Informe

Azahara Martinez, María de los Ángeles Díaz, Álvaro Nieva, Iyán Álvarez, Florencia Pellegrini, óscar Ca

2025-12-27

OBJETIVOS DEL PROYECTO

El objetivo principal de este trabajo es analizar señales de voz a partir de grabaciones propias y desarrollar modelos de clasificación capaces de identificar distintas características de la señal de audio. Para ello, se parte de un conjunto de audios etiquetados y se emplean técnicas de preprocesado, extracción de características y aprendizaje automático, con el fin de obtener modelos que puedan generalizar su comportamiento ante audios no utilizados durante el entrenamiento.

De manera más específica, los objetivos del trabajo son los siguientes:

- Analizar señales de voz grabadas por los propios autores y prepararlas para su posterior tratamiento mediante técnicas de preprocesado.
- Extraer características acústicas relevantes que permitan describir distintas propiedades de la voz de forma numérica.
- Construir un conjunto de datos estructurado a partir de las características extraídas y las etiquetas conocidas de cada audio.
- Desarrollar y entrenar modelos de clasificación capaces de identificar el sexo del hablante a partir de la señal de voz.
- Desarrollar modelos de clasificación para la identificación del acento, distinguiendo entre español neutro, andaluz y argentino.
- Investigar la viabilidad de diferenciar entre voces humanas y voces generadas por inteligencia artificial a partir de descriptores acústicos.
- Evaluar el rendimiento de los modelos propuestos mediante técnicas de validación adecuadas.

PREPROCESADO

Conversión y lectura de audios

Como primer paso del preprocesado, se desarrolló una función para la conversión de archivos de audio desde el formato `.m4a` al formato `.wav`, con el objetivo de unificar el tipo de señal y facilitar su posterior análisis. Para ello, se emplearon las librerías `av` y `tuneR` de R, que permiten la conversión y lectura de señales de audio de forma eficiente.

La función implementada convierte cada archivo a una señal monofónica con una frecuencia de muestreo de 16 kHz y guarda los archivos resultantes en un directorio específico. Además, se automatizó el proceso para convertir de manera recursiva todos los archivos `.m4a` contenidos en una carpeta y sus subdirectorios.

Detección y eliminación de ruido y silencios

Con el objetivo de mejorar la calidad de las señales de voz antes de la extracción de características, se implementó un módulo de detección y eliminación de ruido/silencios basado en medidas temporales por tramas. El audio se segmenta en ventanas cortas (20 ms) con solapamiento del 50%, lo que permite analizar de forma local la actividad de la señal.

Para cada trama se calcularon dos descriptores:

- **Zero Crossing Rate (ZCR):** tasa de cruces por cero, asociada al contenido de alta frecuencia y útil para caracterizar señales ruidosas o fricativas.
- **Energía de corto tiempo (STE):** suma de cuadrados de la señal en cada ventana, normalizada en el intervalo $[0, 1]$, empleada como indicador de presencia de voz frente a silencio o fondo.

La detección de ruido se realizó aplicando un criterio umbral sobre la energía: aquellas tramas con energía inferior a un umbral predefinido se etiquetaron como *Ruido/Silencio*, mientras que el resto se consideraron *Voz/Señal*. De forma opcional, se contempló un segundo criterio basado en ZCR para detectar segmentos con alta tasa de cruces por cero y energía baja-media, característicos de siseos o ruido de viento.

A partir de la máscara temporal resultante, se reconstruyó una señal limpia concatenando únicamente las muestras correspondientes a tramas etiquetadas como voz. Finalmente, se compararon visualmente las formas de onda original y procesada y se exportó el audio resultante en formato `.wav` para su uso en las etapas posteriores del sistema.

EXTRACCIÓN DE CARACTERÍSTICAS

Las señales de voz contienen mucha información, pero trabajar directamente con el audio no resulta práctico para entrenar modelos de clasificación. Por ello, es necesario extraer una serie de características numéricas que describan la señal de forma más simple y manejable.

La extracción de características permite resumir aspectos importantes de la voz, como su tono, su energía o cómo se distribuye el sonido en distintas frecuencias. De esta manera, cada audio puede representarse mediante un conjunto de valores que capturan sus propiedades principales, evitando depender de la duración del archivo o de pequeñas variaciones que no aportan información relevante.

