UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Máster en Big Data y Data Science: ciencia e ingeniería de datos

TRABAJO FIN DE MÁSTER

Reconocimiento de siete emociones en la voz haciendo uso de las características audibles del sonido.

Amanda Rodríguez González
Tutores: Joaquín González-Rodríguez y Doroteo Torre Toledano

Octubre 2021

Reconocimiento de siete emociones en la voz haciendo uso de las características audibles del sonido.

AUTOR: Amanda Rodríguez González
TUTORES: Joaquín González-Rodríguez y Doroteo Torre Toledano

Escuela Politécnica Superior
Universidad Autónoma de Madrid
Octubre de 2021

Resumen

En la presente memoria se realiza un acercamiento al reconocimiento de emociones en la voz, *Speech Emotion Recognition* (SER). Haciendo uso de las características audibles de la voz y redes neuronales profundas. Se emplea un corpus formado por cuatro bases de datos en inglés creadas especialmente para el reconocimiento de emociones en la voz. Las emociones catalogadas en este corpus son neutralidad, felicidad, tristeza, enfado, miedo, disgusto y sorpresa. Para la definición de las mismas se efectúa un estudio de las características audibles del sonido con el fin de conocer cuales influyen más en la búsqueda de un patrón para su clasificación. Tras analizar varias de estas características, se seleccionan las que obtuvieron mejores resultados, componiendo 52 variables. Así un audio queda representado por la media y la desviación típica de 13 coeficientes Cepstrales, 6 tonos y 7 centroides espectrales. Se diseñaron tres modelos de aprendizaje automático con capas CNN, LSTM y BLSTM. El modelo con mejor rendimiento resultó ser el formado por tres capas CNN y una LSTM alcanzando cerca de un 70% de precisión en la predicción de emociones en la voz.

Agradecimientos

Al grupo de investigación AUDIAS. En particular a mi tutor, Joaquín González, por guiar la investigación y transmitirme el afecto hacia este campo, esperemos que te recuperes pronto. A Doroteo Torre, por su dedicación e involucración. Gracias por vuestro tiempo.

A la Universidad Autónoma de Madrid y a los buenos docentes que he tenido la oportunidad de disfrutar del master de Big Data y Data Science. Así mismo, a mis compañeros más cercanos de clase, a Julia.

A mi madre, Pilar. A mi padre, Antonio. Siempre haciendo posible que crezca intelectual y personalmente poniendo toda la confianza en mí y apoyándome incondicionalmente. A mi familia de Valladolid. A mi abuela, Manuela, gracias por enseñarme el gusto por el trabajo y su recompensa. A los que ya no están y aun así siguen conmigo, a ti, Henar.

A mis personas de Madrid, a los que están en las buenas y en las malas. A mis hermanitos de Cuatro Caminos. A mis muchachas, María del Carmen y María José. A mis muchachos, Alberto, Fran y Pablo. Gracias por crear un hogar.

A mis amigos desperdigados por el mundo. A Daniel, a Charles. Gracias por enseñarme a querer pese a la distancia o el tiempo.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

El análisis y la detección de las emociones se lleva ejercitando desde hace años en la industria de la informática con el nombre de computación afectiva. Las técnicas más utilizadas han sido siempre el estudio de la expresión facial y el texto. Sin embargo, estas no sirven si no se posee un contenido visual o si se desea conocer el "cómo se dijo", en vez del "qué dijo"; para resolver este problema, se hace necesario analizar la voz humana y sus características audibles [1].

Speech Emotion Recognition (SER, reconocimiento de emociones en la voz) está empezando a ganar valor y a implementarse en numerosas aplicaciones en las que es importante percibir el estado emocional de la voz. No solo en call centers, sino también en las que haya interacción del humano con una máquina como, por ejemplo, los asistentes con inteligencia artificial de los principales sistemas: Siri, Google Assistant o Alexa. En los cuales, es de gran relevancia que la interacción de estos con la sociedad sea lo más humana y natural posible [2,3]. Según el escritor y experto en marketing digital, Gennady Vaynerchuk, actualmente 1 de cada 4 búsquedas se hacen por voz e irá creciendo en los próximos años. Además, diferentes estudios aseguran que, en la percepción de las emociones, exclusivamente un 7% depende de las palabras, siendo el 38% el tono de voz y el lenguaje corporal [4].

No obstante, la detección de emociones en una señal de audio aún está en vías de desarrollo y se enfrenta a grandes desafíos: (1) categorizar las emociones, (2) extracción de características útiles del sonido para definirlas de una manera precisa y (3) búsqueda de un modelo con arquitectura de red neuronal que generalice correctamente [5].

1.2 Objetivos

El objetivo de este Trabajo de Fin de Máster consiste en el reconocimiento de siete emociones de la voz mediante aprendizaje automático. Empleando metodologías similares a las usadas en el reconocimiento de voz y *Music Emotion Recognition* (MER, reconocimiento de emociones de la música) [6,7].

Las emociones objeto de estudio serán: (1) neutralidad, (2) felicidad, (3) tristeza, (4) enfado, (5) miedo, (6) disgusto y (7) sorpresa. Se intentará encontrar en los audios de cada

emoción las variables de tipo sonido que las consiga clasificar; haciendo uso del paquete de Python Librosa, empleado para el análisis de audio y música [8].

Se comprobará en qué medida son eficientes las redes neuronales profundas buscando patrones en determinadas características audibles de la voz, disponiendo de una base de datos "salvaje" [5]. En definitiva, sin hacer uso de palabras, expresión facial o cualquier otra variable como el sexo o la edad. Se compararán diferentes arquitecturas de redes neurales de clasificación como *Multilayer Perceptrons* (MPL), *Convolutional Neural Network* (CNN), *Long Short-Term Memory Network* (LSTM) y modelos mixtos, con el objetivo final de construir un modelo que mejore la generalización.

1.3 Organización de la memoria

La memoria consta de los siguientes capítulos: (1) descripción del corpus así tanto la procedencia como el etiquetado del mismo; (2) acercamiento a las características del sonido que podrían definir las emociones y la selección de las que obtengan un mejor resultado para definir el modelo, según varias teorías, estudios previos y experimentos propios; (3) arquitectura de la red neuronal y el proceso que se ha seguido para su diseño; (4) resultados y análisis; y (5) conclusiones y líneas futuras.

Los pasos seguidos y forma de trabajo del presente estudio quedan esquematizados según muestra la Ilustración 1.

