicas que más apoyo tenían por parte de la literatura.

De esta manera, una vez se reunieron los conjuntos de datos que atendían a las condiciones establecidas en el capítulo 3, se pasó al siguiente punto de los objetivos específicos donde se debía diseñar una solución cuyo porcentaje de exactitud fuese superior a un 81 %. Los tres mejores modelos desarrollados en este trabajo consiguieron una puntuación por encima de la marca que se propuso, donde la diversidad de datos jugó un papel determinante. Esta estrategia no sólo redujo el overfitting, si no que hizo posible un modelo más rentable, ya que en comparación a las arquitecturas de otros trabajos se logró mejor porcentaje de acierto con diseños más simples. En el desarrollo de los modelos se comprobó que la generación de datos sintéticos podía ser contraproducente (especialmente con ruido blanco). Por otra parte, quedó clara la superioridad del uso de espectrogramas para un clasificador de emociones mono-lenguaje, especialmente con una arquitectura híbrida CNN-LSTM ya que tiene en cuenta la información temporal inherente a la señal de audio.

Por lo que respecta a los experimentos hechos evaluando una lengua extranjera y atendiendo al último punto de los objetivos, se ha llegado a las mismas conclusiones que Tamulevicius (Tamulevicius y col., 2020) donde tras hacer diversas pruebas con diferentes idiomas, únicamente puede reconocer aquellos con los que entrena su modelo. No obstante, tras extraer lo aprendido en este trabajo, se desaconseja tratar la señal acústica con imágenes estáticas (espectrogramas) por no adaptarse correctamente a cómo estas funcionan y perder información importante.

La estrategia que se planteó donde las lenguas extranjeras se dividían según su similitud fonética parece no ser muy relevante coincidiendo con (Pell & Skorup, 2008), donde afirma

que para reconocer el estado emocional en una lengua no aprendida se necesita mayor exposición a esta.

Debido a los ajustados tiempos de entrega, el alcance de este proyecto se ha debido simplificar, por lo que se plantean líneas de trabajo futuras:
En este trabajo se han explorado el reconocimiento de emociones en la voz usando características cepstrales, ya que revisando la literatura, se asumió que eran de las que mejor funcionaban. Aunque los resultados del clasificador sobre una sola lengua estuvieron a la altura de otros estudios, atendiendo al análisis previo en este mismo capítulo se considera que esta no es la mejor manera de abordar ese problema.
Por un lado se cree que sería recomendable no tratar las emociones como categorías discretas, ya que varían enormemente de un lenguaje a otro. Esto requiere un análisis más en profundidad sobre la fonética en el idioma en cuanto a la expresión emocional para conseguir una mayor independencia cultural.
Por otro lado, y aunque el tiempo dedicado a la tesis no lo ha permitido, sería interesante explorar la combinación mecanismos de atención junto al punto anterior, ya que podrían computar segmentos relevantes de la señal de audio y conseguir así mejor generalización.

Bibliografía

Abdul Qayyum, A. B., Arefeen, A. & Shahnaz, C. (2019). Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. *2019 IEEE International Conference on Signal Processing, Information, Communication and Systems, SPICSCON 2019*, (November), 122-125. https://doi.org/10.1109/SPICSCON48833.2019.9065172

Africa, A. D. M., Tabalan, A. R. V. & Tan, M. A. A. (2020). Speech emotion recognition using support vector machines. *International Journal of Emerging Trends in Engineering Research*, *8*(4), 1212-1216. https://doi.org/10.30534/ijeter/2020/43842020

Anvarjon, T., Mustaqeem & Kwon, S. (2020). Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors (Switzerland)*, *20*(18), 1-16. https://doi.org/10.3390/s20185212

Arnold, S. E. (2017). *Bradley Metrock and the Alexa Conference: Alexa As a Game Changer for Search and Publishing : Stephen E. Arnold @ Beyond Search.* http://arnoldit.com/wordpress/2017/02/02/bradley-metrock-and-the-alexa-conference-alexa-as-a-game-changer-for-search-and-publishing/

Atmaja, B. T. & Akagi, M. (2019). Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model. *Proceedings - 2019 IEEE International Conference on Signals and Systems, ICSigSys 2019*, (July), 40-44. https://doi.org/10.1109/ICSIGSYS.2019.8811080

Ayadi, M. E., Kamel, M. S. & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*, 572-587. https://doi.org/10.1016/J.PATCOG.2010.09.020

Bao, M. & Huang, A. (2019). Human vocal sentiment analysis. *arXiv*, 1-16.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. & Weiss, B. (2005). A database of German emotional speech. *9th European Conference on Speech Communication and Technology*, 1517-1520.

Chen, S., Zhu, Y. & Wayland, R. (2017). Effects of stimulus duration and vowel quality in cross-linguistic categorical perception of pitch directions. *PLoS ONE*, *12*, e0180656. https://doi.org/10.1371/journal.pone.0180656

Collier, Z. (2016). *The Story of the Capital One Alexa Skill : Alexa Blogs*. https://developer.amazon.com/es/blogs/alexa/post/c70e3a9b-405c-4fe1-bc20-bc0519d48c97/the-story-of-the-capital-one-alexa-skill

D. G. Childers, R. C. K., D. P. Skinner. (1977). The Cepstrum: A Guide to Processing. *Proceedings of the IEEE*, *65*, 1428-1443.

Davis, S. B. & Mermelstein, P. (1980). *Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences* (inf. téc.).

Davletcharova, A., Sugathan, S., Abraham, B. & James, A. P. (2015). Detection and Analysis of Emotion from Speech Signals. *Procedia Computer Science*, *58*, 91-96. https://doi.org/10.1016/j.procs.2015.08.032

Dellaert, F., Polzin, T. & Waibel, A. (1996). Recognizing emotion in speech. *International Conference on Spoken Language Processing, ICSLP, Proceedings*, *3*(October), 1970-1973. https://doi.org/10.1109/icslp.1996.608022

Effron, L. (2011). *iPhone 4S's Siri Is Lost in Translation With Heavy Accents - ABC News*. https://abcnews.go.com/Technology/siri-lost-translation-heavy-accents/story?id=14834111

Ekman, P. (1992). An Argument for Basic Emotion. http://www.paulekman.com/wp-content/uploads/2009/02/Universality-Of-Emotional-Expression-A-personal-History.pdf

Farouk, M. H. (2014). Emotion Recognition from Speech. *SpringerBriefs in Speech Technology*, 31-32. https://doi.org/10.1007/978-3-319-02732-6_7

Fast Fourier Transformation FFT - Basics. (2016). https://www.nti-audio.com/en/support/know-how/fast-fourier-transform-fft

Fayek, H. (2016). *Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs)*. https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

Fennell, B. (2001). *A History of English: A Sociolinguistic Approach*. Blackwell Publishing.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning* [http://www.deeplearningbook.org]. MIT Press.

