Informe

Azahara Martínez, María de los Ángeles Díaz,Álvaro Nieva, Iyán Álvarez,
Florencia Pellegrini, Óscar Camacho

2026-01-14

INTRODUCCIÓN

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. Los audios generados mediante voz sintética (Voz IA) no se incluyen en este proceso, ya que se encuentran originalmente en formato `.wav` y no requieren conversión adicional.

Detección y eliminación de ruido y silencios

Con el objetivo de mejorar la calidad de la base de datos y eliminar segmentos de audio que no aporten información, se implementó un algoritmo para discriminar entre fragmentos correspondientes al habla y segmentos de ruido o silencio, eliminando estos últimos a partir de un análisis de la *Short Time Energy* (STE) y el *Zero Crossing Rate* (ZCR). Este procedimiento genera nuevos archivos de audio que contienen únicamente la información acústica relevante.

El procesamiento se realizó mediante la creación de ventanas temporales, dividiendo la señal de entrada en tramos de 20 ms con un solapamiento del 50% entre ventanas consecutivas. Estos parámetros se seleccionaron para garantizar la cuasi-estacionariedad de la señal de voz, permitiendo un análisis espectral y temporal preciso sin perder continuidad en los bordes de cada tramo.

Para cada ventana se calcularon los dos descriptores mencionados anteriormente empleando `seewave` y `tuneR`:

- **STE**: Se calculó como la suma de los cuadrados de la amplitud de la señal en cada ventana, normalizando en el intervalo $[0, 1]$. Esta métrica actúa como el discriminador principal, asumiendo que los segmentos de voz presentan una energía significativamente superior frente a los ruidos de fondo o silencios.
- **ZCR**: Se computó la frecuencia con la que la señal cambia de signo dentro de la ventana. Aunque el código permite su uso como criterio secundario para filtrar ruidos con alta frecuencia, en la configuración final se priorizó el criterio del STE.

El criterio de decisión se estableció mediante un umbral de energía definido. Aquellos tramos con una energía normalizada inferior a 0.02 (2% del máximo) se etiquetaron como *Ruido/Silencio*, mientras que las superiores se clasificaron como *Voz/Señal*. Para fijar los umbrales de STE y ZCR se observó los valores típicos normalizados que se obtenían en la mayoría de audios y se fijó a un valor en el que tan solo se perdiera la información menos relevante al eliminar los valores inferiores, dejando tan solo la parte con habla en la señal.

En la siguiente figura se muestra la clasificación para uno de los audios. Los puntos de la STE que superan el umbral corresponden a la señal, mientras que los inferiores corresponden a ruido.

Finalmente, se reconstruyó la señal de audio concatenando temporalmente las muestras correspondientes a los segmentos validos, descartando los tramos de silencio y exportando el resultado limpio a la carpeta de destino para su posterior caracterización.

«««< HEAD

EXTRACCIÓN DE CARACTERÍSTICAS

Para el tratamiento de señales de voz en modelos de clasificación, no resulta adecuado trabajar directamente con la señal de audio, ya que estos modelos requieren datos numéricos estructurados. Por este motivo, se lleva a cabo un proceso de extracción de características que permite representar la información relevante de la señal de forma numérica. Dichas características se emplearán posteriormente para el entrenamiento de los modelos.

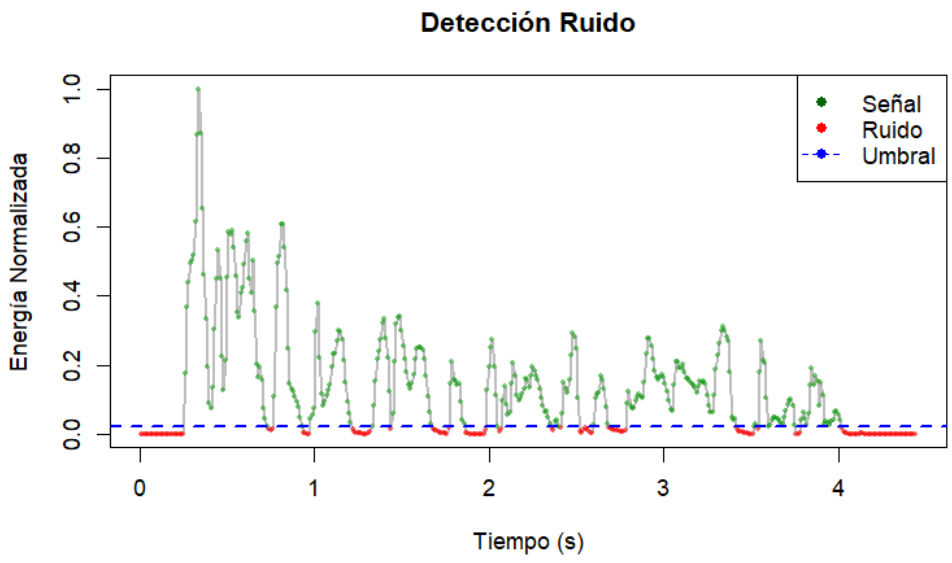

Figure 1: Etiquetado de puntos de ruido o señal a partir de la STE normalizada.

Period pitch

El *period pitch* 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.

Así pues, 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.

De acuerdo con la literatura consultada, los valores típicos de frecuencia fundamental en voz humana se sitúan aproximadamente en el rango comprendido entre 85 y 260 Hz, presentando diferencias claras entre hombres y mujeres ([1] y [2]).