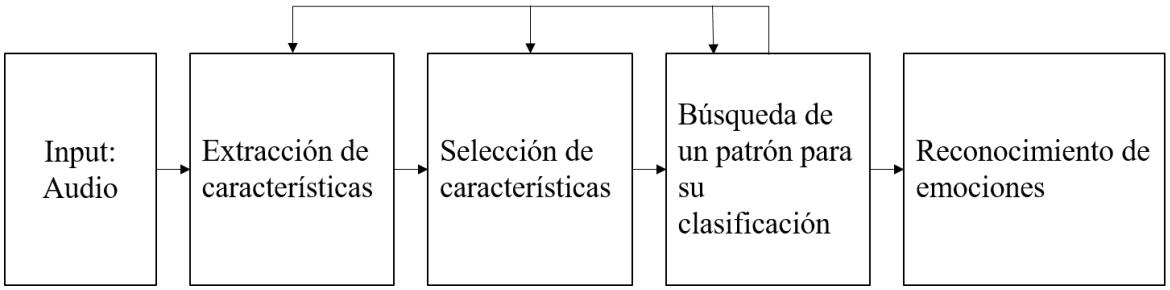


Ilustración 1. Esquema de trabajo

2 Descripción del corpus

El corpus seleccionado es una mezcla heterogénea de partes de cuatro bases de datos famosas utilizadas para el reconocimiento de las emociones en la voz. La Tabla 1 reúne la información relevante de cada base de datos, se puede encontrar más información en su respectiva referencia.

Tabla 1. Composición del corpus

Base de datos	Nº de registros	Composición (%)	Nº de hablantes	Idioma	Contenido	Duración media (s)
CREMA-D [9]	7.442	58,15	91	Inglés	10 frases	2.5
TESS [10]	2.800	21,88	2	Inglés	200 palabras	2
RAVDESS [11]	2.076	16,22	24	Inglés	12 frases	3
SAVEE [12]	480	3,75	4	Inglés	15 frases	3

Con esta combinación de bases de datos se alcanzan del orden de 13.000 registros, un número lo suficientemente grande para poder realizar aprendizaje automático mediante redes neuronales profundas. Aparte del número de registros de cada una, también se han recogido características como el número de hablantes, el idioma, el contenido y la duración. Se intentaron añadir bases de datos con un idioma diferente al inglés, como la EMO-DB [13], en alemán, obteniendo peores resultados en la eficiencia del aprendizaje, marcando una posible relación entre el idioma y las características del sonido como próxima investigación. Si bien estas variables no se van a tener en cuenta en el presente estudio, sí es necesario conocerlas para comprender el cierto grado de diversidad del corpus y poder examinar exclusivamente la influencia de las características del sonido.

La Tabla 2 muestra las etiquetas del corpus con el número de registros de cada una y su composición. A simple vista, y sin entrar en detalle, se puede observar que se trata de una base de datos desbalanceada, ya que se dispone de menor cantidad de registros para la emoción “Sorpresa”. Es importante tenerlo en cuenta para prevenir fallos en el diseño del modelo de aprendizaje. Esto se analizará en detalle en las siguientes secciones.

Tabla 2. Etiquetado del corpus

Emoción	Nº de registros	Composición (%)
Neutralidad	1.794	14,02
Felicidad	2.166	16,92
Tristeza	2.167	16,93
Enfado	2.167	16,93
Miedo	2.046	16
Disgusto	1.863	14,56
Sorpresa	591	4,64

3 Extracción de características

Este capítulo se centra en la definición de las características del sonido que, según diferentes estudios, podrían afectar más a la categorización de las emociones. Tras exponer la teoría básica, se procede a la selección de características del corpus seleccionado para el modelo a crear en concreto.

3.1 Categorización de las emociones

La definición y clasificación de las emociones ha sido siempre un tema de elevada polémica en el ámbito de la psicología. Diferentes autores han clasificado las emociones, principalmente, en dos aproximaciones: (1) categorías discretas de cada emoción y (2) clasificación multidimensional [2,14].

El pionero de esta primera agrupación es el psicólogo Paul Ekman. A través de sus estudios de las emociones y la expresión facial de las mismas, consigue definir 6 emociones básicas en los humanos: felicidad, tristeza, enfado, miedo, disgusto y sorpresa, de las que derivan todas las demás [15].

La aproximación multidimensional define un espacio continuo de emociones. Esto ha llevado a la creación de multitud de diferentes modelos tanto bidimensionales como tridimensionales. El patrón más utilizado es el desarrollado por Russell [16], clasificando 150 emociones en un espectro continúo de 2 D en el que los ejes son la excitación y la tensión. Otros autores también corroboran y amplían esta teoría, por ejemplo, Thayer [17], Whissell o Scholberg con su modelo tridimensional [14].

La Ilustración 2 aplica todas estas teorías al corpus definido en el capítulo anterior, con el que se va a trabajar. Es simplemente una mera clasificación para hacerse a la idea de las características del sonido que pueden ser apropiadas para encontrar un patrón que las delimite. Adicionalmente, se recurrirá a esta figura con el fin de argumentar los aciertos y confusiones en la clasificación del modelo y, así, poder reevaluarlo.

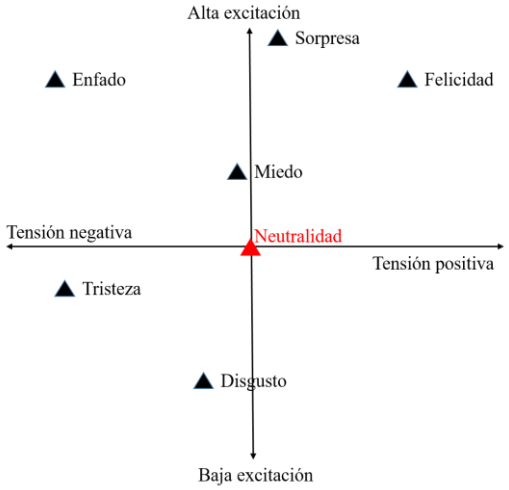


Ilustración 2. Clasificación de las emociones presentes el corpus en función de la tensión y la excitación (elaboración propia en base a los modelos de Russell y Thayer [16,17])

3.2 Características del sonido

Las características audibles de la voz humana son todas aquellas que definen al sonido en el espectro audible. La literatura las ha clasificado en tres grandes bloques: (1) continuas en el tiempo, como el tono o la energía; (2) cualitativas, entre ellas la calidad de la voz; (3) espectrales, por ejemplo, *Mel Frequency Cepstral Coefficients* (MFCCs Coeficientes Cepstrales en las Frecuencias de Mel) [2].

En las siguientes líneas se definen las características que más éxito de acierto obtuvieron en anteriores investigaciones de reconocimiento de emociones en la voz o en la música [2,4,6,7,18].

3.2.1 MFCCs y espectrograma de Mel

Los Coeficientes Cepstrales de las Frecuencias de Mel representan el habla desde el punto de vista de la percepción auditiva humana. Se usan en el llamado espectrograma de Mel, el cual intenta reproducir el funcionamiento del oído humano con mayor resolución en frecuencias bajas. Se desarrollan debido a la necesidad del reconocimiento automático de audio y se usan en sistemas de transcripción de la voz humana e identificación del hablante. Por lo que no significa que sean relevantes en el reconocimiento de emociones. De hecho, hay estudios que muestran la poca efectividad de los MFCCs por sí solos [4].

