Proyecto de Clasificación de Voces de Personas

Luis Albacete Caballero[a], Julio García Bustos[a], Gabriel Ivars Asensio[a], Noé López García[a], José Miguel Palazón Caballero[a], Joan Pedro Bruixola[a]

[a] *Universitat de València, Avinguda de l'Universitat, Burjassot, 46100, Valencia, España*

Abstract

La clasificación de voces de personas es un tema latente en el análisis de señales. Requiriendo extraer particularidades y características de la voz como señal. A lo largo de este trabajo obtendremos distintas características utilizando librerias o implementando nosotros la propia extracción de la característica. Bajo la finalidad de poder clasificar audios dependiendo del género del locutor.

Keywords: Voces, Características, Características de la voz, Clasificación, Aprendizaje Máquina, Análisis de Señales, Identificación de voces

1. Captación de audio

A fin de poder extraer las características de los audios, requerimos en primera instancia de los audios en sí. Para ello hemos captado 70 audios, grabados por 10 compañeros del Máster de Ciencia de Datos.

Para agilizar las labores de preparación del contenido, los audios fueron grabados bajo las mismas condiciones, desde software y hardware de captación de audio como la sala de grabación e instrucciones para la grabación.

Aprovechar para agredecer a todos los colaboradores su intervención.

2. Características de la voz

En el desarrollo de este apartado probaremos distintas técnicas de extracción de características.

2.1. Mel Frecuency Cepstral Coefficients (MFCC)

Esta característica es una de las más comunes para reconocimiento de voces. Combina el análisis de cepstrums con una escala perceptual de las frecuencias.

*Corresponding author

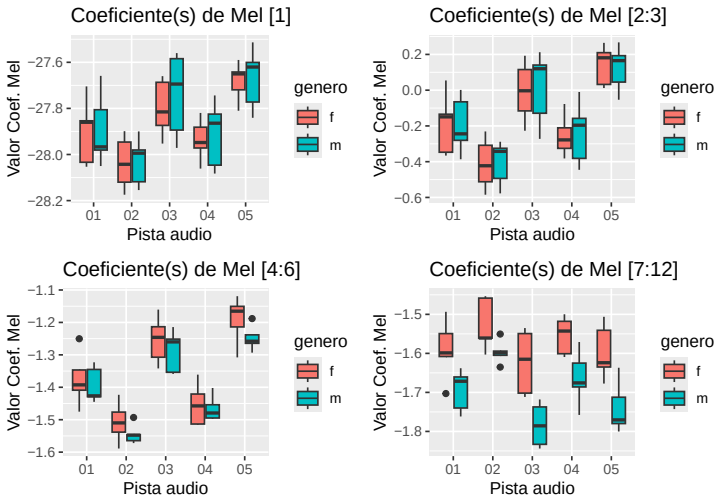
Email addresses: alcaluis@alumni.uv.es (Luis Albacete Caballero), jugarbus@alumni.uv.es (Julio García Bustos), gaia2@alumni.uv.es (Gabriel Ivars Asensio), nologar@alumni.uv.es (Noé López García), jomipaca@alumni.uv.es (José Miguel Palazón Caballero), jopebrui@alumni.uv.es (Joan Pedro Bruixola)

Los coeficientes de Mel se basan en la percepción humana utilizando una escala linealmente separada para los valores de la señal inferior a 1KHz y una logarítmica en caso de ser superior.

El cálculo de los coeficientes se constituye de los siguientes pasos:

1. **Pre-emphasis**. El primer paso consiste en pasar la señal por un filtro paso alto. En este se incrementará la energía de las altas frecuencias. Utilizando la siguiente fórmula: $y(n) = x(n) - a * x(n-1)$ con valores de a entre 0.9 y 1.
2. **Frame blocking**. Consiste en la división de la señal en frames de entre 20 y 30 ms.
3. **Windowing**. Seguidamente se aplica un efecto de "windowing" de manera que los bordes de la señal (frame) sean suavizados (más cercano a 0 en los extremos). La función de ventana que se utiliza es la de Hamming.
4. **Discrete Fourier Transform**. Como deseamos recoger las energías en el dominio de las frecuencias debemos hacer un cambio de dominio mediante la DFT.
5. **Mel-Filter**. Se suma las energías de las componentes espectrales por cada escala de Mel. Al resultado del filtrado se escalará logarítmicamente.
6. **Discrete Cousine Transform**. Finalmente se aplicará la siguiente transformada: $C(n) = \sum Ek * cos(n * (k - 0.5) * \pi/40)$ para volver al dominio del tiempo. Obteniendo los coeficientes de Mel que recogen la energía total escalada por cada banda de frecuencia por cada intervalo.