Estas características sirven como base para entrenar los modelos utilizados en el proyecto, ya que facilitan la comparación entre distintas voces y permiten identificar patrones asociados a diferentes tipos de hablantes o señales. Gracias a este proceso, es posible abordar tareas de clasificación como la identificación del sexo, el acento o la distinción entre voces humanas y generadas por inteligencia artificial.

PERIOD PITCH

Con el objetivo de diferenciar entre voces masculinas y femeninas, se empleó el *period pitch* de la señal de voz como descriptor principal. Este parámetro está directamente relacionado con la frecuencia fundamental de la voz, que corresponde a la vibración periódica más baja producida por las cuerdas vocales y constituye la base sobre la que se construyen el resto de componentes armónicos del habla. La frecuencia fundamental es especialmente relevante porque representa una característica estable de la voz y está estrechamente vinculada a la percepción de si una voz suena más grave o más aguda.

El period pitch se estimó mediante un método basado en la autocorrelación de la señal, que permite detectar la periodicidad dominante en los segmentos sonoros del audio, principalmente en vocales. A partir del retardo correspondiente al máximo de la autocorrelación se obtiene el período fundamental de la señal, y su

inversa proporciona una estimación de la frecuencia fundamental o pitch, expresada en hercios (Hz), es decir, vibraciones por segundo.

Diversos estudios en la literatura han documentado rangos característicos de frecuencia fundamental para voces masculinas y femeninas. En particular, Rabiner y Schafer describen que las voces masculinas suelen situarse aproximadamente entre 85 y 180 Hz, mientras que las voces femeninas presentan valores más elevados, entre 165 y 255 Hz (Rabiner, 1978). Resultados similares se reportan en estudios más recientes, donde se observa que las voces masculinas tienden a concentrarse en el rango de 90 a 150 Hz, mientras que las voces femeninas se sitúan aproximadamente entre 190 y 240 Hz (Basu, 2020).

A partir de estos rangos reportados en la literatura, se definieron umbrales prácticos para la clasificación del sexo basados en el valor del *pitch*. Con el fin de evitar clasificaciones forzadas en zonas de solapamiento entre ambos grupos, se estableció una región intermedia de indeterminación. De este modo, se consideraron voces masculinas aquellas con frecuencias fundamentales comprendidas entre 85 y 170 Hz, y voces femeninas aquellas con valores entre 180 y 240 Hz.

Para calcular el pitch creamos la siguiente función. Con tal de garantizar la fiabilidad de la estimación, se incorporan una serie de comprobaciones previas sobre la señal de entrada. En primer lugar, se verifica que el audio contenga un número suficiente de muestras, ya que señales excesivamente cortas no permiten identificar una periodicidad clara. Asimismo, la búsqueda del período fundamental se restringe a un rango de retardos coherente con los valores esperados de frecuencia fundamental en voz humana, evitando así detecciones erróneas asociadas a componentes no relevantes de la señal.

En aquellos casos en los que la longitud efectiva de la señal no permite cubrir dicho rango de retardos, el cálculo del pitch se considera no válido. Esta decisión se basa en el hecho de que los métodos basados en autocorrelación solo proporcionan estimaciones fiables en segmentos sonoros suficientemente largos y con una periodicidad bien definida, como ocurre principalmente en las partes vocales del habla.

ZCR/Energía

Coeficientes Mel (MFCC) y representación de la señal

Como parte de la extracción de características espectrales, se calcularon los coeficientes cepstrales en escala Mel (MFCC), ampliamente utilizados en tareas de análisis de voz por su capacidad de capturar información relacionada con el timbre y la envolvente espectral. Para ello, se aplicó el procedimiento estándar por ventanas cortas: se dividió cada audio en tramas solapadas (ventana de 25 ms y salto de 10 ms), se realizó la transformada rápida de Fourier (FFT) por trama y se obtuvo el espectro de potencia. Posteriormente, dicho espectro se proyectó sobre un banco de filtros en escala Mel (40 bandas), se aplicó una compresión logarítmica y, finalmente, una transformada discreta del coseno (DCT) para obtener un conjunto compacto de coeficientes (12 MFCC por trama). Adicionalmente, se aplicó un pre-énfasis para compensar la caída de energía en altas frecuencias típica de señales de voz.