Gournay, P., Lahaie, O. & Lefebvre, R. (2018). A canadian French emotional speech dataset. *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018*, 399-402. https://doi.org/10.1145/3204949.3208121

Harar, P., Burget, R. & Dutta, M. K. (2017). Speech emotion recognition with deep learning. *2017 4th International Conference on Signal Processing and Integrated Networks, SPIN 2017*, (February 2017), 137-140. https://doi.org/10.1109/SPIN.2017.8049931

Hautala, L. (2015). *Google Home will go on sale today, shipping November 4*. https://techcrunch.com/2016/10/04/say-hello-to-google-home/

Hellbernd, N. & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, *88*, 70-86. https://doi.org/10.1016/j.jml.2016.01.001

Huang, C., Liang, R., Wang, Q., Xi, J., Zha, C. & Zhao, L. (2013). Practical speech emotion recognition based on online learning: From acted data to elicited data. *Mathematical Problems in Engineering*, *2013*. https://doi.org/10.1155/2013/265819

IBM. (2020). What is Speech Recognition? https://www.ibm.com/cloud/learn/speech-recognition

Introducing OTO Systems Inc. (2018). https://otosystemsinc.medium.com/introducing-oto-99199a9b2c1b

Jackson, P. & ul haq, S. (2011). *Surrey Audio-Visual Expressed Emotion (SAVEE) database*.

Jain, M., Narayan, S., Balaji, P., P, B. K., Bhowmick, A., R, K. & Muthu, R. K. (2020). Speech Emotion Recognition using Support Vector Machine. *International Journal of Emerging Trends in Engineering Research*, *8*(4), 1212-1216. http://arxiv.org/abs/2002.07590

Jet, J. (2017). *Using Amazon's Alexa for Travel*. https://www.huffpost.com/entry/using-amazons-alexa-for-travel_b_59751ff7e4b06b511b02c47f

Jonathan Hui. (2019). Speech Recognition - Phonetics. https://jonathan-hui.medium.com/speech-recognition-phonetics-d761ea1710c0

Kartik, C. (2020). *Understanding Audio data, Fourier Transform, FFT and Spectrogram features for a Speech Recognition System*. https://dropsofai.com/understanding-audio-data-fourier-transform-fft-and-spectrogram-features-for-a-speech-recognition-system/

Khayam, S. A. (2003). *Wayback Machine*.

Koolagudi, S. G. & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*(2), 99-117. https://doi.org/10.1007/s10772-011-9125-1

Kortlandt, F. (1983). FROM PROTO-INDO-EUROPEAN TO SLAVIC.

Kumar, A. & Iqbal, J. L. M. (2019). Machine Learning Based Emotion Recognition using Speech Signal. *International Journal of Engineering and Advanced Technology*, *9*(1S5), 295-301. https://doi.org/10.35940/ijeat.a1068.1291s52019

Langari, S., Marvi, H. & Zahedi, M. (2020). Efficient speech emotion recognition using modified feature extraction. *Informatics in Medicine Unlocked*, *20*, 100424. https://doi.org/10.1016/j.imu.2020.100424

Larman, C. & Basili, V. R. (2003). Iterative and incremental development: A brief history. https://doi.org/10.1109/MC.2003.1204375

Lee, J. & Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2015-January*, 1537-1540.

Lim, W., Jang, D. & Lee, T. (2017). Speech emotion recognition using convolutional and Recurrent Neural Networks. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2016*. https://doi.org/10.1109/APSIPA.2016.7820699

Lisetti, C. L. (1998). Affective computing. *Pattern Analysis and Applications*, *1*(1), 71-73. https://doi.org/10.1007/bf01238028

Livingstone, S. & Russo, F. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE*, *13*. https://doi.org/10.5281/zenodo.1188976

Matthew, R. (2019). *Deep Learning Neurons versus Biological Neurons*. https://towardsdatascience.com/deep-learning-versus-biological-neurons-floating-point-numbers-spikes-and-neurotransmitters-6eebfa3390e9

McFee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E. & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python, 18-24. https://doi.org/10.25080/Majora-7b98e3ed-003

Mizak, A., Park, M., Park, D. & Olson, K. (2017). Amazon .Alexa"Pilot Analysis Report Front Porch Center for Innovation and Wellbeing.

Mocherman, A. (2015). *Nuance to Acquire Loquendo*. https://web.archive.org/web/20150521135427/http://www.nuance.com/company/news-room/press-releases/Press-Release---Nuance-to-Acquire-Loquendo_FINAL-v2.doc

Mustaqeem & Kwon, S. (2020). A CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors (Switzerland)*, *20*(1). https://doi.org/10.3390/s20010183

Nair, P. (2018). The dummy's guide to MFCC — Medium. https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

Nielsen, M. (2015). Neural Networks and Deep Learning. *The Machine Age of Customer Insight*, 91-101. https://doi.org/10.1108/978-1-83909-694-520211010

Nuance. (2019). *Nuance - Conversational AI for Healthcare and Customer Engagement — Nuance*. https://www.nuance.com/index.html

OTO-Systems. (2020). Voice AI: From Personal Assistants and Beyond. https://www.oto.ai/blog/voice-ai-from-personal-assistants-and-beyond

Pell, M. D. (2001). Influence of emotion and focus location on prosody in matched statements and questions. *The Journal of the Acoustical Society of America*, *109*(4), 1668-1680. https://doi.org/10.1121/1.1352088

Pell, M. D., Monetta, L., Paulmann, S. & Kotz, S. A. (2009). Recognizing Emotions in a Foreign Language. *Journal of Nonverbal Behavior*, *33*(2), 107-120. https://doi.org/10.1007/s10919-008-0065-7

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A. & Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: A comparison of four languages. *Journal of Phonetics*, *37*(4), 417-435. https://doi.org/10.1016/j.wocn.2009.07.005

Pell, M. D. & Skorup, V. (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication*, *50*(6), 519-530. https://doi.org/10.1016/j.specom.2008.03.006

Pichora-Fuller, M. K. & Dupuis, K. (2020). *Toronto emotional speech set (TESS)*. https://doi.org/10.5683/SP2/E8H2MF

Plutchik, R. (2001). The Nature of Emotions. https://www.jstor.org/stable/27857503?seq=1

Poorjam, A. H. (2018). *Why we take only 12-13 MFCC coefficients in feature extraction?*

Ram, R., Kumar Palo, H. & Mohanty, N. (2013). Emotion Recognition with Speech for Call Centres using LPC and Spectral Analysis. *Article in International Journal*

of Advanced Computer Research International Journal of Advanced Computer Research, (3), 2249-7277. https://www.researchgate.net/publication/299563254

Rana, M. & Miglani, S. (2014). Performance Analysis of MFCC and LPCC Techniques in Automatic Speech Recognition. *International Journal Of Engineering And Computer Science*, *3*(7727), 7727-7732.