El vector final consta de 52 coeficientes: del 1 al 13 son los valores reales de los Coeficientes Cepstrales; del 14 al 26 representan la velocidad de cambio, “deltas”; del 27 al

39 la aceleración, los “deltas-deltas”; del 40 al 52 son los terceros coeficientes diferenciales [18].

3.2.2 Cromagrama

La característica cromática, *Chroma*, representa los perfiles de tono, se trata de una de las herramientas más poderosas para analizar la música. Se categorizan 12 tonos en la escala musical occidental. Usado en previos trabajos como una de las características para el reconocimiento de emociones en la música [7].

El vector final está compuesto por 12 componentes, uno por cada tono y semitono.

3.2.3 Centroide tonal

El centroide tonal, *Tonnetz*, clasifica un archivo de audio en 6 tonos, a diferencia del *Chroma* que lo hace en 12. Se usa en MER (*Music Emotion Recognition*), junto al *Chroma* [7,19]. Aunque esto no asegura el correcto funcionamiento para SER (*Speech Emotion Recognition*).

El vector final está compuesto por 6 componentes, uno por tono.

3.2.4 Contraste espectral

Calcular el nivel de energía de un audio es una tarea difícil debido a que la energía de la frecuencia varía con el tiempo. Esto es lo que consigue el contraste espectral, *Contrast*, midiendo la energía de frecuencia en cada marca de tiempo [19].

El vector final está compuesto por 7 componentes que representan la energía.

3.2.5 Tempo

El *Tempo* hace referencia a la velocidad de una pieza musical. Medido como la frecuencia del ritmo musical o los latidos por minuto, bpm.

El tempo puede ir variando a lo largo del audio en conjunto. Para poder medirlo se representa una matriz de características que muestran la prevalencia del tempo en cada momento, llamado tempograma.

3.3 Selección de características

La correcta selección de características juega un papel fundamental en la eficiencia de la clasificación. Se trata de una parte importante del diseño del sistema, por lo que será

preciso redefinirlas siempre que sea necesario y el objetivo sea tanto optimizar como mejorar el modelo.

En este apartado se realiza una selección de las características que mejor pueden clasificar las emociones del corpus (Tabla 2) entre las expuestas anteriormente. Para realizar esta elección se han llevado a cabo una serie de experimentos que miden el porcentaje de clasificación variando las características empleadas. Se define un modelo de clasificación básico, el perceptrón multicapa (MLP Classifier). Optimizado el modelo con los parámetros de la Tabla 3, se procede a entrenar el modelo con diferente combinación de variables. Estas variables no se estudiarán en ningún caso por si solas debido a que se obtiene una notable mejora en la tasa de acierto cuando se usan combinadas [4,5,7]. La extracción de características se realiza con la librería Librosa [8] y se anota tanto la media como la desviación estándar del vector representante de cada característica, debido a encontrar mejores resultados con la unión de ambas. Se separa el conjunto de datos de *train* del de *test* con la proporción 75-25% y se entrena el modelo con diferentes características. Los resultados quedan recogidos en la Tabla 4.

Tabla 3. Parámetros optimizados del modelo MLP

alpha	batch_size	epsilon	hidden_layer_sizes	learning_rate	activation	solver
0,1	256	1e-08	(50, 100, 50)	constant	relu	adam

Tabla 4. Porcentaje de acierto medio de la clasificación para el conjunto de train y test con diferente combinación de las características

Características	Nº variables usadas (media y desviación)	Acierto en train (%)	Acierto en test (%)
MFCC, Chroma, Tonnetz, Contrast, Tempo	898	82	71
MFCC, Tonnetz, Contrast, Tempo	874	83	70,5
MFCC, Tonnetz, Contrast, Tempo	820	85	70
MFCC, Tonnetz, Contrast	**52**	**74**	**62**
MFCC, Chroma, Tonnetz, Contrast	76	73	61
Tonnetz, Contrast	26	69	59
Tonnetz, Contrast, Tempo	794	74	59
MFCC, Mel, Chroma	50	70	59

MFCC, Mel, Chroma, Contrast, Tempo	306	70	58
Tempo	768	67	57

Tabla 5. Número máximo de variables que generan las características del sonido

Características	**MFCC**	**Mel**	**Chroma**	**Tonnetz**	**Contrast**	**Tempo**
Nº máximo de variables	53	128	12	6	7	768

Se comenzó usando las variables más utilizadas en la literatura para el reconocimiento de emociones en la música y transcripción de audio: MFCC, Mel y *Chroma*, obteniendo una eficiencia de predicción relativamente baja, del 59% de acierto y un modelo de menos del 75% de precisión (Tabla 4). En otras investigaciones se obtienen resultados similares para otros corpus y estas características [4,6].

Para que mejorara la actuación del modelo, se buscó otras características que pudieran estar más relacionadas con la excitación y la tensión (Ilustración 2). Se probó con Tonnetz y Contrast, obteniendo un resultado similar al anterior, pero con menos variables. Por esta razón se sustituyen las características Mel y *Chroma* por *Tonnetz* y *Contrast*. Se probó a eliminar la característica MFCC obteniendo resultados ligeramente peores, pues, aunque por sí sola no produzca una buena clasificación, junto con las demás sí [4]. Se comprobó que el rendimiento mejoraba si en vez de emplear los 53 coeficientes Cepstrales se usaban solo los 13 primeros. La Tabla 5 recoge el número máximo de dimensiones que puede contener el vector de salida de las características. Siendo uno de los mejores resultados el modelo formado por MFCC, *Tonnetz* y *Contrast* con un 62% de efectividad en test y exclusivamente 52 variables.

En la búsqueda de nuevas variables para mejorar el resultado se escogió el *Tempo*. Individualmente produce mejores resultados que las demás variables, puede representar la excitación y contiene una información considerable, 768 variables (Tabla 5).

La mejor práctica resultó considerar la media y la varianza de los 13 primeros coeficientes de MFCC, el *Tonnetz*, *Contrast* y *Tempo*. Aunque no sea el porcentaje más alto disponible en la Tabla 4, es el que tiene menor número de variables, 820, y el modelo funciona

moderadamente mejor que los otros, 85%, con un porcentaje de predicción mayor, del 70%.

Con el modelo y las características optimizadas, se hizo un cambio en el etiquetado del corpus con el objetivo de balancear la base de datos. Este cambio se basa en la categorización de las emociones presentada en la sección 3.1 (Ilustración 2) y queda recogido en la Tabla 6. El resultado fue de un modelo funcionando para *train* al 86% y en *test* al 75%. Esto no se trata de un buen resultado, pues los aciertos no mejoran en exceso y se está reduciendo de 7 a 4 clases. Se descarta esta clasificación y se mantiene la clase desbalanceada debido a que no se observa que produzca problemas graves al modelo.

