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, mostró la superioridad 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 se han animado 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, en [19] se 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.

En estudios más recientes, [2] 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.
Siguiendo por el enfoque de CNN, en [16] exploran una arquitectura basada en CNN 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 [21], 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.

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 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.

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, 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 de la revisión de 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.
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 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 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 las 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 por su similitud a cómo el sistema auditivo humano procesa el sonido y las 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, 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

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, que 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 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 conviene el balance entre la precisión y el *recall*.

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

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

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

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 más 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.

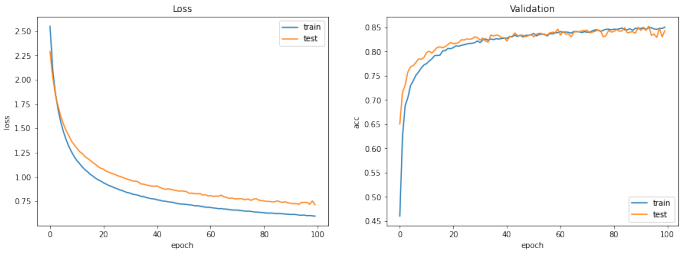

Figura 1: Rendimiento del modelo CNN 1D con los datos TESS y SAVEE.

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. Añadiendo además para afinar el modelo 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

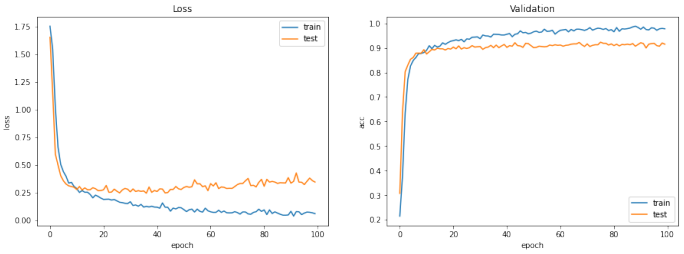

Figura 2: Rendimiento del modelo CNN 2D con los datos TESS y SAVEE.

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

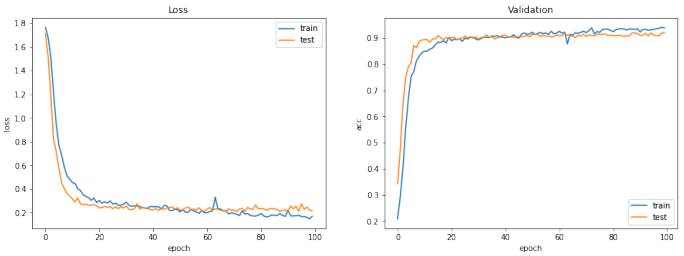

Figura 3: Rendimiento del modelo CNN-LSTM con los datos TESS y SAVEE.

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 que 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 sugiere 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 [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 mejor modelo resultante (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, si no 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 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:

Por la revisión del estado del arte, 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).