A partir de estos rangos reportados, se definieron umbrales prácticos para la clasificación del sexo. Con el fin de evitar clasificaciones forzadas en las zonas de solapamiento entre ambos grupos, se estableció una región intermedia de indeterminación. De este modo, se consideraron voces masculinas aquellas con frecuencias fundamentales inferiores a 170 Hz, voces femeninas aquellas con valores superiores a 180Hz y como indeterminadas aquellas comprendidas entre ambos umbrales.

Para calcular el pitch creamos la función `period_pitch` y, con el fin 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.

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. Los MFCC (coeficientes Mel) son características que resumen el timbre de la voz a partir de su espectro, usando una escala de frecuencias similar a la percepción humana (escala Mel).

En el código, para calcular estos coeficientes, se usa la función `melfcc()` de la libreria `tuneR`. Esta función divide la señal en ventanas temporales solapadas de duración 25 ms con un salto 10 ms. Para cada ventana se obtiene el espectro (vía FFT), se proyecta sobre un banco de filtros Mel con 40 bandas, se aplica compresión logarítmica y posteriormente una transformación tipo DCT para obtener un número reducido de coeficientes. En este caso se conservan 12 coeficientes por ventana.

Además se calculan Δ y $\Delta\Delta$ son las derivadas temporales de los MFCC: Δ mide cómo cambian de un frame al siguiente (dinámica) y $\Delta\Delta$ mide el cambio de ese cambio.

Centroide espectral

Centroide espectral global

Con el objetivo de caracterizar la distribución de la energía en el dominio de la frecuencia, se calculó el centroide espectral global de cada señal de voz. Este descriptor puede interpretarse como el "centro de gravedad" del espectro, ya que corresponde a la media de las frecuencias ponderada por su energía, de modo que valores más elevados indican una mayor concentración de energía en frecuencias altas. Previamente a su cálculo, se eliminó el valor medio de la señal, correspondiente a la componente de continua, ya que su presencia introduce un componente espectral en 0 Hz que no aporta información acústica relevante y puede sesgar artificialmente el centroide hacia bajas frecuencias.

De esta forma, el centroide se obtiene a partir del espectro de potencia calculado mediante la transformada rápida de Fourier, considerando únicamente las componentes de frecuencia positivas. De este modo, cada señal queda representada por un único valor que resume su contenido espectral global.

Segmentación temporal en 12 partes

Con el fin de capturar información espectral a lo largo del tiempo, se desarrolló una segunda función que divide cada señal de audio en doce segmentos temporales consecutivos de igual duración. Para cada uno de estos segmentos se calcula de forma independiente el centroide espectral siguiendo el mismo procedimiento que en el caso global. Esta estrategia permite representar cada audio mediante un vector de doce valores, cada uno correspondiente a un segmento temporal distinto, describiendo así la evolución del centroide espectral a lo largo de la señal.

La segmentación temporal proporciona una descripción más detallada del contenido espectral, al incorporar información temporal que no queda reflejada en una medida global única.

CONSTRUCCIÓN DEL CONJUNTO DE DATOS

Conjunto de datos general

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, pitch, centroide y centroide en 12 tramos. 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.

Conjunto de datos para detección de IA: MFCC

Para cada archivo de audio se aplicó un preprocesado común con el objetivo de asegurar la comparabilidad entre señales. En concreto, se convirtió cada audio a mono mediante la función `to_mono()`, promediando los canales izquierdo y derecho en caso de que la señal fuese estéreo. Después, se homogeneizó la frecuencia de muestreo a un valor fijo (16 kHz) mediante la función `resample_if_needed()`, ya que el cálculo de MFCC depende directamente de la escala de frecuencias y no es comparable si los audios están a distintas tasas de muestreo.

Una vez normalizada la señal, se extrajeron los coeficientes MFCC utilizando una configuración fija de ventanas temporales (duración de ventana y salto constantes) y un número determinado de coeficientes cepstrales, para ello se uso la función `melfcc()` como se ha comentado anteriormente. Dado que los MFCC se calculan por frames y generan una matriz temporal, se incorporó además información dinámica calculando las derivadas temporales: (Δ) (delta) y ($\Delta\Delta$) (delta-delta), mediante la función `delta_simple()`, capturando así los cambios espectrales a lo largo del tiempo.

Para transformar estas matrices (MFCC, (Δ) y ($\Delta\Delta$)) en un conjunto de variables de tamaño fijo por observación, se resumió cada una mediante estadísticas agregadas. En particular, para cada coeficiente se calcularon su media y su desviación estándar a lo largo de todos los frames del audio. Además, se incluyó un control de calidad eliminando valores no finitos (NaN o infinitos) y descartando los audios excesivamente cortos, con el fin de evitar que la extracción de características generase vectores inestables o no representativos.

Para incorporar toda esta información en el conjunto de datos, se utilizó la función `extract_feat_from_wav()`, la cual recibe un archivo de audio, realiza el preprocesado descrito, calcula MFCC, (Δ) y ($\Delta\Delta$), obtiene los estadísticos (medias y desviaciones), asigna la etiqueta correspondiente y devuelve un vector final de características. Así, cada fila del dataframe corresponde a un audio individual, con sus variables MFCC resumidas y su clase asociada (real o IA), además de metadatos como la ruta del archivo y la frecuencia de muestreo final.

Para automatizar la creación del dataframe completo, se implementó la función `build_feat_dataframe()`, que recorre todos los archivos *.wav* de las carpetas real y fake, aplica `extract_feat_from_wav()` a cada uno y une los resultados en un único dataframe.

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.