Agruparemos los coeficientes para un análisis más conciso basandonos en el concepto del oído humano. Escalando poco a poco las agrupaciones. En la siguiente gráfica veremos los coeficientes entre hombres y mujeres para las 5 primeras pistas de audio.



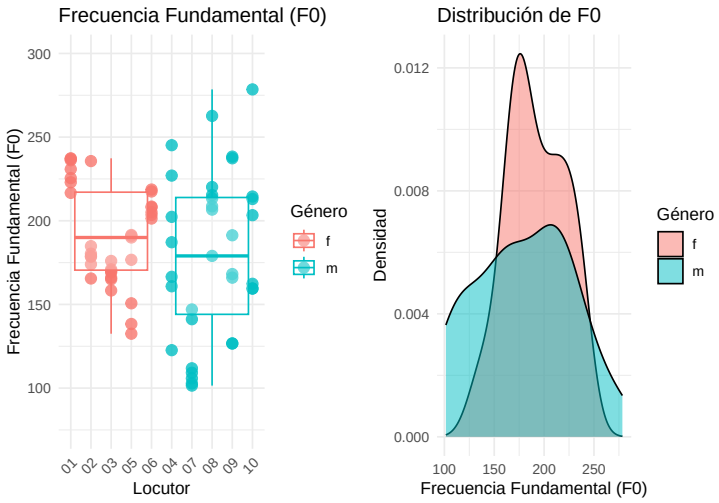
Empezando por el primer coeficiente podemos observar que el género masculino tiene los valores más altos y distribuidos. Para los coeficientes que van del

segundo al sexto, podemos ver que los valores femeninos van ganando relevancia según aumentamos el coeficiente, es decir, según captamos mayores frecuencias. Finalmente para los últimos coeficientes vemos los valores para los hombres son claramente inferiores y menos distribuidos.

2.2. Frecuencia fundamental

La frecuencia fundamental es una característica ampliamente utilizada en el análisis de secuencias de audio, especialmente para diferenciar géneros y hablantes. Esto se debe a que representa el tono base de la voz, el cual tiende a ser más bajo en hombres que en mujeres. Su variabilidad entre individuos también lo convierte en un indicador relevante para la identificación de locutores. Por su relación directa con la fisiología vocal, esta métrica es clave en estudios relacionados con la prosodia y la fonética.

Su cálculo se basa en dividir la señal en ventanas (de 2 segundos en este caso) y realizar la Transformada de Fourier en cada una. La frecuencia fundamental en cada ventana se determina como la frecuencia con la máxima amplitud en el espectro resultante, considerando solo la mitad positiva del espectro. Para obtener un único valor representativo de toda la señal, se calcula la media ponderada de estas frecuencias, asignando un peso mayor a las primeras ventanas y eliminando la contribución de los últimos 4 segundos, ya que suelen ser menos relevantes debido a que el locutor ya ha terminado de hablar en la mayoría de los casos.



Nuestro caso se muestra paradójico con respecto a esta característica, ya que, aunque las mujeres presentan frecuencias fundamentales altas y similares, los hombres se distribuyen en un rango más amplio, desde frecuencias bajas hasta las más altas. Este comportamiento no es el esperable teóricamente, pues se ha demostrado que las mujeres tienen una frecuencia fundamental superior a la de los hombres, lo que en nuestro caso no ocurre.

Esta discrepancia podría explicarse por factores propios de los datos, como diferencias en las condiciones de grabación o la inclusión de locutores con características atípicas, como hombres con frecuencias fundamentales altas o mujeres con frecuencias bajas.

2.3. Formantes

Los formantes se encuentran caracterizados por aquellas intensidades en el espectro de sonido que destacan en un señal. Altamente relacionado con el tracto vocal de la persona, produciendo una gran cantidad de formantes. Siendo los más relevantes los primeros. En concreto los dos primeros relacionados con el sonido de las vocales. A partir del segundo formante, las frecuencias le dan color a nuestra voz, hasta el punto de tener frecuencias no distinguibles para el oído humano.