Rashid, S. A. A. & Alang, N. K. (2018). Some Commonly Used Speech Feature Extraction Algorithms. (tourism), 13. https://www.intechopen.com/books/advanced-biometric-technologies/liveness-detection-in-biometrics

Rowe, A. (2018). *Interactive Storytelling App Novel Effect Just Raised A 3M dollar Series A*. https://www.forbes.com/sites/adamrowe1/2018/05/22/storytelling-app-novel-effect-3m-series-a/?sh=287bda11120d

Sandesara, A., Parikh, S., Sapovadiya, P. & Rahevar, M. (2020). A Comparative Study On Speech Emotion Recognition. *International Journal of Research in Engineering, Science and Management*, *3*(11), 25-35. https://doi.org/10.47607/ijresm.2020.366

Shriberg, E., Tsiartas, A., Smith, J. & Wagner, V. (2018). Crowdsourcing Emotional Speech. https://www.researchgate.net/publication/327805915_Crowdsourcing_Emotional_Speech

Smith, J. O. (2011). Spectral Audio Signal Processing. *Center for Computer Research in Music and Acoustics (CCRMA)*, 1-674.

Spectrogram SciPy v1.7.0 Manual. (2008). https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.spectrogram.html

Stevens, S. S. & Volkmann, J. (1940). The Relation of Pitch to Frequency: A Revised Scale. *The American Journal of Psychology*, *53*(3), 329-353. http://www.jstor.org/stable/1417526

Tamulevicius, G., Korvel, G., Yayak, A. B., Treigys, P., Bernataviciene, J. & Kostek, B. (2020). A study of cross-linguistic speech emotion recognition based on 2d feature spaces. *Electronics (Switzerland)*, *9*(10), 1-13. https://doi.org/10.3390/electronics9101725

Valverde-Albacete, F. J., Carrillo-de-Albornoz, J. & Pelaez-Moreno, C. (2013). A Proposal for New Evaluation Metrics and Result Visualization Technique for Sentiment Analysis Tasks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8138 LNCS*, 41-52. https://link.springer.com/chapter/10.1007/978-3-642-40802-1_5

Varshney, L. & Sun, J. (2013). Why do we perceive logarithmically? *Significance*, *10*. https://doi.org/10.1111/j.1740-9713.2013.00636.x

Verma, P. & Smith, J. O. (2018). Neural Style Transfer for Audio Spectrograms. https://youtu.be/UlwBsEigcdE

Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J. & Tarokh, V. (2020). Speech emotion recognition with dual-sequence LSTM architecture. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2020-May*, 6474-6478. https://doi.org/10.1109/ICASSP40776.2020.9054629

Wyse, L. (2017). Audio Spectrogram Representations for Processing with Convolutional Neural Networks. http://arxiv.org/abs/1706.09559

X. Huang, A. Acero & Hon, H. (2001). Spoken Language Processing: A guide to theory, algorithm, and system developmen.

A. Apéndices

Reconocimiento de Emociones en la Lengua no Aprendida

Luisa Sánchez Avivar

Universidad Internacional de la Rioja, Logroño (España)

Julio del 2021

RESUMEN

En este estudio se llevó a cabo un reconocimiento emocional de la voz multi-lingüístico. Para ello, se implementaron tres modelos distintos entrenados en inglés, y posteriormente fueron evaluados en dos lenguas extranjeras que no formaron parte del entrenamiento (francés y alemán). Las características cepstrales de la escala de Mel se extrajeron a partir de las muestras de audio y fueron usadas en los tres clasificadores con una arquitectura basada en redes convolucionales. El uso de espectrogramas en una arquitectura híbrida de redes convolucionales y LSTM se mostró superior frente a los otros, consiguiendo un 92.06 % de exactitud en una clasificación monolingüística. Por otro lado, la clasificación multi-lingüística no arrojó resultados satisfactorios aplicando el mismo método.

PALABRAS CLAVE

CNN-LSTM, MFCC, Reconocimiento de emociones en la voz, Señal acústica

I. INTRODUCCIÓN

El espectro emocional que una persona esconde en su discurso es un factor esencial de la comunicación humana y ofrece información adicional sin alterar el contenido lingüístico. Las tecnologías orientadas a convertir la voz en texto (*speech to text*) no tienen una forma segura de medir la calidad del diálogo de su interlocutor, impactando en negocios que hacen uso de estos avances (por ejemplo, centros de atención al cliente donde miden su grado de satisfacción).

La importancia de la interacción con las máquinas a través de comandos de voz, se ha visto acentuada gracias a la aparición de asistentes inteligentes como Siri en Apple [7] o Alexa en Amazon [3] , que han explotado las diferentes áreas del análisis de la voz con el objetivo de mejorar la experiencia de usuario. Sin embargo, a pesar de los avances tecnológicos, estos asistentes de voz normalmente carecen de la habilidad de reconocer el estado emocional del usuario, y cerrar esta brecha podría ser un gran avance en las industrias ya mencionadas.

Algunas compañías privadas ya han decidido a integrar el reconocimiento de emociones con técnicas del procesamiento del lenguaje natural [20], permitiendo mayor eficiencia del procesamiento de la conversación al detectar -por ejemplo- irritabilidad o frustración en el usuario. Estos ejemplos, impulsan la motivación de crear un sistema capaz de crear una respuesta, no sólo coherente en el plano semántico, sino también sensible al estado emocional del usuario.

Con este estudio se pretende entender mejor la relación entre emoción e idioma y arrojar luz a preguntas como ¿Hay emociones que son más fácilmente reconocibles indistintamente del lenguaje? ¿Hay lenguas donde es más fácil reconocer ciertas emociones? ¿Cómo influye la elección de la base de datos?¿Plantean las técnicas más populares un enfoque adecuado?.