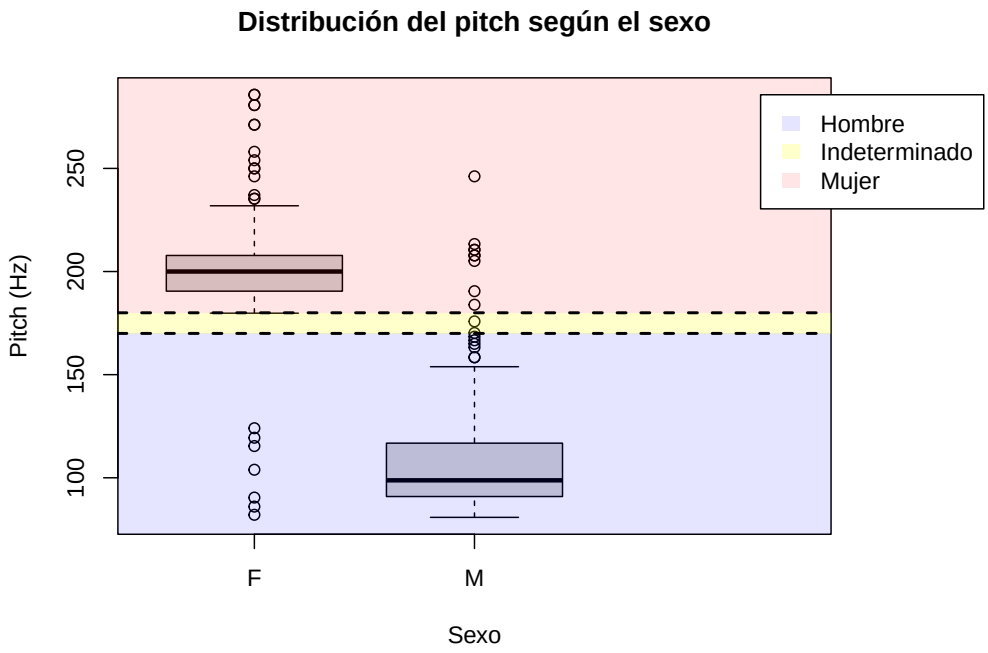

Figure 2: Distribución del pitch según el sexo.

El análisis de correlación muestra que la duración y la ZCR están fuertemente relacionadas, mientras que el pitch es prácticamente independiente del resto de variables, aportando información complementaria. De manera similar, el centroide espectral muestra baja correlación con las demás características, lo que indica que también contribuye con información adicional útil para la clasificación.

```
##              duracion  zcr energia_rms period_pitch centroide
## duracion         1.00 0.99        0.20         0.01      0.07
## zcr              0.99 1.00        0.21         0.04      0.16
## energia_rms      0.20 0.21        1.00         0.00      0.15
## period_pitch     0.01 0.04        0.00         1.00      0.43
## centroide        0.07 0.16        0.15         0.43      1.00
```

En cuanto al centroide espectral global, se observa una diferencia clara entre las voces humanas y las generadas por inteligencia artificial, tanto en el grupo masculino como en el femenino. En ambos casos, las voces generadas por IA presentan valores de centroide más elevados, lo que indica una mayor concentración de energía en frecuencias altas. Esta diferencia es especialmente marcada en las voces masculinas, donde el solapamiento entre clases es reducido. En el caso de las voces femeninas, aunque la tendencia se mantiene, se aprecia un mayor solapamiento entre ambas distribuciones.

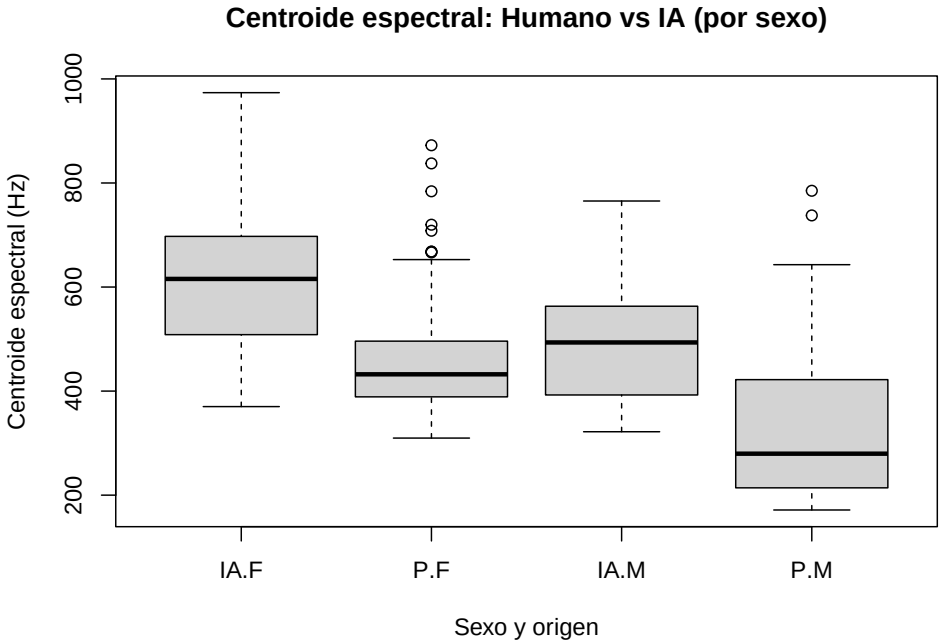

Figure 3: Centroide espectral: Humano vs IA por sexo.