Con el objetivo de una sencilla visualización, a continuación obtendremos los formantes a partir de los sonidos de las vocales "a" y "e". Tras haber recortado algunos de los audios grabados con el objetivo de investigar la característica.

De los formantes obtenidos hemos realizado la media por género. Por lo general, esperamos valores más altos para las mujeres. Cabe destacar que al haber recortado manualmente los audios, los fonemas no son puros. Además el tamaño de nuestra muestra es bastante reducido. Es por eso que no esperamos obtener resultados perfectos, en caso de querer comparar con otros estudios, pero sí pudiendo diferenciar.

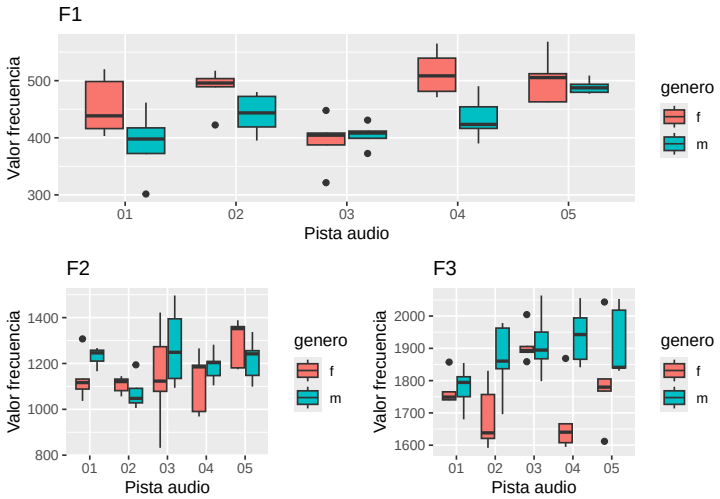
Table 1: Formantes en las vocales A y E

Genero	F1 (vocal A)	F2 (vocal A)	F1 (vocal E)	F2 (vocal E)
Femenino	629.554	1767.900	476.29	2228.64
Masculino	627.156	1571.176	384.37	2065.97

A partir de los datos obtenidos, podemos diferenciar claramente el género del locutor. Para el caso de la vocal E, los formantes 1 y 2, para las mujeres son más altos. Sin embargo, si luego observamos en el caso de la vocal A, para el formante 1, tenemos valores muy similares, es solo en el formante 2 donde podemos diferenciar. Teniendo otra vez valores inferiores en los hombres.

Debido al coste computacional implicado en el cálculo de formantes, hemos decidido precalcularlo utilizando la aplicación PRAAT. De cada audio extraemos una palabra con ciertos fonemas caraterísticos. Teniendo en cuenta la complejidad que buscamos con nuestro clasificador, bajo el objetivo de clasificar el género del locutor, resumiremos los datos. En vez de utilizar los formantes en su total extensión en el tiempo calcularemos estadísticos, como la media y la desviación estándar para los 3 primeros formantes.

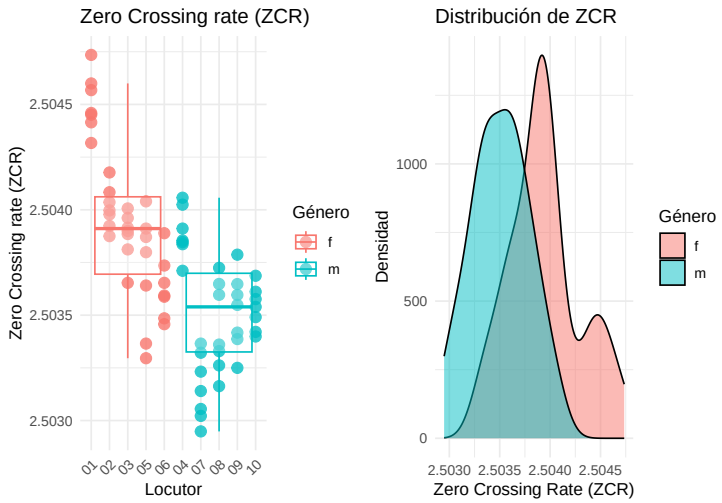
Si analizamos los valores de los formantes visualizamos en parte lo esperado. Para el primer formante tenemos frecuencias más altas en las mujeres. Según pasamos a los siguientes a los siguientes formantes notamos una evolución donde las frecuencias en los hombres van tomando más protagonismo.