Tabla 6. Reducción de la catalogación del corpus

Nueva clasificación	**Emociones**
Clase 1	Sorprendido, felicidad
Clase 2	Neutralidad
Clase 3	Miedo, enfado
Clase 4	Tristeza, disgusto

4 Diseño de la red neuronal

Las redes neuronales profundas son conocidas por su uso en multitud de aplicaciones de inteligencia artificial. Las redes neuronales profundas se caracterizan por disponer de varias capas entre las de entrada y salida. Estas técnicas han permitido reducir la tasa de error en campos como el reconocimiento de voz y de imágenes, incluso menor que el margen de error humano en este último caso [20]. A priori, se considera como una buena solución para la categorización de emociones en la voz [4,6].

En este apartado se explica, a modo de introducción, los principales clasificadores básicos usados en otros trabajos para el reconocimiento de emociones tanto en la música como en la voz. Finalmente, se compara la eficiencia de diversos modelos mixtos, con diferente arquitectura, en el corpus seleccionado. Para la ejecución se ha empleado la librería Scikit-learn [21] y Keras con TensorFlow.

4.1 Clasificadores

Con el objetivo de diseñar la red neural óptima para el presente problema, se realizan tres experimentos previos para medir la eficiencia de las capas: (1) MLP, (2) CNN y (3) LSTM. También se efectuaron pruebas con la capa *Gated Recurrent Units* (GRU). Sin embargo, no se añadieron a la memoria por su poca eficiencia por sí solas. Se seleccionan MFCC, *Tonnetz* y *Contrast* como características por ser las que generaron el resultado más eficiente. Así pues, se estará trabajando con 52 variables (Tabla 4).

4.1.1 Perceptrón Multi Capa (MLP Multi-Layer Perceptron)

El perceptrón multicapa es una clase de red neuronal artificial particularmente sencilla. Como se dijo en el capítulo 3.3, es la estructura de red neuronal profunda más básica. La componen una serie de capas totalmente conectadas. Cada capa es un conjunto de funciones no lineales de suma ponderada de todas las salidas de la anterior.

Los parámetros óptimos del modelo están recogidos en la Tabla 3, obtenidos tras aplicar el método GridSearchCV, disponible en la librería Scikit-learn. El procedimiento de trabajo también fue descrito en la sección 3.3.

La Tabla 7 recoge el porcentaje de acierto para los tres experimentos, con resultados similares en los tres. La Ilustración 5 (a) muestra la matriz de confusión.

4.1.2 Red Neuronal Convolucional (CNN Convolutional Neural Network)

Los modelos CNN son conocidos por su gran uso en el tratamiento de imágenes y vídeos. Sus capas no están completamente conectadas como en el caso de MLP, pues solo extraen las características simples de la entrada mediante las denominadas operaciones de convolución. Cada capa es un conjunto de funciones no lineales de sumas ponderadas en diferentes coordenadas, pero, esta vez, solamente de las salidas que están espacialmente cerca de la capa anterior y, así, pueden reutilizar los pesos.

Tras separar el conjunto de *train* del de *test* y dotarlos de aleatoriedad, se realizaron unos ajustes previos en ambos para poder hacer uso del modelo básico CNN mostrado en la Ilustración 3. Se empleó la funcionalidad LabelEncoder con el fin de codificar las etiquetas de las emociones en valores numéricos de forma que el modelo lo entienda y cree un mapeado entre ellos en el entrenamiento. También fue necesario cambiar la dimensión del conjunto de variables, x, de una a dos para poder tratarlas como un tensor.

El modelo está formado por una capa convolucional con una forma de entrada de (52,1) y 256 filtros de salida. Se dota de una función de activación ReLU, ya que se recomienda su uso en las capas ocultas [20]. Posteriormente, se aplica un *dropout* de 0,1 para evitar el sobreajuste. Antes de la capa densa, se recurre a una de tipo Flatten de modo que la entrada bidimensional se transforma en unidimensional. Finalmente, se usa la capa densa de 7 unidades, una por cada emoción, con una función de activación Softmax para la distribución categórica. La Ilustración 3 muestra el resumen de la composición del modelo.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_17 (Conv1D)           (None, 52, 256)           1536
_________________________________________________________________
activation_45 (Activation)   (None, 52, 256)           0
_________________________________________________________________
dropout_12 (Dropout)         (None, 52, 256)           0
_________________________________________________________________
flatten_9 (Flatten)          (None, 13312)             0
_________________________________________________________________
dense_25 (Dense)             (None, 7)                 93191
_________________________________________________________________
activation_46 (Activation)   (None, 7)                 0
=================================================================
Total params: 94,727
Trainable params: 94,727
Non-trainable params: 0
_________________________________________________________________
```

Ilustración 3. Arquitectura del modelo básico CNN

Tal y como se observa en la Tabla 6, el modelo CNN resulta ser el más eficiente de los tres, por lo que será la base de la arquitectura del modelo final. Igualmente, la Figura 5 (b) muestra la matriz de confusión.

4.1.3 Long Short-Term Memory (LSTM)

La capa *Long Short-Term Memory* introdujo el concepto de memoria en las redes neuronales. De esta forma, pueden aprender de experiencias importantes que pasaron anteriormente. Se usa fundamentalmente en reconocimiento de voz y en composición musical.

El tratamiento previo de los datos se realiza de la misma manera que para el modelo CNN. La composición también es análoga, sustituyendo la primera capa convolucional por una LSTM. Asimismo, su resumen se presenta en la Ilustración 4.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_13 (LSTM)               (None, 256)               264192
_________________________________________________________________
activation_17 (Activation)   (None, 256)               0
_________________________________________________________________
flatten_2 (Flatten)          (None, 256)               0
_________________________________________________________________
dense_12 (Dense)             (None, 7)                 1799
_________________________________________________________________
activation_18 (Activation)   (None, 7)                 0
=================================================================
Total params: 265,991
Trainable params: 265,991
Non-trainable params: 0
_________________________________________________________________
```

Ilustración 4. Arquitectura del modelo básico LSTM

Pese a que el resultado es el más bajo (Tabla 7), su uso junto con las capas CNN podría mejorar la eficiencia de los modelos [4,6,20].