Al analizar la desviación típica del centroide espectral obtenida a partir de la segmentación temporal, se esperaba observar una mayor capacidad de separación entre voces humanas y generadas por inteligencia artificial, bajo la hipótesis de que las voces humanas presentan una mayor variabilidad espectral a lo largo del tiempo. Sin embargo, los resultados muestran un solapamiento considerable entre ambas clases, tanto en voces masculinas como femeninas, lo que indica que esta medida agregada, considerada de forma aislada, no resulta suficiente para discriminar de manera fiable entre ambos tipos de voz. Por si solo, los parametros del centroide segmentado no dan mucha información, no obstante, después los analizaremos todos juntos para ver si entre todos ellos en conjunto pueden proporcionar clasificaciones más robustas.

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), centroide, 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 usando la función `summary()`, se observó que las únicas variables estadísticamente significativas para la clasificación del sexo eran el period pitch y el centroide, 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 y centroide frente al modelo completo mediante un ANOVA, con el fin de comprobar si la complejidad adicional estaba justificada.

El contraste mediante ANOVA entre el modelo reducido (period pitch y centroide) y el modelo completo muestra que la incorporación de estas variables produce una reducción significativa de la devianza. No obstante,vimos en el análisis individual de los coeficientes que el pitch y el centroide aportan información estadísticamente significativa de forma independiente, mientras que el resto de variables no contribuyen de manera relevante. Por ello, se optó por utilizar el modelo reducido para la clasificación del sexo, ya que combina descriptores complementarios, fáciles de calcular a partir de la señal de voz y con menor riesgo de sobreajuste, proporcionando una base robusta y generalizable.

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 0.727, lo que sugiere que aproximadamente el 72.7% de la variabilidad del conjunto de entrenamiento queda explicada por el modelo basado únicamente en el period pitch y el centroide. 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 0.9014, evidenciando una buena concordancia entre las predicciones y las etiquetas de sexo. Veámoslo gráficamente:

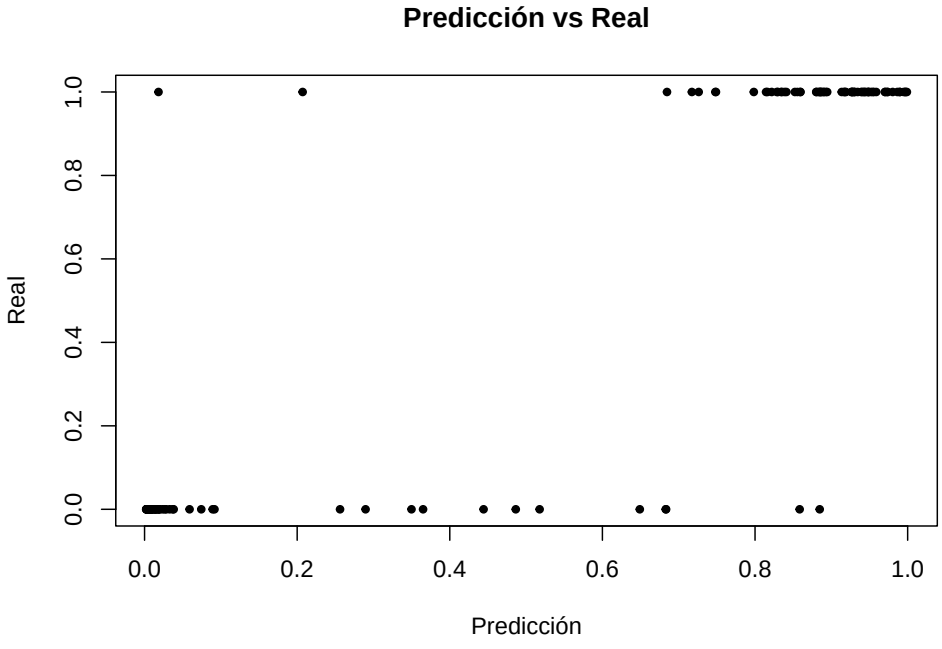

Figure 4: Predicción vs real del sexo.

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 origen (IA vs Humano): Coeficiente de Mel

Para tratar este problema de clasificación, se probaron diferentes datasets y dataframes, con los que se entrenó un modelo *SVM con kernel RBF*.

Se eligió entrenar un *SVM con kernel RBF* porque es un modelo especialmente adecuado para el tipo de datos que hemos usado.

En nuestro caso, cada audio no se mantiene como una secuencia temporal completa, sino que se transforma en un vector fijo de características: medias y desviaciones estándar de los MFCC y de sus derivadas Δ y $\Delta\Delta$. Esto convierte el problema en un dataset tabular (una fila por audio y un número moderado de variables, en torno a 72 predictores), similar a un problema clásico de clasificación con variables numéricas.

Con este tipo de variables, es razonable esperar que la separación entre audios reales y audios generados por IA no sea perfectamente lineal. Es decir, las dos clases pueden mezclarse en el espacio de características y necesitar una frontera de decisión más flexible que una simple recta o plano.

Por este motivo se utiliza un SVM con kernel RBF, ya que este modelo:

- Puede capturar relaciones no lineales entre las variables.
- Suele funcionar muy bien cuando hay un número medio de características (ni muy pocas ni miles)[3]
- Es un baseline clásico y robusto en tareas de voz y audio cuando se trabaja con MFCC y estadísticas agregadas.[4]

Proceso de entrenamiento Para entrenar el clasificador se definió un procedimiento de validación y ajuste de hiperparámetros que permite estimar el rendimiento de forma robusta y reducir el riesgo de sobreajuste.