2.4. Zero Crossing Rate

La tasa de cruces por cero (ZCR, por sus siglas en inglés) mide cuántas veces la señal cruza el eje cero en un periodo de tiempo determinado. Es mayor en voces femeninas debido a su mayor frecuencia fundamental, lo que evidencia una estrecha relación entre ambas características. Este parámetro se utiliza frecuentemente en el análisis de señales para distinguir patrones relacionados con la tonalidad y la estructura armónica de la voz.

Para su cálculo, se emplea la función zcr en R, utilizando ventanas con un solapamiento del 50%. Posteriormente, se realiza el promedio de los valores obtenidos en cada ventana para obtener un único valor representativo por señal, similar al enfoque utilizado para calcular la frecuencia fundamental. Esta media permite consolidar la información de toda la señal en un solo parámetro útil para el análisis.



En esta característica se observa un comportamiento más acorde con lo esperado teóricamente. El Zero Crossing Rate (ZCR) tiende a ser mayor en las mujeres que en los hombres, lo cual es consistente con las propiedades de las señales de voz. Este patrón se aprecia tanto en la distribución general del ZCR por género como en los locutores individuales, donde, por norma general, las mujeres presentan valores de ZCR superiores. Este resultado refleja que el ZCR es una característica adecuada para diferenciar géneros en nuestro caso de estudio.

2.5. Linear Predictive Coding

El Linear Predictive Coding (LPC) es un modelo matemático que describe una señal de audio en función de una combinación lineal de sus valores pasados. En lugar de representar la señal de audio directamente, se modela como una secuencia de coeficientes que predicen el valor futuro de la señal basándose en los valores previos. Los coeficientes LPC son los parámetros que caracterizan este modelo de predicción lineal. En el contexto de la LPCC, los coeficientes son derivados a partir de los coeficientes LPC mediante una transformación matemática, y son usados frecuentemente como una representación compacta y eficiente de la información espectral de la señal. Nuestro objetivo es analizar si los coeficientes obtenidos para los audios con locutores hombres tienen patrones distintos a los obtenidos para las locutoras.

Para obtener los coeficientes LPCC seguiremos los siguientes pasos:

1. Preprocesamiento de la señal: La señal de audio se segmenta en ventanas de corto tiempo (frames) para su análisis.
2. Cálculo de los coeficientes LPC: A partir de cada segmento de la señal, se calcula un conjunto de coeficientes LPC (usaremos la función lpc). Estos coeficientes describen cómo la señal futura puede ser predicha a partir de los valores anteriores.

3. Transformación a LPCC: Una vez obtenidos los coeficientes LPC, se aplican transformaciones matemáticas (como la transformada cepstral) para obtener los coeficientes LPCC.

Usaremos orden 4, lo que significa que obtendremos 4 coeficientes por ventana. Por cada audio calcularemos la media y desviación estándar de los coeficientes.

A continuación, analizaremos si es posible identificar patrones en los coeficientes que permitan discernir el género del locutor.

Vamos a filtrar por la pista de audio y veremos si se aprecian diferencias entre la media y la desviación estándar en función del género del locutor. Hemos ignorado el valor del primer coeficiente pues siempre es 1.

Table 2: Media y desviación del LPCC para la pista 2

genero	C2_M	C2_SD	C3_M	C3_SD	C4_M	C4_SD
f	-4.466702	0.0349284	10.20331	0.139762	-21.89433	0.3816098
m	-4.451493	0.0519720	10.14217	0.207878	-21.72749	0.5664334

Table 3: Media y desviación del LPCC para la pista 3

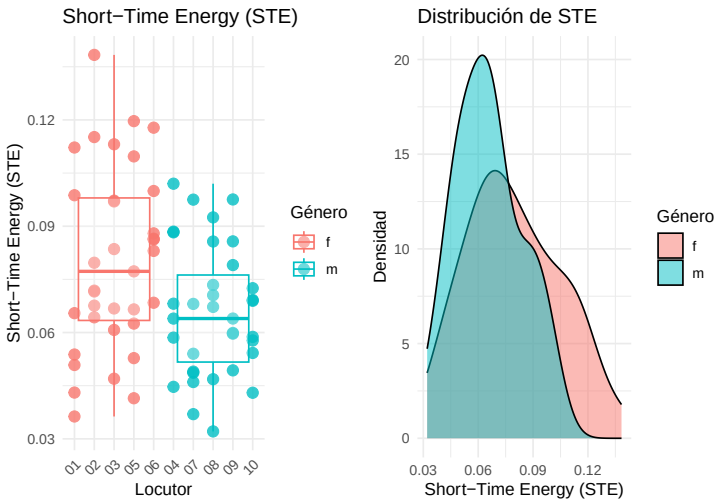
genero	C2_M	C2_SD	C3_M	C3_SD	C4_M	C4_SD
f	-4.467829	0.0350627	10.20784	0.1403122	-21.90678	0.3830371
m	-4.453387	0.0521274	10.14982	0.2085455	-21.74846	0.5680478