Con el fin de incorporar información dinámica, se calcularon también las derivadas temporales de primer orden (Δ) y segundo orden ($\Delta\Delta$), que aproximan respectivamente la *velocidad* y la *aceleración* de los MFCC a lo largo del tiempo.

Para adaptar esta representación por tramas a modelos clásicos de aprendizaje automático, cada audio se transformó en un único vector de características agregando estadísticas globales: media y desviación típica de los MFCC, así como de sus derivadas Δ y $\Delta\Delta$. De este modo, cada clip queda representado por un vector fijo que resume tanto el contenido espectral promedio como su variabilidad temporal.

Centroide espectral

Con el objetivo de diferenciar entre voces humanas y voces generadas por inteligencia artificial, se analizó el centroide espectral de las señales de audio. Este descriptor proporciona una medida de la frecuencia media ponderada por la energía del espectro y está relacionado con la distribución espectral de la señal.

El análisis mostró que las voces generadas por IA tienden a presentar centroides espectrales más estables y, en muchos casos, más bajos que los de las voces humanas. Esto se debe a que las señales sintéticas suelen carecer de micro-ruido, irregularidades y fenómenos de fricación propios de la producción vocal humana, presentando un espectro más limpio y concentrado en determinadas bandas de frecuencia. Por el contrario, la voz humana exhibe una mayor dispersión espectral debido a la respiración, turbulencia y variabilidad natural del habla.

A partir de estas observaciones, se estudió la separabilidad entre clases utilizando diagramas de caja, diferenciando además por sexo. En el caso de las voces masculinas, se observó una clara separación entre las clases humana e IA, sin solapamiento apreciable, lo que permitiría una clasificación casi perfecta mediante un umbral adecuado. En cambio, para las voces femeninas se detectó cierto solapamiento entre ambas clases, implicando la posible aparición de errores de clasificación.

Para formalizar este proceso, se calibraron umbrales óptimos del centroide espectral de manera independiente para voces masculinas y femeninas, maximizando métricas de rendimiento como la exactitud y la medida F1. Finalmente, se evaluó el rendimiento del clasificador basado en umbrales mediante una matriz de confusión, confirmando un buen desempeño global, especialmente en el caso de las voces masculinas.

CONSTRUCCIÓN DEL CONJUNTO DE DATOS

Para organizar y analizar los 466 audios, se creó un dataframe que contiene las características y etiquetas de cada audio, donde denotamos:

- M: masculino / F: femenino
- P: persona / IA: IA
- AN: andaluz / AR: argentino / N: neutro

Primero, se creó un dataframe que etiquetara cada persona con su sexo, origen y acento, mediante la función `etiquetar_por_nombre()`, que compara el nombre del archivo con los nombres del dataframe de reglas y devuelve las etiquetas correspondientes.

A continuación, para cada audio se incluyeron las siguientes variables: duración, zcr, energía rms y pitch. Para ello se utilizó la función `añadir_voz()` la cuál recibe un archivo de audio, extrae sus características, asigna las etiquetas usando la función anterior y añade toda la información como una nueva fila al dataframe. Así, cada fila del dataframe corresponde a un audio individual, con sus características numéricas y etiquetas.

Cabe destacar que todas las características consideradas en el conjunto de datos están definidas como medias o medidas normalizadas respecto a la longitud de la señal, de modo que no dependen de la duración total del audio. En consecuencia, el proceso de eliminación de silencios únicamente reduce las regiones no informativas de la señal, sin afectar a la coherencia ni a la comparabilidad de las características extraídas entre distintos audios.