Tabla 7. Porcentaje de acierto medio de la clasificación de train y test para los distintos modelos básicos escogidos

Modelo básico	Train accuracy (%)	Test accuracy (%)
MLP	73,8	62
CNN	65,2	63
LSTM	62,7	61

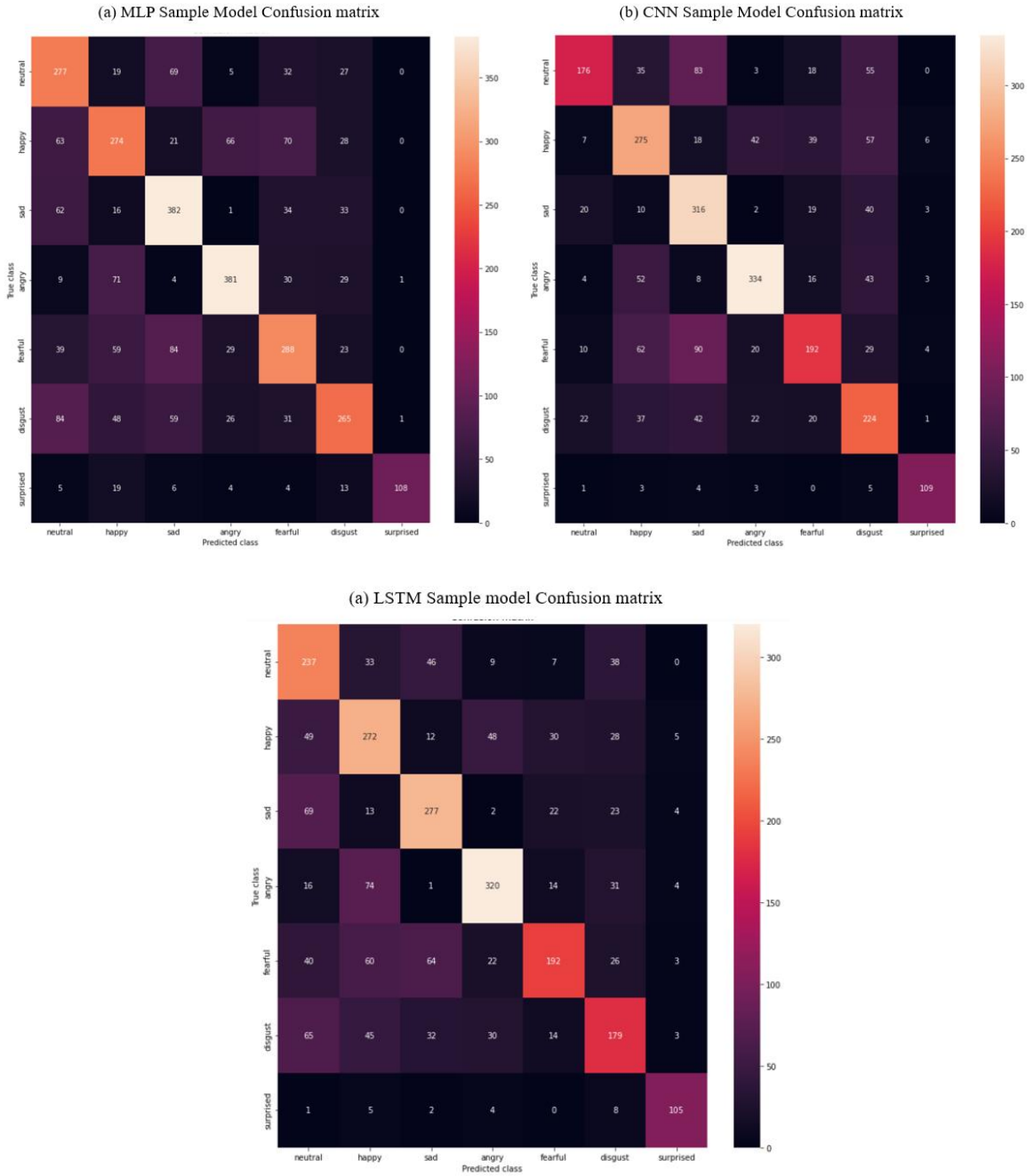


Ilustración 5. Matrices de confusión de los modelos básicos: (a) MLP, (b) CNN y (c) LSTM

4.2 Modelos mixtos

Tras haber adquirido una idea de la actuación de las capas básicas para el entrenamiento (Tabla 6 e Ilustración 5), se procede a realizar pruebas en modelos que incluyan capas de los tres tipos y/o varias del mismo tipo. La capa CNN fue la que mejor resultado aportó, por lo que se usará de base para la elección del modelo. Existía incertidumbre con el uso de las capas LSTM por sí solas. Son empleadas en multitud de proyectos otorgando exitosos resultados, como en el reconocimiento de emociones en la música [6], y no tan buen

comportamiento en el reconocimiento de emociones en la voz [5]. Se ejecutó una prueba usando varias capas LSTM y se observó cómo las capacidades de generalización eran inferiores que en los modelos donde había varias capas CNN, corroborando los resultados de [5] para un corpus diferente.

Siguiendo las praxis de las investigaciones anteriores, se diseñaron tres modelos: (1) exclusivamente formado por capas CNN; (2) modelo mixto sustituyendo una capa CNN por una LSTM a las del (1); y (3) modelo mixto bidireccional (BLSTM, *Bidirectional Long Short-Term Memory*) formado por 3 capas CNN y 2 BLSTM [21].

Los parámetros con los que se compilan y entrenan los modelos se presentan en la Tabla 8. La función de pérdida usada es la de entropía cruzada (categorical_crossentropy), la empleada en las tareas de clasificación en las que un ejemplo solo puede pertenecer a una de las categorías posibles y el modelo debe decidir cuál. El optimizador se emplea para minimizar la función de coste encontrando los pesos adecuados para cada parte de la red, asegurando una mejor generalización, se escoge Adam por ser el más reciente y esperar un rendimiento superior [20]. Las métricas serán por tasa de acierto y el tamaño del lote de 20 o 100 durante 250 o 300 épocas, dependiendo del modelo.

Tabla 8. Parámetros para la compilación y entrenamiento de los modelos

loss	**optimizer**	**metrics**	**batch_size**	**epochs**
categorical_crossentropy	adam	accuracy	20/100	250/300

Para el diseño de la red formada exclusivamente por capas CNN unidimensionales, se observó que se producía sobreajuste notable cuando solo había una capa o más de 4. La Ilustración 6 muestra la arquitectura del modelo CNN más preciso. Se trata de una arquitectura similar a la del modelo básico convolucional (Ilustración 3) con mayor cantidad de capas CNN, 4 en total. El tratamiento previo del conjunto de datos es el propio explicado en la sección 4.1.2. Además de las capas usadas en el modelo básico, se incluye una capa Max Pooling para reducir el número de variables y, por ende, el sobreajuste y el coste computacional del modelo. La Ilustración 6 muestra el detalle de su arquitectura. Tal y como detalla la Tabla 9 y la figura 9 (a), es el modelo más deficiente de los tres.