Notamos que ciertos patrones se repiten para todas las pistas de audio. La media de los coeficientes 2 y 4 correspondientes a los chicos es mayor que la de las chicas en todas las pistas. Mientras que con el coeficiente 3 pasa al contrario. También podemos destacar que la desviación estándar de los coeficientes es en todos los casos superior para los locutores chicos que para las chicas.

2.6. Short-Time Energy

La característica Short-Time Energy (STE) mide la energía contenida en una señal dentro de pequeñas ventanas temporales. Puede ser muy útil para analizar variaciones de energía a lo largo del tiempo. Para nuestro caso, puede ser particularmente interesante porque las variaciones de energía a lo largo del tiempo pueden reflejar características importantes de la voz humana, como la intensidad, los patrones de habla y las pausas.

El procedimiento que hemos seguido consiste en tomar cada señal (audio) y dividirla en ventanas temporales. Luego, calculamos la energía cuadrática media para cada ventana y la normalizamos. Finalmente, promediamos los valores de STE normalizados obtenidos en cada audio para obtener un único valor representativo.



Observamos que las locutoras tienen una mayor variabilidad en la STE en comparación con los locutores hombres. Esto podría deberse a características propias de las voces femeninas, como el rango de modulación, que es más amplio. Además, las locutoras presentan valores de STE más altos en general. Esto se refleja en una mediana de aproximadamente 0.10, mientras que en los locutores hombres la mediana se encuentra alrededor de 0.07. Esto puede tener relación con la frecuencia fundamental. Las voces femeninas suelen tener una frecuencia fundamental más alta que las voces masculinas. Dado que la STE mide la energía de la señal acústica, una frecuencia fundamental más alta puede generar más energía en las frecuencias superiores, lo que puede reflejarse en un valor más alto de STE.

Las curvas de densidad muestran que las voces femeninas tienen una distribución más amplia y desplazada hacia valores más altos de STE, lo que confirma lo que hemos visto en el boxplot. Los locutores hombres, por otro lado, presentan una distribución más concentrada y con menor dispersión. Aunque hay diferencias entre géneros, también se observa cierto solapamiento en las densidades de STE. Esto indica que, si bien hay tendencias generales, la STE por sí sola puede no ser completamente discriminativa entre géneros en todos los casos.

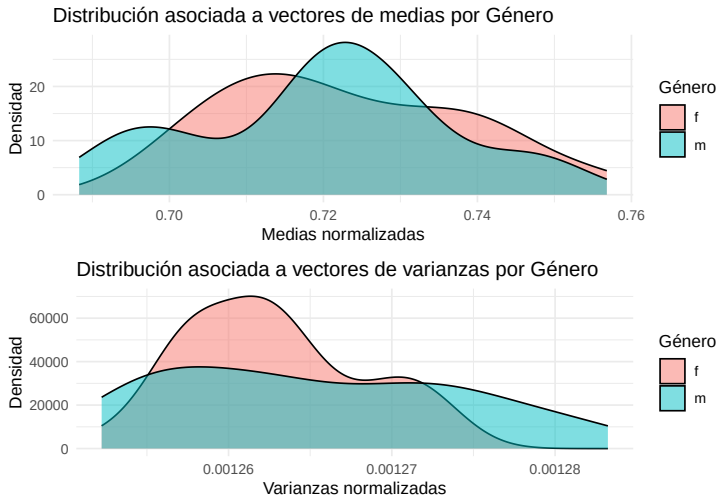
En conclusión, los resultados sugieren que la Short-Time Energy es una métrica que capta diferencias significativas entre géneros, aunque también está influenciada por características individuales.

2.7. Piecewise Gaussian Modeling

El PGM es un método que utiliza una estructura a largo plazo para representar las características, llamadas Ventanas de Tiempo de Integración (ITW). En cada ITW, se tiene un vector de medias y un vector de varianzas.