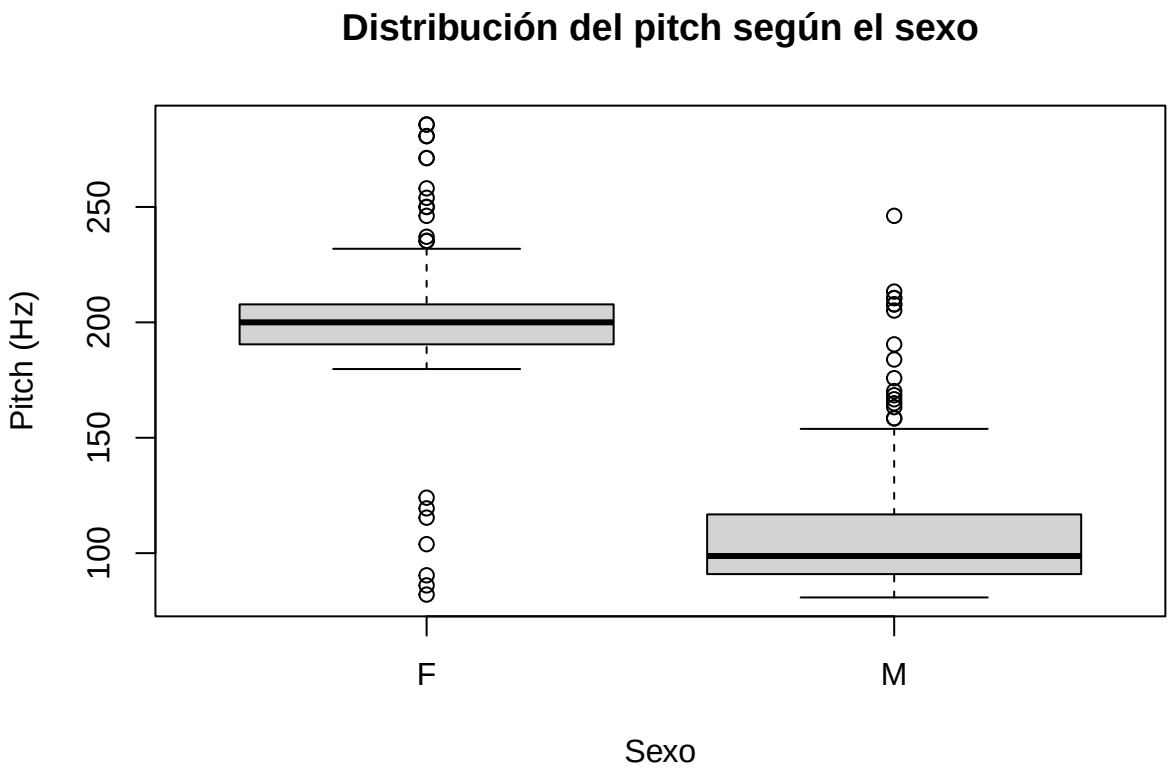
Para automatizar la creación del dataframe, se implementó la función `df_carpeta()`, que recorre todos los archivos de la carpeta con los audios limpios y aplica `añadir_voz()` a cada uno. De forma que, se obtiene un dataframe completo, voces_df, con toda la información lista para análisis y clasificación.

ANÁLISIS EXPLORATORIO

Antes de aplicar los métodos de clasificación, se realizó un análisis exploratorio del conjunto de datos con el objetivo de comprender la distribución de las variables y detectar posibles patrones relevantes. Así, el conjunto de datos final está compuesto por 466 audios, con una distribución equilibrada entre voces femeninas (240) y masculinas (226), lo que reduce la posibilidad de sesgos derivados del desbalanceo de clases.

Tras hacer un análisis descriptivo usando `summary()`, se puede ver que la duración de los audios después de limpiarlos y eliminar los silencios presenta una variabilidad moderada, con valores comprendidos aproximadamente entre 1.5 y 5 segundos. Por otro lado, las características ZCR y energía RMS presentan rangos consistentes y valores medios estables, indicando señales con contenido sonoro continuo y sin presencia dominante de ruido.

En cuanto a la frecuencia fundamental (pitch), se observa un rango amplio de valores (aproximadamente entre 80 y 285 Hz), consistente con voces humanas. El análisis por sexo revela que las voces femeninas presentan valores de pitch más elevados que las masculinas. Si lo representamos usando un diagrama de cajas, podemos ver un solapamiento entre ambas distribuciones, lo que indica que probablemente existan errores de clasificación en algunos audios concretos al emplear únicamente esta característica.