En primer lugar, se estableció un esquema de validación mediante la función `trainControl()`. En este caso se utilizó validación cruzada repetida (repeated cross-validation), dividiendo el conjunto de entrenamiento en 5 particiones (5 folds), de forma que en cada iteración se entrena el modelo con 4 folds y se valida con el fold restante. Además, el proceso completo se repite repeats = 2 veces.

Una vez fijado el esquema de validación, se entrenó el modelo mediante la función `train()` (method="svmRadial") del paquete `caret`, obteniendo el modelo entrenado: `svm_fit`.

En total se entrenaron 2 modelos, con 2 datasets diferentes: un dataset extraido de internet (enlace al dataset: dataset for-norm) y otro dataset formado por nuestros audios y generados con IA.

Proceso de testeo Para evaluar el rendimiento del modelo una vez entrenado, en primer lugar, se generaron las predicciones del modelo sobre el conjunto de test mediante la función `predict()`. Por un lado, se solicitaron las probabilidades por clase, extrayendo concretamente la probabilidad asociada a la clase "IA" y por otro lado, se obtuvo también la predicción final de clase.

Además, se calcularon métricas de clasificación a partir de estas predicciones. En primer lugar, se utilizó `confusionMatrix()` para construir la matriz de confusión comparando las clases predichas y se evaluó el modelo desde un punto de vista probabilístico mediante la curva ROC, empleando la función `roc()` del paquete `pROC`.

Se realizó el testeo sobre datos del dataset extraido de internet y sobre datos extraidos del dataset creado por nosotros.

Centroide espectral Dado que el centroide espectral mostró diferencias entre voces humanas y generadas por IA, y aunque sus valores aislados no permiten una separación perfecta, resulta interesante evaluar su capacidad discriminativa en conjunto para clasificar el origen de la voz. Para ello, se entrenaron dos modelos de clasificación basados en random forest con el objetivo de evaluar la capacidad discriminativa de las características derivadas del centroide espectral. El primer modelo utiliza exclusivamente el centroide espectral global, mientras que el segundo incorpora los centroides espectrales obtenidos mediante segmentación temporal, junto con su desviación típica. En ambos casos, el análisis se realizó de forma independiente para voces masculinas y femeninas, por lo que se obtuvieron matrices de confusión separadas para cada grupo.

Las matrices de confusión del modelo con el centroide global son las siguientes:

Table 1: Matrices de confusión del modelo basado en el centroide espectral global, separadas por sexo

	Pred_M_IA	Pred_M_P	Pred_F_IA	Pred_F_P
Real_M_IA	11	5	18	7
Real_M_P	16	36	9	38

El modelo de random forest entrenado únicamente con el centroide espectral global presenta una capacidad moderada de clasificación entre voces humanas y generadas por IA. En concreto, se obtiene una precisión del 69,1 % en las voces masculinas y del 77,8 % en las voces femeninas, lo que indica un rendimiento desigual según el sexo. Aunque el centroide global muestra una separación clara en el análisis exploratorio, los resultados del modelo ponen de manifiesto la existencia de errores de clasificación relevantes, especialmente en la identificación de voces humanas etiquetadas como IA. Esto sugiere que, si bien el centroide espectral global captura diferencias estructurales en la distribución espectral de las señales, su uso como única variable explicativa resulta insuficiente para una clasificación robusta, particularmente en presencia de solapamiento entre clases.

Table 2: Matrices de confusión (12 centroides + sd12), separadas por sexo

	Pred_M_IA	Pred_M_P	Pred_F_IA	Pred_F_P
Real_M_IA	12	4	15	10
Real_M_P	11	41	3	44

Al incorporar la información temporal del centroide espectral mediante la segmentación en doce partes, junto con su desviación típica, se observa una mejora consistente en el rendimiento del modelo de clasificación. En particular, la precisión aumenta hasta 77,9 % en las voces masculinas y 81,9 % en las voces femeninas, lo que confirma una ganancia respecto al modelo basado únicamente en el centroide global. A pesar de que ni los centroides segmentados ni la desviación típica presentan una separación clara cuando se analizan de forma individual, su uso conjunto permite al modelo capturar patrones más complejos en la evolución temporal del espectro. Esto se traduce en un aumento de la precisión global, especialmente en la correcta identificación de voces humanas, reduciendo el número de falsos positivos asociados a la clase IA.

En conjunto, los resultados muestran que, aunque el centroide espectral global parece diferenciar de manera más clara entre voces humanas y generadas por IA en un análisis univariado, la utilización conjunta de los centroides espectrales segmentados y su desviación típica ofrece un mejor rendimiento en el contexto de modelos de clasificación multivariantes. Este comportamiento pone de manifiesto que la información relevante no reside únicamente en un descriptor agregado, sino en la combinación de múltiples características que capturan tanto la estructura global como la variabilidad temporal del espectro. Por tanto, el centroide espectral resulta especialmente útil cuando se integra como parte de un conjunto de características más amplio, más que como un descriptor aislado.

Por último, al comparar estos resultados con los obtenidos mediante otros descriptores espectrales más complejos, como los coeficientes cepstrales en escala Mel (MFCC), se observa que estos últimos ya incorporan de forma más eficiente información espectral y temporal relevante para la clasificación. En consecuencia, la inclusión del centroide espectral en modelos que ya utilizan MFCC no aporta una mejora sustancial adicional.