II. ESTADO DEL ARTE

El reconocimiento de emociones en el habla es una disciplina en inteligencia artificial que trata de reconocer y clasificar emociones a través de la señal de voz. Originalmente este tema se ha planteado desde la psicología [6] [8] y ha seguido por estudios que relacionan las emociones con las propiedades fonéticas en el habla [5]. El uso de técnicas basadas en aprendizaje profundo y redes artificiales, han logrado abordar el reconocimiento de emociones en el discurso sin la necesidad de entender el contexto, atendiendo a la información emocional que las señales de audio transportan [14].

Usando la base de datos IEMOCAP (inglés), J.Wang [23] propone un modelo dual LSTM donde cada sonido se procesa con características MFCC y espectrogramas de MEL simultáneamente llegando a un 72.7 % de exactitud.

Siguiendo con un modelo híbrido, en [15] se estudia el resultado de un sistema que combina de CNN y LSTM para clasificar emociones en una secuencia de audio en alemán (EMO-DB), consiguiendo un 88.01 % de precisión.

Con el fin de eliminar el preprocesado de la señal, A.Qayyum presenta un modelo de redes convolucionales para una clasificación de emociones en el idioma inglés. Este utiliza la base de datos SAVEE, donde obtiene finalmente un 81.63 % de precisión [19].

En estudios más recientes, se aborda el problema del Reconocimiento de Emociones en el Habla con una red CNN computacionalmente eficiente que es alimentada con espectrogramas de Mel; es decir, se consigue una representación en 2D de la señal de audio aprovechando mejor las ventajas de una red de este tipo. El sistema es probado en con dos datasets distintos independientemente, IEMOCAP (en inglés, y eliminando 'frustración') y EMO-DB (alemán) consiguiendo un 77.01 % y 92.02 % de precisión respectivamente [2].
Siguiendo por el enfoque de CNN, en [16] exploran una arquitectura de este tipo compuesta por siete capas bidimensionales que es alimentada con espectrogramas MFCC, y lleva a cabo una clasificación de cinco emociones evaluando el resultado en RAVDESS donde consiguen un 81.0 % de precisión e IEMOCAP con un 84.00 %.
Finalmente, es obligatorio hablar del trabajo de G.Tamulevicius, que en una línea más cercana al objetivo de este proyecto, emplea una red neuronal convolucional bidimensional de tres capas entrenado en lituano, para reconocer seis lenguas (lituano, inglés, serbio, español, alemán y polaco). Aunque sus resultados no son realmente señalables, ya que no consiguen reconocer las emociones en la lengua extranjera, insisten en la importancia del uso de características en dos dimensiones, ya que proveen información temporal además de las características acústicas de las emociones [21].

III. OBJETIVOS Y METODOLOGÍA

El objetivo general de este trabajo es hacer un estudio comparativo del reconocimiento de emociones por voz, a través de lenguajes no aprendidos (lenguajes que no hayan formado parte del entrenamiento), una vez se haya conseguido un modelo capaz de clasificar en una lengua conocida con un porcentaje de acierto superior al 81 %. Esto implica:

- Hacer un estudio del estado del arte sobre diferentes métodos, técnicas, y conjunto de datos utilizados en el reconocimiento de emociones a través de la voz.
- Conseguir al menos tres datasets pertenecientes a tres idiomas diferentes donde uno de ellos será usado como referencia, y los otros deberán cumplir las siguientes condiciones: uno de los conjuntos de datos restantes deberá tener raíces fonéticas distintas al corpus de referencia, y el otro tener raíces fonéticas similares.
- Diseñar una solución en la que el conjunto de datos de referencia tenga un porcentaje de acierto superior al 81 % en la clasificación de emociones. Esta referencia ha sido marcada por los resultados reportados en la revisión del estado del arte.
- Aplicar el modelo diseñado en el paso anterior a los otros conjuntos de datos.
- Evaluar la tasa de acierto para cada uno de eso conjuntos y comparar los resultados obtenidos.

Para ello, la metodología elegida para este proyecto se divide en dos partes: una fase inicial y otra iterativa donde en cada iteración se diseñan unas modificaciones y capacidades funcionales que son añadidas en función de la etapa anterior.

1. Fase inicial:
   a) Revisión de la literatura sobre el reconocimiento de emociones en el habla, así como los métodos usados y los resultados obtenidos. Este paso permite una mayor comprensión del problema y su alcance.
   b) Análisis y recolección de posibles conjuntos de datos en diferentes idiomas que son aptos para los experimentos que se quieren realizar.
2. Elaboración:
   a) Identificación y redacción de una serie de pruebas iniciales con los diferentes métodos y técnicas descritas en la literatura, aplicados según el análisis de las bases de datos.
   b) Implementación en Python de las pruebas diseñadas con las técnicas y arquitecturas identificadas.
   c) Ajuste de los parámetros así como del balance de los datos con el fin de conseguir un mejor resultado.
   d) Evaluación: Se evalúan los resultados obtenidos de la implementación antes de decidir la iteración por finalizada.
3. Evaluación y comparación de los resultados.

IV. CONTRIBUCIÓN

El objetivo de este estudio es contrastar los resultados obtenidos tras aplicar el mismo sistema de reconocimiento de emociones en la voz entrenado con un lenguaje de referencia, con los otros dos lenguajes extranjeros. Mediante esta comparativa se pretende responder a la pregunta de si es posible reconocer emociones en un idioma que en principio se desconoce.

A. Conjunto de datos

A continuación se presentan los datos que se usan en este estudio. Con el fin de establecer unas dimensiones coherentes entre las bases de datos, se extraerán de los conjuntos originales seis emociones para clasificar en este trabajo: enfado, asco, tristeza, miedo, felicidad y neutral.

A.1. Idioma de referencia: Inglés

El idioma de referencia será el que se use en el entrenamiento.

- SAVEE es un conjunto de datos aplicado al reconocimiento de emociones que consiste en grabaciones de 480 frases en total en inglés británico ejecutadas por cuatro actores profesionales masculinos modulando siete emociones distintas (enfado, asco, tristeza, alegría, miedo, sorpresa y neutral) [11].
- TESS es un conjunto de datos compuesto por 2800 archivos de audio donde dos actrices de 26 y 64 años cuya lengua materna es el inglés americano, articulan 200 frases cada una y modulándolas en siete emociones (enfado, asco, tristeza, alegría, miedo, sorpresa, y neutral) [18].