Por último, el análisis de correlación entre las variables numéricas muestra una alta correlación entre la duración y la ZCR, mientras que el pitch presenta una correlación prácticamente nula con el resto de variables. Esto sugiere que el pitch aporta información complementaria y relevante para la tarea de clasificación.

```
##              duracion  zcr energia_rms period_pitch
## duracion         1.00 0.99        0.20         0.01
## zcr              0.99 1.00        0.21         0.04
## energia_rms      0.20 0.21        1.00         0.00
## period_pitch     0.01 0.04        0.00         1.00
```

MODELOS DE CLASIFICACIÓN Y VALIDACIÓN

Clasificación del sexo

A partir del conjunto de características extraídas y del análisis exploratorio realizado, tenemos como objetivo la clasificación del sexo del hablante. De esta forma, se plantea el problema como una clasificación binaria y se emplea un modelo de regresión logística. Este tipo de modelo permite relacionar las características extraídas con la probabilidad de que una voz corresponda a un hablante masculino o femenino.

Para ello, se codificó la variable objetivo `sexo` en formato numérico, asignando el valor 1 a las voces femeninas y 0 a las masculinas. Esta transformación permite el uso de modelos probabilísticos basados en regresión logística.

Inicialmente, se ajustó un modelo de regresión logística multivariante que incluía todas las características disponibles en el conjunto de datos: duración, tasa de cruces por cero (ZCR), energía RMS, frecuencia fundamental (pitch), así como las variables categóricas de origen y acento. Este primer modelo tiene como objetivo evaluar la contribución conjunta de todas las variables y explorar posibles relaciones entre ellas y la variable respuesta.

Al analizar el resumen del modelo de regresión logística completo mediante la función `summary()`, se observó que la única variable estadísticamente significativa para la clasificación del sexo era la frecuencia fundamental (period pitch), mientras que el resto de variables no aportaban evidencia significativa una vez incorporado dicho descriptor. A partir de esta observación, se decidió contrastar un modelo reducido basado exclusivamente en el pitch frente al modelo completo mediante un ANOVA, con el fin de comprobar si la complejidad adicional estaba justificada.