Para la implementación de este método, primero calculamos el MFSC (Mel Frequency Spectral Coefficient). Es similar al MFCC obtenido anteriormente pero sin la parte de DCT (Discrete Cosine Transform). Una vez calculado, se

obtiene una matriz $N \times T$ donde T se refiere al número de vectores espectrales contenidos en una ITW (Ventana de Tiempo de Integración) y N al número de ventanas ITW. A continuación se modela un conjunto de T vectores MFSC consecutivos mediante un modelo gaussiano. Es decir, $N \times T$ de MFSC serán modelados por N gaussianas. Al calcular la matriz de covarianza, solo tomamos el índice diagonal de la matriz de covarianza para generar el vector de varianza. Finalmente, las características del PGM son la concatenación de la media y la varianza normalizadas por sus respectivos máximos y mínimos. Para obtener una característica de media y otra de varianza por audio, hemos calculado la media de cada vector.



Las curvas de densidad muestran que para la característica de las medias normalizada, las voces masculinas tienen una menor dispersión que las femeninas y un mayor valor en promedio, pero parecen mezclarse entre sí en un mismo rango de valores. A priori no parece una buena característica para discriminar

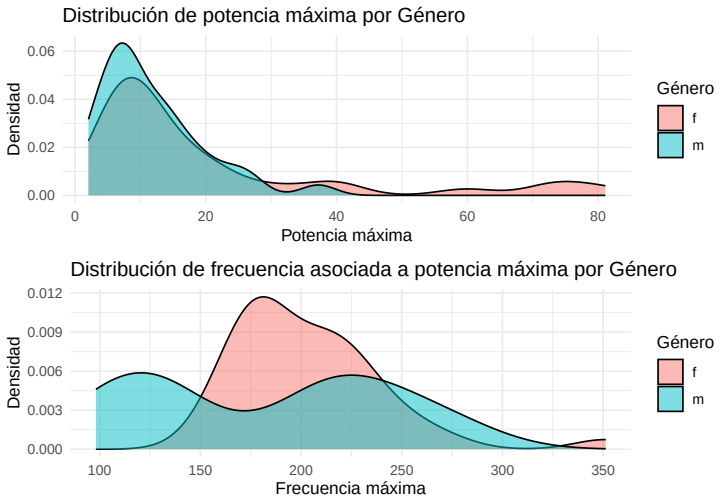
La característica asociada a las varianzas normalizadas podría ser un mejor discriminante tampoco parece una buena característica para discrimanr. La dispersión de las voces masculinas es muy grande, asemejandose mucho a una distribución uniforme continua (prácticamente no aportan información). En cambio, la dispersión de las voces femeninas es menor y los valores se concentran mas en torno a la media.

2.8. Power Spectrum

Esta característica nos indica qué tan fuertes están presentes diferentes frecuencias en una señal. Para estimar el Power Spectrum, se utiliza un estimador llamado periodograma. Los pasos a seguir son los siguientes:

- **Transformada Rápida de Fourier (FFT):** Se aplica la transformada para cambiar al dominio de las frecuencias.

- **Cálculo del periodograma:** Se calcula el módulo al cuadrado de la transformada de Fourier en cada frecuencia. El resultado es el periodograma, que representa la distribución de la potencia de la señal en función de la frecuencia.
- **Frecuencia de máxima potencia en el espectro:** Se identifica la frecuencia en la que se encuentra el valor máximo del espectro de potencia. Este valor de potencia máxima, junto con su frecuencia asociada, se utilizan como características clave para la clasificación de cada audio.



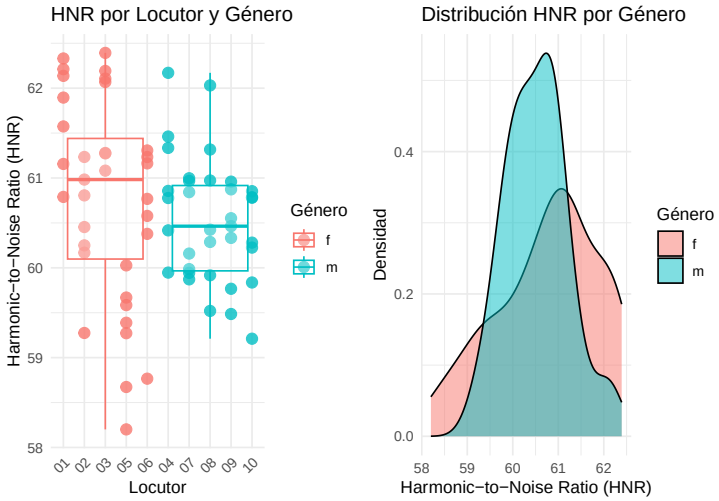
Se puede observar una mayor concentración de valores más bajos de potencia máxima en las voces masculinas mientras que en las voces femeninas se muestra una distribución más dispersa, con algunos valores más altos de potencia máxima.