A.2. Idiomas de test: Alemán y Francés

Estas bases de datos conforman el conjunto de test, los cuales son probados en los modelos resultantes de este trabajo.

- EMO-DB es una base de datos alemana que incluye una colección de 800 grabaciones interpretadas por diez actores (cinco hombres y cinco mujeres) matizando seis emociones (enfado, asco, tristeza, alegría, miedo y neutral) grabadas en una cámara anecoica [1].
- CaFE es una base de datos canadiense en idioma francés donde seis hombres y seis mujeres, pronuncian un total de seis frases interpretando siete emociones (enfado, asco, tristeza, alegría, miedo, sorpresa y neutral) [9].

Cuadro 1: Distribución de los conjunto de datos.

Emoción	TESS y SAVEE	EMO-DB	CaFE
enfado	460	92	92
asco	460	92	92
miedo	460	92	92
felicidad	460	92	92
tristeza	460	92	92
neutral	520	92	92

Distribución resultante de las clases pertenecientes a las bases de datos presentadas, después de haber sido balanceadas.

B. Extracción de características

Uso de características MFCC De la revisión del estado del arte, se concluye que MFCC es uno de los mejores algoritmos para capturar características de la señal de audio. Esto se debe a su similitud al sistema auditivo humano en cuanto al procesamiento de sonidos y frecuencias, por lo que su efectividad se ha visto reportada y discutida a lo largo de otros estudios [16] [23]. La librería usada para la manipulación de audio Librosa ofrece la posibilidad de extraer características MFCC de un archivo de audio. En cuanto a la configuración, se extraerán 13 características MFCC usando el rango de muestreo del propio archivo de audio [4].

Uso de espectrogramas Por otro lado, el uso de espectrogramas hace referencia a la conversión de la señal a imagen. El objetivo de esta técnica es aprovechar las fortalezas de las redes convolucionales en las imágenes aplicándolas a un problema de señal de audio. En concreto, para este trabajo se hará uso de los espectrogramas de las características MFCC de la señal, cuyo proceso constará de dos partes:

1. Generación de espectrogramas como imagen.
2. Lectura y procesado de las imágenes que alimentarán la red.

Para generar estos espectrogramas, se hará con la ayuda del paquete Librosa, especificando en los correspondientes parámetros la extracción 13 características MFCC; una vez generadas, se guardan en disco recortando el padding 0.05 pulgadas y en formato jpg. Finalmente, las imágenes generadas son leídas con la ayuda de OpenCV, donde se transforma su canal de color a RGB y son re-dimensionadas con un tamaño de 40 x 30 píxeles.

C. Pre-procesado de los datos

El primer paso en el preprocesado será su estandarización, que consiste en el cociente entre la media artimética de los valores de los datos de entrenamiento, y la desviación normal de los de test. En esta técnica los valores son centrados con respecto a su media con una desviación estándar, consiguiendo una mejora en la estabilidad numérica del modelo. Ya que a lo largo de las pruebas se usarán combinaciones de aumentos de datos e incluso mezclas entre distintos datasets, es recomendable aplicar este paso.
Posteriormente se categorizarán cambiando el formato de los datos para su uso en el modelo con keras. En este caso se utilizará la codificación *One Hot* que representa los enteros en secuencias de bits.
La división de los datos en entrenamiento y test se hará con el algoritmo *StratifiedShuffleSplit* de la librería de Python *sklearn*, que además de encargarse de barajar de manera aleatoria los datos previamente, asigna

cantidades equitativas a las clases cuando se divide en entrenamiento y test.
Como último punto, para las bases de datos extranjeras que serán usadas como test, se aplicará la técnica de aumento de datos basada en el cambio del tono o modulación. Esta consiste en cambiar el tono de un sonido sin variar su velocidad. Para implementar este método se usará la librería Librosa que ofrece un método específico.

D. Arquitectura

Modelo CNN-LSTM

- 3 capas convolucionales unidimensionales con 64 filtros de 3 x 3 y activación Relu, seguidas de una capa Max Pooling con tamaño 2 para la pool.
- Una capa Flatten con un Dropout del 25 %
- 2 capas LSTM unidimensionales con 50 y 20 unidades respectivamente y un Dropout del 50 %. Sólo se permitirá a la primera capa LSTM devolver el estado oculto de salida por cada entrada de tiempo. Ya que a este nivel se cambia a redes unidimensionales, se deberá redimensionar la entrada a 1 x 960.
- Capa de salida densa de 7 nodos y función Softmax.

Modelo CNN 2D

- 3 capas convolucionales bidimensionales con 32 filtros y un tamaño del kernel de 4 x 10. Como función de activación se usa Relu y pading establecido a 'same'.
- A las todas las capas convolucionales les sigue una capa Max Pooling con tamaño para la *pool* de 3, y posteriormente se aplica un Dropout del 20 %.
- Una capa Flatten, seguida de la capa densa de salida con 7 nodos y activación Softmax.

Modelo CNN 1D

- 2 capas convolucionales unidimensionales con activación Relu. El número de filtros es de 128 y el tamaño del kernel de 5. En las dos capas convolucionales se usa regularización de tipo L2 para aplicar una penalización a las capas del kernel con un valor de 0.01 y corregir así el overfitting.
- La primera capa convolucional está seguida de una capa Dropout del 0.5 y una capa Max Pooling con un tamaño 8.
- A la segunda capa convolucional se sigue otra capa de Dropout con un valor del 25 % y una capa Flatten.
- Por último, esta arquitectura cierra con una capa densa con 7 nodos (número de clases) con una función de activación Softmax.

E. Criterios de éxito

Las dos principales métricas que se usarán para decidir cómo de buena es la predicción del modelo serán:

- **Exactitud o *Accuracy*** Establece una comparación entre los resultados predichos y los obtenidos, determinando cómo de preciso es el algoritmo cuando se trata de identificar las clases.

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

- **F1 score** Siendo el *recall* la fracción de elementos relevantes que son recuperados (el cociente de las predicciones positivas y el número de clases positivas en el conjunto de test), la medida de F1 Score representa el balance entre la precisión y el *recall*.

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

Debido a la paradoja de la exactitud donde puede existir un sesgo por una distribución desigual de las clases, es aconsejable elegir una métrica de evaluación adicional además de la *accuracy*. Por ello se contará con F1 que es la media armónica entre el recall y la precisión.

F. Diseño de los experimentos

Originalmente para llegar a los resultados, se han efectuado más pruebas de las que aquí se exponen. No obstante, debido a la extensión de estas, se ha decidido incluir en este artículo únicamente las más relevantes.