Resultados Se evaluó el rendimiento del sistema bajo tres configuraciones experimentales, combinando distintos conjuntos de entrenamiento y test con el objetivo de analizar tanto el desempeño en condiciones controladas como su capacidad de generalización:

1. Entrenamiento con el dataset de Internet y test con el dataset de Internet, para medir el rendimiento del modelo en el mismo dominio de datos con el que ha sido entrenado.
2. Entrenamiento con el dataset de Internet y test con nuestro dataset, para evaluar la capacidad de generalización del modelo al aplicarlo sobre audios obtenidos en un entorno distinto (cambio de dominio).
3. Entrenamiento con nuestro dataset y test con nuestro dataset, para estimar el rendimiento del modelo cuando se entrena y evalúa específicamente en el contexto y características de nuestro conjunto de datos.

Caso 1: Entrenado con dataset de Internet y testeado con dataset de Internet

Predicción\Real	**real**	**ia**
real	2063	106
ia	201	2264

En esta configuración el modelo se evalúa en el mismo dominio en el que fue entrenado. La matriz de confusión muestra un número bajo de errores (106 falsos "real" y 201 falsos "IA"), y se obtiene una `accuracy` ≈ 0.934, por lo que esta claro que no es azar.

Además, la curva ROC aparece muy próxima al vértice superior (comportamiento de AUC elevado), lo cual confirma que el modelo separa bien ambas clases cuando el test comparte distribución con el entrenamiento.

Estos resultados son buenos y razonables, y se explican por dos factores principales:

- El dataset de Internet es grande y suele incluir variabilidad suficiente para que el SVM aprenda una frontera no lineal estable en el espacio MFCC (medias/desviaciones + Δ y $\Delta\Delta$).
- No existe cambio de dominio ya que las condiciones de grabación, tipos de voces de IA y características del preprocesado son similares entre train y test.

Caso 2: Entrenado con dataset de Internet y testeado con nuestro dataset

Predicción\Real	**real**	**ia**
real	177	3
ia	26	105

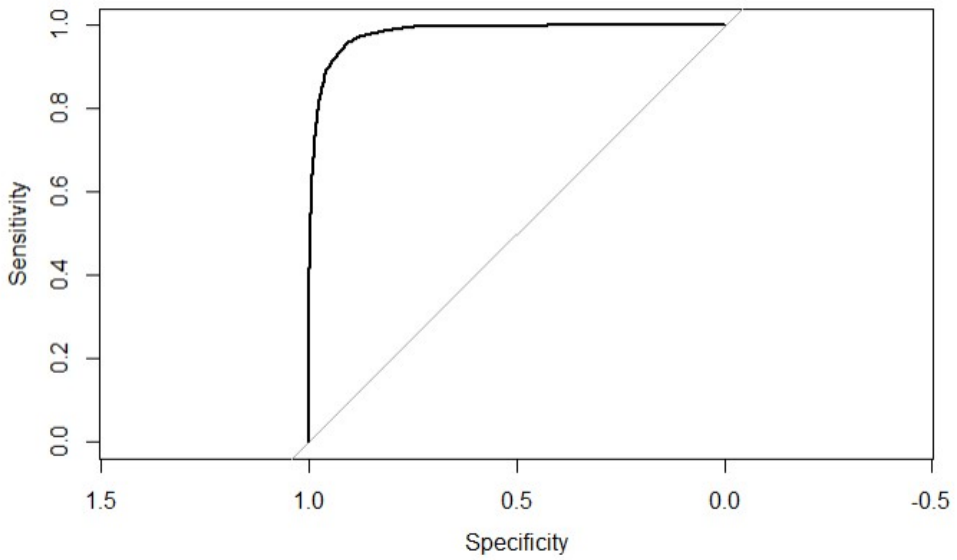

Figure 5: Curva ROC del modelo.

Aquí se mide la generalización a un dominio distinto puesto a que son audios grabados por nosotros y audios IA generados también por nosotros. Aun así, el rendimiento sigue siendo alto: `accuracy` ≈ 0.907. La curva ROC sigue mostrando una separación clara.

En la matriz de confusión se observa un patrón típico de cambio de dominio:

- Apenas hay falsos "real" cuando el audio es IA (solo 3 casos), es decir, el modelo detecta muy bien IA en este test.
- El principal error es etiquetar como IA algunos audios reales (26 casos), lo que sugiere que ciertas características de nuestros audios reales (grabación, limpieza de silencios, energía residual, etc.) pueden parecerse a rasgos presentes en parte del material sintético del dataset de Internet.

Estos resultados son lógicos: al cambiar de dominio suele aparecer una degradación moderada porque cambian distribuciones (calidad del micrófono, ruido de fondo, compresión, etc.). Aun así, que el rendimiento se mantenga por encima del 90% sugiere que los MFCC agregados y sus deltas están capturando rasgos relativamente transferibles entre datasets.

Caso 3: Entrenado con nuestro dataset y testeado con nuestro dataset

Predicción\Real	**real**	**ia**
real	153	0
ia	0	108

En esta combinación se obtiene accuracy = 1.0 (sin ningún error) y una ROC perfecta. Este resultado es poco habitual para un escenario realista de detección IA vs humano, y normalmente indica que la clasificación, ha quedado "demasiado fácil" para el modelo.

Esto puede deberse a que el diseño del dataset introduce una estructura muy repetitiva: mismas frases, solo una variedad de 6 personas (por lo que se repiten mucho las mismas voces) y misma cadena de grabación y limpieza, además de realizar el mismo procedimiento de generación IA usando también las mismas frases.