Layer (type)	Output Shape	Param #
conv1d_128 (Conv1D)	(None, 52, 256)	1536
activation_202 (Activation)	(None, 52, 256)	0
conv1d_129 (Conv1D)	(None, 52, 128)	163968
activation_203 (Activation)	(None, 52, 128)	0
dropout_53 (Dropout)	(None, 52, 128)	0
max_pooling1d_34 (MaxPooling	(None, 6, 128)	0
conv1d_130 (Conv1D)	(None, 6, 128)	82048
activation_204 (Activation)	(None, 6, 128)	0
conv1d_131 (Conv1D)	(None, 6, 128)	82048
activation_205 (Activation)	(None, 6, 128)	0
activation_206 (Activation)	(None, 6, 128)	0
flatten_36 (Flatten)	(None, 768)	0
dense_52 (Dense)	(None, 7)	5383
activation_207 (Activation)	(None, 7)	0

Total params: 334,983
Trainable params: 334,983
Non-trainable params: 0

Ilustración 6. Arquitectura del modelo CNN

Con el objetivo de mejorar el modelo anterior corrigiendo el sobreajuste, se le añade una capa LSTM componiendo el modelo de la Ilustración 7. Este modelo obtiene uno de los mejores porcentajes de acierto (Tabla 9). Las curvas de pérdida y precisión de la Ilustración 9 (b) confirman un buen rendimiento de trabajo.

Layer (type)	Output Shape	Param #
conv1d_46 (Conv1D)	(None, 52, 128)	768
activation_94 (Activation)	(None, 52, 128)	0
conv1d_47 (Conv1D)	(None, 52, 256)	164096
activation_95 (Activation)	(None, 52, 256)	0
dropout_24 (Dropout)	(None, 52, 256)	0
conv1d_48 (Conv1D)	(None, 52, 256)	327936
activation_96 (Activation)	(None, 52, 256)	0
max_pooling1d_12 (MaxPooling	(None, 6, 256)	0
lstm_36 (LSTM)	(None, 32)	36992
activation_97 (Activation)	(None, 32)	0
flatten_19 (Flatten)	(None, 32)	0
dense_35 (Dense)	(None, 7)	231
activation_98 (Activation)	(None, 7)	0

Total params: 530,023
Trainable params: 530,023
Non-trainable params: 0

Ilustración 7. Arquitectura del modelo CNN-LSTM

Finalmente, se contempla un modelo con capas BLSTM debido a la mejora que produce en las redes neuronales profundas [22]. En este experimento se obtuvieron resultados similares al modelo formado por capas CNN, con un *batch* de 100 en lugar de 20 (Tabla 8). Su composición está descrita en la Ilustración 8. En este modelo las curvas de pérdida y precisión de *train* y *test* coinciden más, Ilustración 9 (c), aunque sin logar el porcentaje de precisión del modelo CNN-LSTM (Tabla 9).