F.1. Búsqueda del mejor modelo

Estos experimentos giran en torno a la búsqueda del modelo en el que más tarde se validarán los lenguajes extranjeros.

- **Experimento 3**: SAVEE y TESS con el modelo CNN 1D con características MFCC
- **Experimento 5**: Ensamblado con SAVEE y TESS con modelo CNN 2D usando espectrogramas.

- **Experimento 6**: Ensamblado con SAVEE y TESS con modelo CNN-LSTM usando espectrogramas.

F.2. Pruebas con lenguajes extranjeros

En estas pruebas se evaluarán los mejores modelos de la sección anterior con lenguajes que no han sido vistos en su entrenamiento.

- **Experimento 7**: Este experimento evalúa los tres modelos del bloque anterior con el idioma alemán (base de datos EMO-DB).

- **Experimento 8**: Este experimento evalúa los tres modelos del bloque anterior con el idioma francés (base de datos CaFE).

V. DESCRIPCIÓN DE LOS RESULTADOS

A. Evaluación mono-lingüística

A.1. Experimento 3

Se muestran los resultados del experimento donde se explora el comportamiento del modelo CNN 1D con la combinación del conjunto de datos SAVEE y TESS.

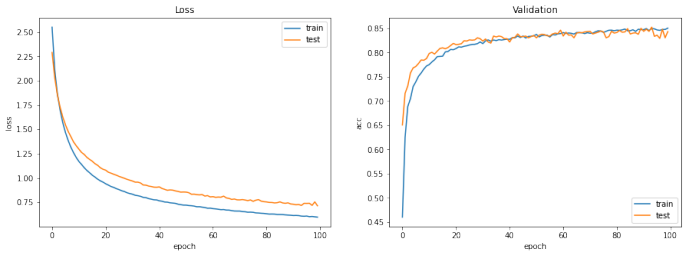


Figura 1: Rendimiento del modelo CNN 1D con los datos TESS y SAVEE.

La figura 1 corresponde al rendimiento del modelo que utilizó un optimizador RMSprop con una tasa de aprendizaje de 0.00005, valor de ϵ a 'None' y entropía cruzada categórica (*categorical crossentropy*) como función de coste. Para afinar el modelo, se añadieron los callbacks ReduceLROnPlateau con un factor de reducción de la tasa de aprendizaje de 0.9 minimizando la *val loss*, y EarlyStopping maximizando la *val accuracy*, ambos con una paciencia de 20 épocas. El modelo se entrenó durante 100 épocas con un batch de tamaño 32, resultando en los datos que se exponen en la tabla 2

Cuadro 2: Resultados del modelo CNN 1D.

Clase	TESS y SAVEE	
	acc	F1
enfado	0.47	0.64
asco	1.00	0.85
miedo	1.00	0.79
felicidad	0.97	0.89
tristeza	1.00	0.87
neutral	1.00	0.79
accuracy	**88.22 %**	

A.2. Experimentos 5 y 6

Estos experimentos exploraron el comportamiento de dos modelos basados en arquitecturas de redes convolucionales alimentados con espectrogramas a partir del conjunto de datos SAVEE y TESS.

Por un lado el **experimento 5** usó una arquitectura CNN 2D

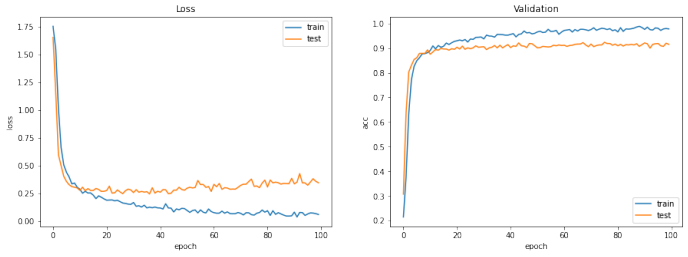


Figura 2: Rendimiento del modelo CNN 2D con los datos TESS y SAVEE.

Y el **experimento 6**, una arquitectura CNN-LSTM:

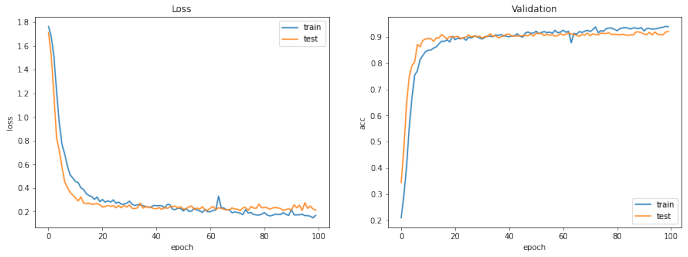


Figura 3: Rendimiento del modelo CNN-LSTM con los datos TESS y SAVEE.

En ambos experimentos, la estrategia de entrenamiento que se siguió fue un optimizador Adam con los parámetros por defecto que ofrece Keras y la entropía cruzada categórica (*categorical crossentropy*) como función de pérdida. Se entrenó durante 100 épocas con un batch de tamaño 30.

Cuadro 3: Resultados de los experimentos 5 y 6.

Clase	CNN 2D		CNN-LSTM	
	acc	F1	acc	F1
enfado	0.86	0.89	0.96	0.94
asco	0.90	0.91	0.67	0.80
miedo	0.94	0.93	0.99	0.93
felicidad	0.94	0.94	1.00	0.95
tristeza	0.96	0.92	0.99	0.91
neutral	0.88	0.89	0.96	0.91
accuracy	**90.50 %**		**92.06 %**	

Resultados de los modelos que fueron alimentados por espectrogramas.

B. Evaluación en la lengua extranjera

B.1. Experimento 7

Este experimento evaluó los tres modelos del bloque anterior en el idioma alemán, usando como conjunto de test la base de datos EMO-DB.

Cuadro 4: Resultados del experimento 7.

Clase	CNN 1D		CNN 2D		CNN-LSTM	
	acc	F1	acc	F1	acc	F1
enfado	0.16	0.26	0.17	0.18	0.18	0.21
asco	0.25	0.20	0.06	0.04	0.10	0.12
miedo	0.35	0.21	0.30	0.40	0.22	0.19
felicidad	0.56	0.36	0.21	0.25	0.24	0.26
tristeza	0.36	0.08	0.26	0.10	0.38	0.22
neutral	0.26	0.17	0.21	0.17	0.16	0.16
accuracy	**32 %**		**20 %**		**22 %**	

Compara el resultado de los tres modelos entrenados en inglés evaluados en el idioma alemán.