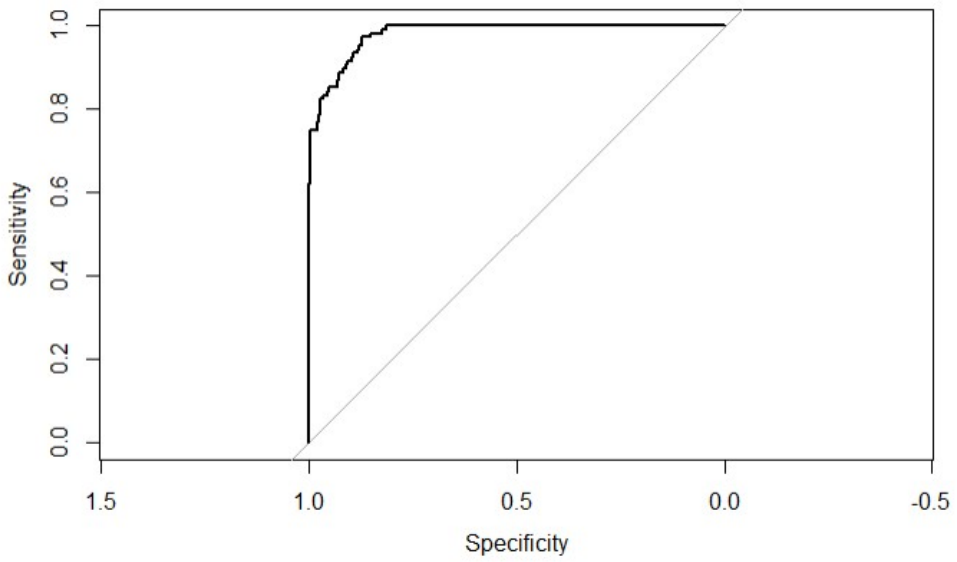

Figure 6: Curva ROC del modelo.

Esto puede provocar que el modelo no esté aprendiendo "humanidad vs IA" en un sentido general, sino artefactos muy específicos de nuestro contexto.

En conclusión, los resultados obtenidos no son una evidencia sólida de que el modelo generalice a "cualquier" voz humana vs "cualquier" IA, porque el test probablemente comparte demasiada estructura con el entrenamiento.

Clasificación del acento

Además de la clasificación por sexo, se intentó abordar la identificación del acento de la voz, considerando tres categorías: castellano neutro, andaluz y argentino.

Comenzamos haciendo un estudio exploratorio de las características acústicas para comprobar si alguna podía, por sí sola, ayudar a distinguir entre acentos. Nos centramos en la tasa de cruces por cero (ZCR) ya que está muy relacionada con el contenido de altas frecuencias y con la presencia de consonantes fricativas o sonidos sordos, que pueden variar entre acentos, ya sea por mayor aspiración, articulación más "suave" o más "marcada", etc.

Se representó la distribución del ZCR por acento mediante un diagrama de cajas, del cual se observa que:

Las tres categorías (AN: andaluz, AR: argentino, N: neutro) presentan rangos de ZCR parcialmente solapados, lo que indica que no hay una separación clara solo con esta característica. Sin embargo, el acento neutro (N) muestra una mediana de ZCR ligeramente más alta y una dispersión mayor, mientras que AN y AR parecen algo más compactos y con medianas algo inferiores.

Existen algunos outliers en AR y N (valores de ZCR más altos), que podrían corresponder a audios concretos con más ruido, más fricación o peor calidad de grabación.

En conjunto, el boxplot indica que el ZCR contiene cierta información sobre el acento (las distribuciones no son idénticas), pero la superposición tan grande entre cajas apunta a que no es suficiente por sí solo para discriminar con fiabilidad los acentos. Esto motivó a combinar el ZCR con otras características en un modelo multinomial.

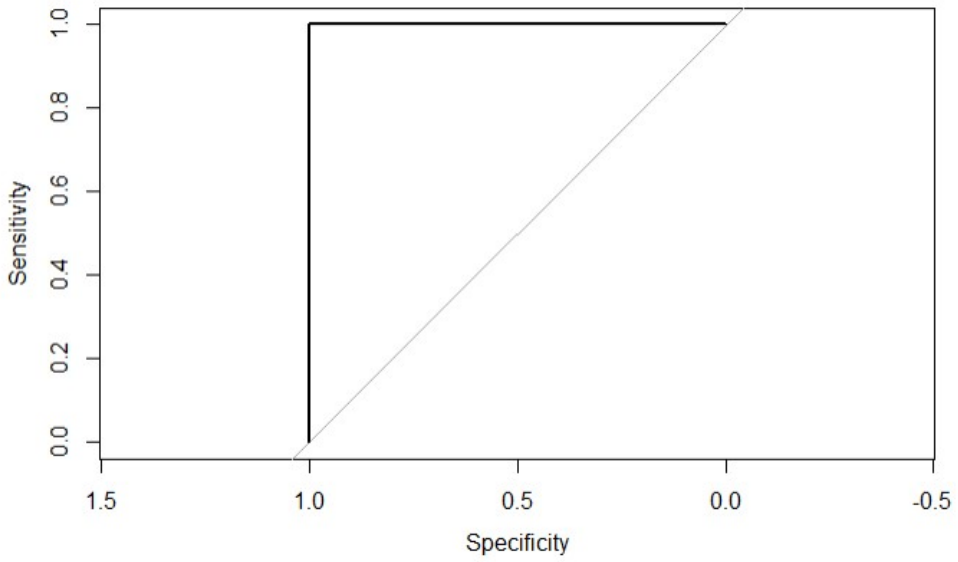

Figure 7: Curva ROC del modelo.