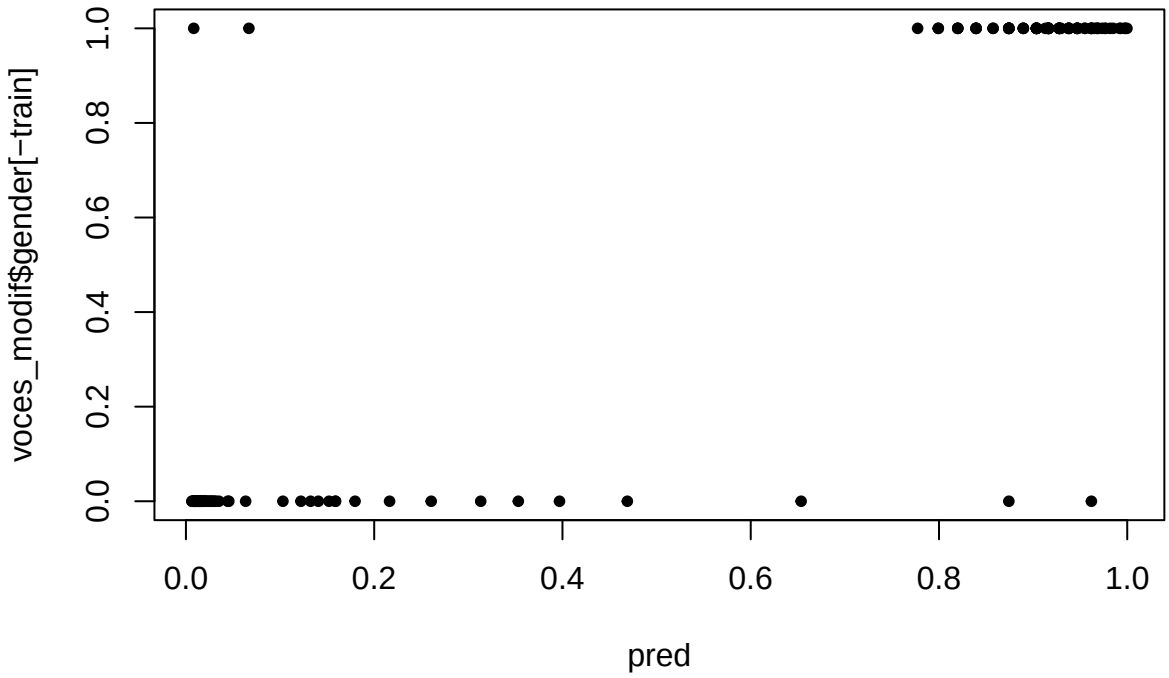
```
## Analysis of Deviance Table
##
## Model 1: gender ~ period_pitch
## Model 2: gender ~ sexo + origen + acento + duracion + zcr + energia_rms +
##     period_pitch
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       464     186.78
## 2       457       0.00  7   186.78 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El contraste mediante ANOVA entre el modelo reducido y el modelo completo muestra que incorporar el resto de variables produce una reducción significativa de la devianza ($p < 0.001$). No obstante, el análisis individual de los coeficientes revela que únicamente el period pitch resulta estadísticamente significativo, mientras que el resto de variables no aportan información relevante de forma independiente.

A continuación, se evaluó el desempeño del modelo reducido en un esquema de entrenamiento y prueba. Para ello, se seleccionó aleatoriamente un 70% de las observaciones para ajustar el modelo y se reservó el 30% para validación.

La bondad de ajuste del modelo se reflejó en un pseudo-R^2 de 0.697, lo que sugiere que aproximadamente el 69.7% de la variabilidad del conjunto de entrenamiento queda explicada por el modelo basado únicamente en el period pitch. Posteriormente, se aplicó el modelo al conjunto de prueba para evaluar su capacidad predictiva.

La correlación entre las probabilidades predichas y las clases reales alcanzó un valor de 0.918, evidenciando una buena concordancia entre las predicciones y las etiquetas de sexo.



Al observar la gráfica de predicciones, se aprecia que la mayoría de los puntos correspondientes a voces masculinas se agrupan cerca de 0, mientras que las voces femeninas se concentran alrededor de 1, lo que indica que el modelo reconoce correctamente la mayoría de los casos. Sin embargo, algunos puntos se sitúan en valores intermedios, mostrando incertidumbre en la predicción.

Sin embargo, en el análisis exploratorio vimos que el diagrama de cajas del pitch tenía zonas donde había solapamiento entre los rangos de frecuencia fundamental de hombres y mujeres. Por lo tanto, es probable que los audios con pitch en el rango intermedio sean precisamente los que generan errores de clasificación, reflejando la variabilidad natural del tono de voz y características individuales que no se capturan únicamente con el period pitch.

En conjunto, estos resultados refuerzan que el pitch es el descriptor más relevante para la clasificación del sexo, pero que la superposición natural de frecuencias humanas introduce un margen de error inevitable, explicando los puntos intermedios en la gráfica de predicciones.

Clasificación del acento

Además de la clasificación por sexo, se abordó la identificación del acento de la voz, considerando tres categorías: español neutro, andaluz y argentino. Para esta tarea se empleó un modelo de regresión logística multinomial, adecuado para problemas de clasificación con más de dos clases.

Como variables predictoras se utilizaron descriptores acústicos de bajo nivel, concretamente la tasa de cruces por cero (ZCR), la energía RMS y la duración del audio. Estas características se seleccionaron por su relación con aspectos articulatorios y prosódicos que pueden variar entre acentos.

Con el fin de explorar la separabilidad entre clases, se realizó un análisis gráfico de la distribución del ZCR en función del acento, observándose diferencias apreciables entre algunas categorías. Posteriormente, se ajustó el

modelo multinomial y se evaluó su capacidad explicativa mediante el pseudo-R^2 de McFadden, comparando la verosimilitud del modelo completo con la de un modelo nulo.

Aunque los resultados muestran que estas características contienen información relevante para la clasificación del acento, el rendimiento del modelo está condicionado por el tamaño reducido del conjunto de datos, por lo que se considera necesario ampliar el número de muestras para obtener conclusiones más robustas.

Clasificación del origen (IA vs Humano)