B.2. Experimento 8

Se exponen los resultados del experimento 8 que evaluó los modelos del bloque anterior en el idioma francés, usando como conjunto de test la base de datos CaFE.

Cuadro 5: Resultados del experimento 8.

Clase	CNN 1D		CNN 2D		CNN-LSTM	
	acc	F1	acc	F1	acc	F1
enfado	0.14	0.22	0.20	0.30	0.20	0.26
asco	0.26	0.15	0.17	0.13	0.21	0.22
miedo	0.00	0.00	0.13	0.12	0.11	0.10
felicidad	0.23	0.18	0.17	0.15	0.18	0.25
tristeza	0.35	0.11	0.14	0.05	0.21	0.11
neutral	0.10	0.04	0.18	0.10	0.35	0.11
accuracy	**18 %**		**18 %**		**21 %**	

Compara el resultado de los tres modelos entrenados en inglés evaluados en el idioma francés.

VI. DISCUSIÓN O ANÁLISIS DE RESULTADOS

A. Análisis mono-lingüístico

Este conjunto de experimentos tenía como objetivo evaluar diferentes configuraciones (datos, arquitecturas y técnicas) con el fin de conseguir un modelo óptimo en el idioma de referencia. Como ya se ha mencionado, únicamente se han expuesto aquellas pruebas realmente relevantes. Concretamente en el experimento 3 se comprobó que la combinación de los conjuntos SAVEE y TESS fue realmente eficiente debido a que proporcionaba una gran diversidad de datos y por lo tanto más capacidad de aprendizaje a la red. Esta arquitectura fue notablemente simplificada, lo que junto a la penalización de las capas de salida que se aplicaron a las dos primeras capas convolucionales (regularización L2), ayudaron a reducir el *overfitting*.

Cabe resaltar que el desempeño del modelo con redes convolucionales unidimensionales compite con los resultados reportados por [19], [16], y [2], los cuales con arquitecturas más complejas consiguen menor exactitud.

Los experimentos 5 y 6 exploraron el uso de espectrogramas MFCC como estrategia. Este enfoque logró los mejores resultados: 92.06 % en CNN-LSTM en primer lugar y CNN con un 91.50 % en una clasificación de seis emociones en inglés. Aquí se pudo comprobar la efectividad del uso de espectrogramas en un clasificador mono-lingüístico, que también se refleja en la estabilidad de los valores de accuracy y F1 ya que, presentan menor variación entre ellos.

Llama la atención el hecho de que en todas las pruebas que se hicieron en un modelo basado únicamente en redes convolucionales entrenado y evaluado en inglés, el Enfado fue la emoción más complicada de distinguir. Esto podría sugerir un patrón para este lenguaje, lo que encaja con los resultados encontrados en la revisión del estado del arte [19] [16][2]. La comparación de estas observaciones con las de C.Huang [10] indica que el factor más distinguible respecto al Enfado es cómo sus características varían en el tiempo, lo que tiene sentido con el hecho de que este patrón no se continúe en los resultados del experimento 6 con el modelo CNN-LSTM, ya que este sistema permite aprender la relevancia temporal de cada secuencia del habla.

B. Análisis en la lengua extranjera

Finalmente, tal y como se predijo para los experimentos 7 y 8, se obtuvo un porcentaje de accuracy un poco mayor en la evaluación con alemán (lengua con raíces fonéticas más próximas al idioma que se usó en el entrenamiento), que en la evaluación con francés (lengua con raíces fonéticas más lejanas al idioma que se usó en el entrenamiento).

No obstante, ni en el experimento 7 ni en el experimento 8 se pudieron hacer las mismas observaciones que en las pruebas con un enfoque mono-lingüístico, ya que al contrario de lo que ocurría con los modelos evaluados en inglés:

- No fue posible encontrar un patrón sobre la emoción más difícil (o más fácil) de reconocer como ocurría con el Enfado.
- En el experimento 7 (evaluación en alemán), el modelo basado en redes convolucionales unidimensionales fue el que mayor porcentaje de exactitud obtuvo con un 32 % de exactitud.

En la tabla 6 se comparan los resultados de este trabajo con el de [21] tras haber evaluado un modelo entrenado en una lengua con otras distintas usando espectrogramas.

Cuadro 6: Comparación de los modelos evaluados en lenguas extranjeras.

Idioma	**Tamulevicius 2020**		**CNN-LSTM**	
	Accuracy	F1	Accuracy	F1
Serbio	0.37	0.18	-	-
Polaco	0.21	0.17 9	-	-
Alemán	0.49	0.27	**0.19**	**0.19**
Español	0.3	0.19	-	-
Francés	-	-	**0.20**	**0.18**

Compara el trabajo de Tamulevicius donde se han usado espectrogramas con el modelo propuesto (CNN-LSTM).

Atendiendo a los datos que se muestran en la tabla 6, no parece que se pueda establecer una relación de proximidad fonética entre distintos idiomas para la clasificación de emociones en la lengua extranjera siguiendo un enfoque basado en redes convolucionales. El trabajo de Tamulevicius tampoco presenta un porcentaje de acierto proporcional entre lenguas fonéticamente similares, ya que el polaco (lengua eslava) debería ser la que puntúa más alto teniendo en cuenta que es evaluado en un modelo entrenado con lituano (lengua báltica) [13].
Haciendo una revisión de lo aprendido en este trabajo, se hacen algunas observaciones sobre por qué, usando espectrogramas que han resultado ser tan exitosos para la clasificación en una sola lengua, no se comportan de la misma manera en otros idiomas:

- Los objetos visuales y los sonidos no se agrupan en una imagen de la misma manera. Se asume que un píxel en una imagen pertenece a un determinado objeto por ser de un color específico, mientras que las características del sonido (tono o intensidad) no se separan en capas distinguibles [24].
- Los ejes X e Y de una imagen no tienen el mismo significado que en un espectrograma, ya que en una imagen, la información representada **no cambia su significado**. Sin embargo esto no ocurre de la misma manera en un espectrograma donde estas dos dimensiones representan unidades distintas: frecuencia y tiempo [22].
- Desde el punto de vista humano, la forma en la que se procesan esas señales no es comparable. A la hora de localizar un objeto de manera visual en el entorno, se escanea lo que hay alrededor varias veces, ya que los objetos que lo conforman son estáticos. Por otro lado, un sonido toma la forma física de la presión manifestada en la onda, y desde el punto de vista de quien la percibe, esa determinada onda con un determinado estado sólo existe en un momento específico. Dicho de otro modo, la información que contiene una onda de audio está dispuesta de manera secuencial [12].