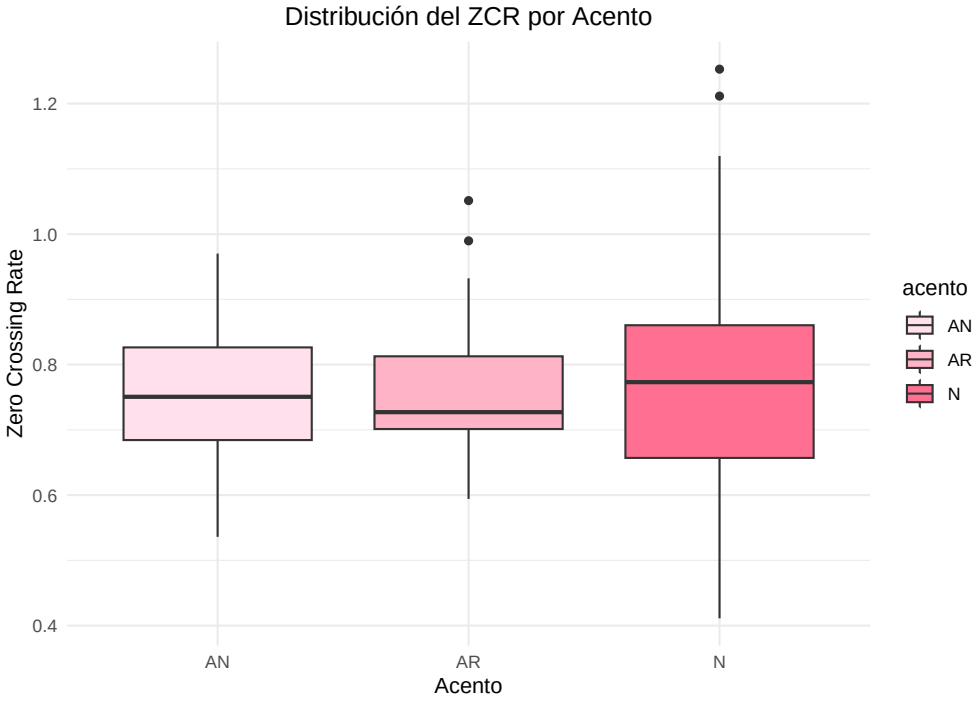

Figure 8: Distribución del ZCR por Acento

Dado lo visto y teniendo en cuenta que se trata de una clasificación multiclase, se planteó un modelo de regresión logística multinomial utilizando como predictores varios descriptores acústicos de bajo nivel: ZCR, la energía RMS, la duración del audio y el pitch period. Estas características se eligieron por su relación con aspectos articulatorios y prosódicos que suelen diferir entre acentos:

La energía RMS resume la intensidad media de la señal y puede reflejar diferencias en el patrón de acentuación

o en la "fuerza" con la que se articula. Por otro lado, la duración del audio captura diferencias en el tempo o velocidad de habla, ya que algunos acentos tienden a ser más rápidos o más pausados. Además, el period pitch recoge variaciones en la entonación y altura de la voz, otro rasgo típicamente distintivo entre acentos.

En conjunto, estas variables, permiten modelar tanto aspectos segmentales (contenido en altas frecuencias, presencia de fricativas, etc.) como suprasegmentales (ritmo, intensidad y entonación), que son precisamente los elementos en los que suelen diferir los acentos.

Para evaluar la capacidad explicativa del modelo se utilizó el pseudo-R^2 de McFadden. En regresión logística el R^2 clásico no es adecuado porque no existe una descomposición de la varianza análoga a la de la regresión lineal, y por ello se recurre a índices basados en la verosimilitud. El pseudo-R^2 de McFadden es uno de los más utilizados porque compara de forma directa el modelo con predictores frente a un modelo nulo, que solo incluye el intercepto. Valores más altos de R^2_{McFadden} indican que el modelo con predictores mejora de forma apreciable la explicación de los datos frente al modelo nulo, mientras que valores cercanos a cero sugieren que las variables empleadas apenas añaden capacidad discriminativa entre acentos.

Tras analizar los resultados, vimos que, aunque el modelo consigue ajustar los datos ($R^2_{\text{McFadden}} = 0.5296$), los resultados no pueden considerarse fiables. El número de muestras por acento no es equilibrado: el grupo cuenta con tres personas con acento castellano neutro (dos hombres y una mujer), dos con acento andaluz (hombre y mujer) y una sola con acento argentino (mujer). La variabilidad interna de cada clase es muy limitada al usar varios audios de las mismas voces. Esto hace que, en la práctica, el modelo tiende a aprender rasgos de cada persona más que del acento en sí.

Por lo tanto, este análisis no resulta concluyente y se incluye unicamente en el informe como demostración del estudio realizado.

References

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[2] A. Basu, P. Ahuja, and D. Dahiya, "Estimation of pitch and fundamental frequency variation between normal males, females and intersex population," *Indian Journal of Forensic Medicine & Toxicology*, vol. 14, pp. 1042–1048, 2020.

[3] MathWorks. (n.d.) fitcsvm — train support vector machine (svm) classifier for one-class and binary classification. [Online]. Available: https://es.mathworks.com/help/stats/fitcsvm.html

[4] A. Temko and C. Nadeu, "Classification of acoustic events using svm-based clustering schemes," *Pattern Recognition*, vol. 39, pp. 682–694, 2006.