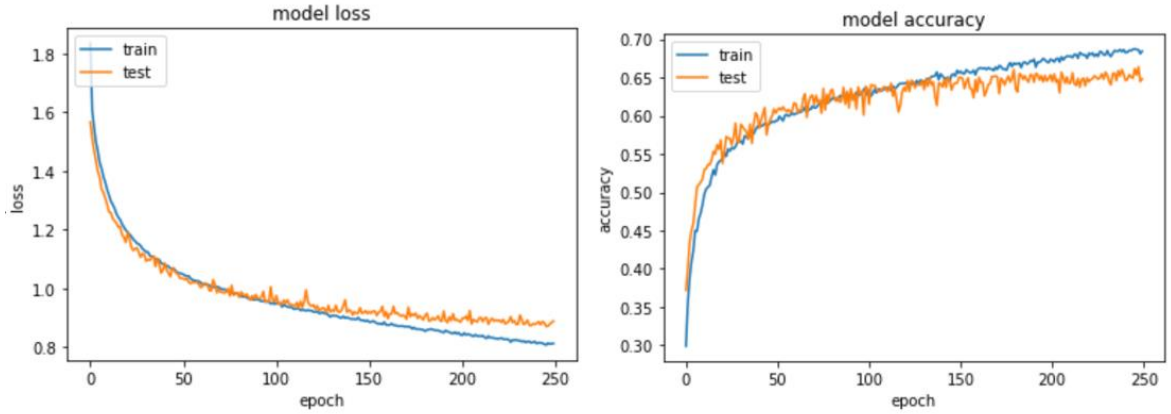
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_118 (Conv1D)          (None, 52, 128)           768
_________________________________________________________________
activation_187 (Activation)  (None, 52, 128)           0
_________________________________________________________________
conv1d_119 (Conv1D)          (None, 52, 256)           164096
_________________________________________________________________
activation_188 (Activation)  (None, 52, 256)           0
_________________________________________________________________
dropout_49 (Dropout)         (None, 52, 256)           0
_________________________________________________________________
conv1d_120 (Conv1D)          (None, 52, 256)           327936
_________________________________________________________________
activation_189 (Activation)  (None, 52, 256)           0
_________________________________________________________________
max_pooling1d_30 (MaxPooling (None, 6, 256)            0
_________________________________________________________________
bidirectional_8 (Bidirection (None, 6, 128)            164352
_________________________________________________________________
bidirectional_9 (Bidirection (None, 128)               98816
_________________________________________________________________
activation_190 (Activation)  (None, 128)               0
_________________________________________________________________
flatten_33 (Flatten)         (None, 128)               0
_________________________________________________________________
dense_49 (Dense)             (None, 7)                 903
_________________________________________________________________
activation_191 (Activation)  (None, 7)                 0
=================================================================
Total params: 756,871
Trainable params: 756,871
Non-trainable params: 0
_________________________________________________________________
```

Ilustración 8. Arquitectura del modelo CNN-BLSTM

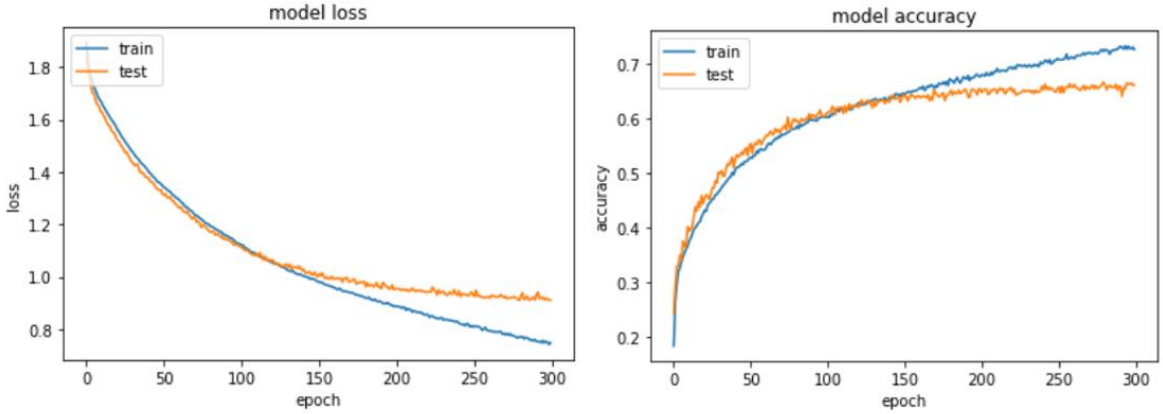
Tabla 9. Porcentaje de acierto medio de la clasificación de train y test para los distintos modelos mixtos

Modelo	Train accuracy (%)	Test accuracy (%)
CNN	68,4	64,78
CNN-LSTM	72	67
CNN-BLSTM	65,27	65,59

(a) Curvas Modelo CNN



(b) Curvas Modelo CNN-LSTM



(c) Curvas Modelo CNN-BLSTM

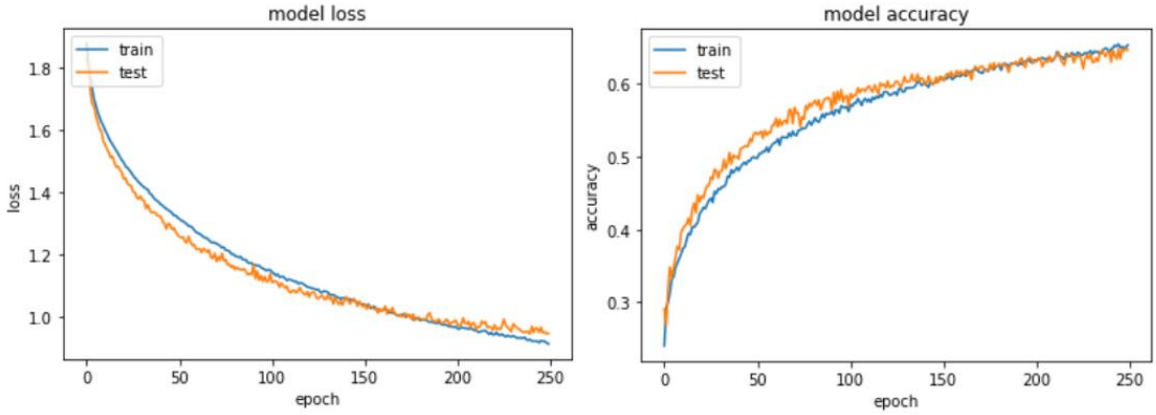


Ilustración 9. Curvas de pérdida y precisión de los modelos: (a) CNN, (b) CNN-LSTM y (c) CNN-BLSTM

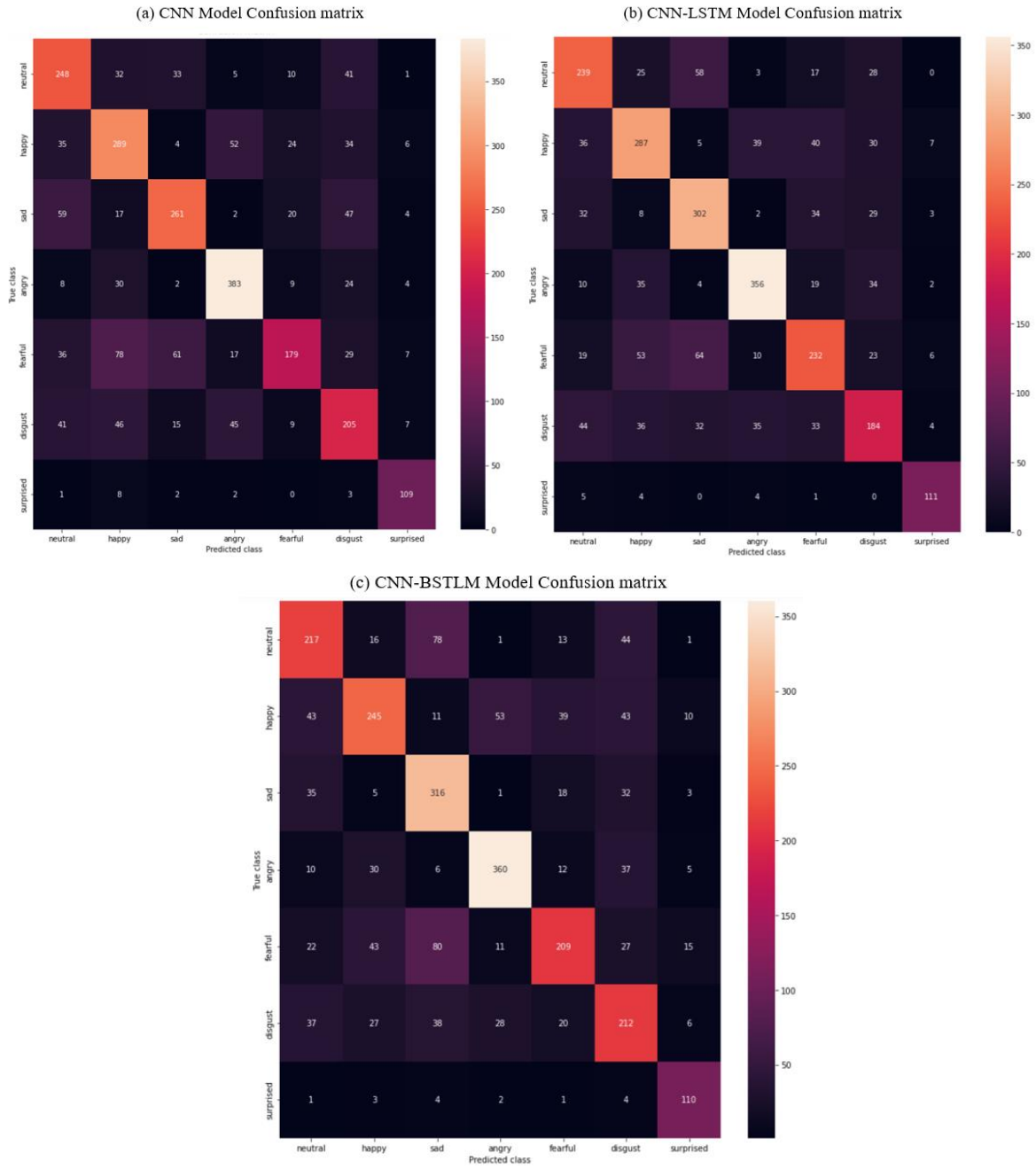


Ilustración 10. Matrices de confusión de los modelos: (a) CNN, (b) CNN-LSTM y (c) CNN-BLSTM

5 Resultados y análisis

Todos los modelos fueron entrenados con 52 variables por audio del corpus descrito en el capítulo 2. Se intentaron evitar altas cargas computacionales, ya que la ejecución era local, tardando del orden de 1 hora por cada entrenamiento.

La Tabla 10 recoge el número de emociones reales de cada tipo y la predicción propuesta por cada modelo. Esto se estudia haciendo uso de las matrices de confusión de la Ilustración 10. Los modelos analizados operan de manera similar tanto en el acierto como en el fallo. El porcentaje de reconocimiento de los tres modelos es alto para las emociones: enfado, tristeza, felicidad y neutralidad y es menor para miedo, disgusto y sorpresa. Causado, probablemente, por una mezcla entre disponer de menos registros para estos últimos y que las características seleccionadas no consiguen clasificarlos con tanta precisión como los primeros.

El modelo que mejor comportamiento presentó es el formado por 3 capas CNN y una LSTM, alcanzando cerca de un 70% de precisión. Al tratarse de una red neuronal pesada el tiempo de ejecución incrementa, aunque menos que el del modelo CNN-BLSTM.

Tabla 10. Número de emociones de cada tipo real y predicción de los modelos

Modelo	**Neutralidad**	**Felicidad**	**Tristeza**	**Enfado**	**Miedo**	**Disgusto**	**Sorpresa**
Nº Total	370	444	410	460	407	368	125
CNN	428	500	378	506	251	383	138
CNN-LSTM	385	448	465	449	376	328	133
CNN-BLSTM	365	369	533	456	312	399	150

6 Conclusiones y trabajo futuro

Queda claro que el reconocimiento de emociones en la voz, SER, es uno de los grandes desafíos que están aún sin resolver. El reto que desencadena esta disciplina es, principalmente, la búsqueda de las características audibles de la voz que clasifiquen las emociones. Gracias a la existencia de las redes neuronales profundas es posible hacer varias aproximaciones para resolver este problema. No obstante, aún queda lejos de alcanzar los rendimientos de otros sistemas más avanzados, como el reconocimiento de voz. Se abre una línea de trabajo conjunto que engloba a expertos de sonido, estadísticos, psicólogos e informáticos, entre otros, para hacer frente a todos los desafíos.