VII. CONCLUSIONES

Para finalizar este artículo, se reúnen a continuación las conclusiones en base a los objetivos establecidos en el capítulo 3.
Respecto al primer objetivo específico, el cual consistía en una revisión en profundidad de la literatura actual, se abordó en el capítulo 2. Se llegó a la conclusión de que no existía una línea definida sobre qué métodos, técnicas y datos eran los más adecuados, para abordar este problema. Debido a esto, se decidió apostar por las técnicas que más apoyo tenían por parte de la literatura.
Una vez se reunieron los conjuntos de datos que atendían a las condiciones establecidas en el capítulo 3, se pasó a diseñar una solución cuyo porcentaje de exactitud fuese superior a un 81 %. Los tres mejores modelos desarrollados en este trabajo consiguieron una puntuación por encima de la marca que se propuso, donde la diversidad de datos jugó un papel determinante. Esta estrategia no sólo redujo el overfitting, sino que hizo posible un modelo más rentable ya que, en comparación a las arquitecturas de otros trabajos se consiguió mejor porcentaje de acierto con diseños más simples. Quedó clara la superioridad del uso de espectrogramas para un clasificador de emociones mono-lenguaje, especialmente con una arquitectura híbrida CNN-LSTM ya que tiene en cuenta la información temporal inherente a la señal de audio.
Por lo que respecta a los experimentos hechos evaluando una lengua extranjera y atendiendo al último punto de los objetivos, se ha llegado a las mismas conclusiones que Tamulevicius [21], donde tras hacer diversas pruebas con diferentes idiomas, únicamente puede reconocer aquellos con los que entrena su modelo. No obstante, extrayendo lo aprendido en este trabajo, se

desaconseja tratar la señal acústica con imágenes estáticas (espectrogramas) por no adaptarse correctamente a cómo estas funcionan y perder información importante.
La estrategia que se planteó donde las lenguas extranjeras se dividían según su similitud fonética parece no ser muy relevante coincidiendo con [17] donde afirma que para reconocer el estado emocional en una lengua no aprendida, se necesita mayor exposición a ésta.
El alcance de este proyecto se ha debido simplificar, por lo que se plantean líneas de trabajo futuras:
Revisando las evidencias publicadas, se asumió que las características *cepstrales* eran la mejor opción para abordar el problema que se planteaba, sin embargo, como se han mencionado en el análisis, se considera que este no es el mejor enfoque. Por un lado se cree que sería recomendable no tratar las emociones como categorías discretas, ya que varían enormemente de un lenguaje a otro. Esto requiere un análisis más en profundidad sobre la fonética en el idioma en cuanto a la expresión emocional para conseguir una mayor independencia cultural. Por otro lado, sería interesante explorar la combinación mecanismos de atención junto al punto anterior, ya que podrían computar segmentos relevantes de la señal de audio y conseguir así mejor generalización.

Referencias

[1] "A database of German emotional speech". En: *9th European Conference on Speech Communication and Technology* (2005).

[2] T. Anvarjon, Mustaqeem y S. Kwon. "Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features". En: *Sensors (Switzerland)* 20.18 (2020).

[3] S.E. Arnold. *Bradley Metrock and the Alexa Conference: Alexa As a Game Changer for Search and Publishing*. Feb. de 2017.

[4] M. Bao y A. Huang. "Human vocal sentiment analysis". En: *arXiv* (2019).

[5] A. Davletcharova y col. "Detection and Analysis of Emotion from Speech Signals". En: *Procedia Computer Science* 58 (2015).

[6] F. Dellaert, T. Polzin y A. Waibel. "Recognizing emotion in speech". En: *International Conference on Spoken Language* 3 (1996).

[7] L. Effron. *iPhone 4S's Siri Is Lost in Translation With Heavy Accents - ABC News*. Oct. de 2011.

[8] P. Ekman. *An Argument for Basic Emotion*. 1992.

[9] P. Gournay, O. Lahaie y R. Lefebvre. "A canadian French emotional speech dataset". En: *9th ACM Multimedia Systems Conference, MMSys* (jun. de 2018).

[10] C. Huang y col. "Practical speech emotion recognition based on online learning: From acted data to elicited data". En: *Mathematical Problems in Engineering* (2013).

[11] Philip Jackson y Sana ul haq. *Surrey Audio-Visual Expressed Emotion (SAVEE) database*. Abr. de 2011.

[12] Jonathan Hui. *Speech Recognition - Phonetics*. 2019.

[13] F. Kortlandt. "FROM PROTO-INDO-EUROPEAN TO SLAVIC". En: (2002).

[14] S. Langari, H. Marvi y M. Zahedi. "Efficient speech emotion recognition using modified feature extraction". En: *Informatics in Medicine Unlocked* 20 (2020).

[15] W. Lim, D. Jang y T. Lee. "Speech emotion recognition using convolutional and Recurrent Neural Networks". En: *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA* (2016).

[16] Mustaqeem y S. Kwon. "A CNN-assisted enhanced audio signal processing for speech emotion recognition". En: *Sensors (Switzerland)* 20.1 (2020).

[17] Marc D. Pell y V. Skorup. "Implicit processing of emotional prosody in a foreign versus native language". En: *Speech Communication* 50.6 (2008).

[18] M. Pichora-Fuller y K. Dupuis. *Toronto emotional speech set (TESS)*. 2020.

[19] A. Qayyum, A. Arefeen y C. Shahnaz. "Convolutional Neural Network (CNN) Based Speech-Emotion Recognition". En: *IEEE International Conference on Signal Processing, Information, Communication and Systems* (2019).

[20] E. Shriberg y col. "Crowdsourcing Emotional Speech". En: (2018).

[21] G. Tamulevicius y col. "A study of cross-linguistic speech emotion recognition based on 2d feature spaces". En: *Electronics (Switzerland)* 9.10 (2020).

[22] P. Verma y J. O. Smith. "Neural Style Transfer for Audio Spectrograms". En: (ene. de 2018).

[23] J. Wang y col. "Speech emotion recognition with dual-sequence LSTM architecture". En: *IEEE International Conference on Acoustics, Speech and Signal Processing* (2020).

[24] L. Wyse. "Audio Spectrogram Representations for Processing with Convolutional Neural Networks". En: (jun. de 2017).