Clasificar las emociones no es tarea fácil, pues no es posible medirlas cuantitativamente. Por lo que detectar emociones en la voz no solo depende de las características del sonido, sino también del ser humano y de otros factores, como el idioma o la cultura. Sin embargo, en esta memoria se realiza un acercamiento al intento de clasificarlas cuantitativamente con características audibles. Se escogieron un total de 52 variables, formadas por la media y la desviación típica de los 13 primeros coeficientes de Cepstral, 6 tonos y 7 centroides de espectro extraídos del audio. Este conjunto mejoró del orden de un 10 % los resultados obtenidos en otras investigaciones. Se propone usar otros estadísticos para la medición de estas variables y encontrar la forma de evaluar la característica Tempo de manera más reducida, ya que se observó que podría contener información valiosa para la categorización.

El diseño de los modelos estuvo relacionado con el corpus utilizado. Al estar compuesto por un conjunto de bases de datos diferentes se estudió la posibilidad de tener contradicciones en la catalogación de las emociones. Aun así, se obtuvo una sustancial mejora en la predicción de emociones gracias, probablemente, al incremento del número de audios catalogados en inglés. Se plantea el uso de un corpus con más registros que incluyan otro tipo de variables para estudiar la influencia que estas tienen en la detección de emociones y, así, mejorar la eficiencia de los modelos.

Finalmente, se corrobora la importancia de un buen diseño de la red neuronal. Se subraya las mejoras de rendimiento y predicción cuando trabajan en conjunto las capas CNN con las LSTM para los sistemas de reconocimiento. Se concluye con la posible mejora que

supondría analizar a la par la transcripción de la voz y las características del audio teniendo optimizados ambos sistemas.

Referencias

[1] Petrushin, V. (1999, November). Emotion in speech: Recognition and application to call centers. In *Proceedings of artificial neural networks in engineering* (Vol. 710, p. 22).

[2] Sezgin, M. C., Gunsel, B., & Kurt, G. K. (2012). Perceptual audio features for emotion detection. *EURASIP Journal on Audio, Speech, and Music Processing*, *2012*(1), 1-21.

[3] Navarro, E. C. (2013). El lenguaje no verbal: un proceso cognitivo superior indispensable para el ser humano. *Revista comunicación*, *20*(1 (2011)), 46-51.

[4] Vidrascu, L., & Devillers, L. (2007, August). Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features. In *Proc. Inter. workshop on Paralinguistic Speech between models and data, ParaLing*.

[5] Parry, J., Palaz, D., Clarke, G., Lecomte, P., Mead, R., Berger, M., & Hofer, G. (2019, September). Analysis of Deep Learning Architectures for Cross-Corpus Speech Emotion Recognition. In *INTERSPEECH* (pp. 1656-1660).

[6] Atmaja, B. T., & Akagi, M. (2020, November). On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers. In *2020 IEEE REGION 10 CONFERENCE (TENCON)* (pp. 968-972). IEEE.

[7] Atmaja, B. T., Shirai, K., & Akagi, M. (2019, November). Speech emotion recognition using speech feature and word embedding. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 519-523). IEEE.

[8] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015, July). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference* (Vol. 8, pp. 18-25).

[9] Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, *5*(4), 377-390.

[10] Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto emotional speech set (TESS).

[11] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, *13*(5), e0196391.

[12] Haq, S. U. (2011). *Audio visual expressed emotion classification*. University of Surrey (United Kingdom).

[13] Zhang, L., Walter, S., Ma, X., Werner, P., Al-Hamadi, A., Traue, H. C., & Gruss, S. (2016, December). "BioVid Emo DB": A multimodal database for emotion analyses validated by subjective ratings. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1-6). IEEE.

[14] Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. *Information Fusion*, *59*, 103-126.

[15] Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, *6*(3-4), 169-200.

[16] Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161.

[17] Thayer, R. E. (1990). *The biopsychology of mood and arousal*. Oxford University Press.

[18] Breebaart, J., & McKinney, M. F. (2004). Features for audio classification. In *Algorithms in Ambient Intelligence* (pp. 113-129). Springer, Dordrecht.

[19] Hildebrand, C., Efthymiou, F., Busquet, F., Hampton, W. H., Hoffman, D. L., & Novak, T. P. (2020). Voice analytics in business research: Conceptual foundations, acoustic feature extraction, and applications. *Journal of Business Research*, *121*, 364-374.

[20] Sze, V., Chen, Y. H., Yang, T. J., & Emer, J. S. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, *105*(12), 2295-2329.

[21] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

[22] Frinken, V., & Uchida, S. (2015, August). Deep BLSTM neural networks for unconstrained continuous handwritten text recognition. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)* (pp. 911-915). IEEE.