La distribución de la frecuencia de máxima potencia en el espectro, a priori, parece un poco mas determinante para la tarea de clasificación. En las voces femeninas hay una mayor densidad en frecuencias medias/altas (alrededor de 190 Hz) mientras que en las voces masculinas pueden diferenciarse dos grupos de frecuencias (bajas y altas), por lo que no parecen estar tan concentradas alrededor de una única frecuencia media.

2.9. Harmonic-to-Noise Ratio (HNR)

Esta característica mide la proporción entre la energía de la componente armónica de la señal respecto al ruido presente. Aunque su principal uso es en el área de la sanidad para detectar patologías mediante la voz, también nos puede ayudar a distinguir la claridad y ciertas particularidades de las voces.

Para ello, nos ayudaremos de la descomposición mediante wavelets para separar las Aproximaciones (componente de baja frecuencia de la señal normalmente interpretada como la parte harmónica) de los Detalles (componente de alta frecuencia relacionada con el ruido).



El gráfico de distribución de densidades muestra una clara diferencia de tendencia entre géneros; mientras que las voces masculinas suelen encontrarse más cerca de la media cercana a 62, el HNR de las voces femeninas está distribuido de forma más uniforme en el rango de valores. De esta forma, valores de HNR cercanos a los extremos podrían indicar mayor probabilidad de que la voz grabada sea femenina y valores cercanos a la media lo contrario. No obstante, esta diferencia de tendencias según el género no es muy evidente o conlcuyente, por lo que este indicador por si mismo puede que no sea suficiente pero potencialmente puede aportar información relevante a un algoritmo de clasificación.

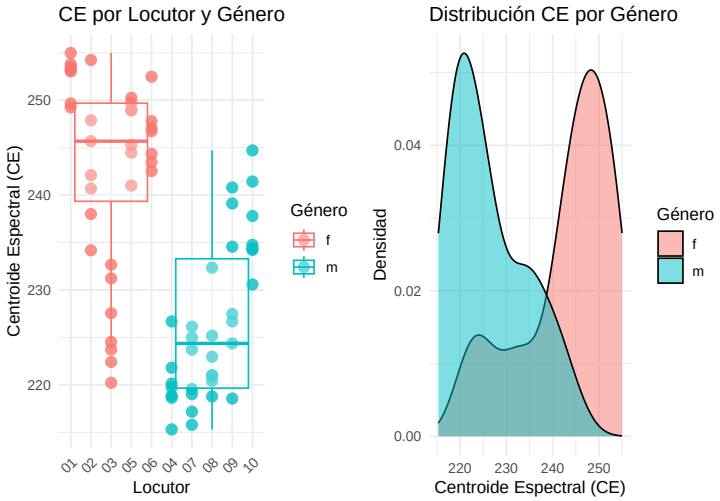
Centrándonos en el HNR según locutor, podemos ver que en el caso del género femenino hay cierta constancia en los valores, agrupándose en un rango no muy amplio que podría ayudar a su identificación. En cuánto al género masculino, sus HNR se encuentran distribuidos en intervalos semejantes por lo que podría ser considerablemente difícil discernir entre locutores masculinos.

2.10. Centroides Espectrales (CE)

Esta medida pretende caracterizar el espectro del audio señalando el "centro de masa" de su energía. Dicho centro se calcula realizando la media ponderada de las frecuencias encontradas usando una transformada de Fourier y usando sus amplitudes como pesos.

El centroide espectral de un audio tiene profunda relación con el timbre del mismo, en particular con su "brightness"; esta característica puede ser muy útil a la hora de diferenciar entre géneros e incluso de identificar patrones únicos en cada locutor.

Para su implementación en R nos ayudaremos de la librería seewave, la cual cuenta con esta medida dentro del resumen de estadísticas presentes en la función "specprop".



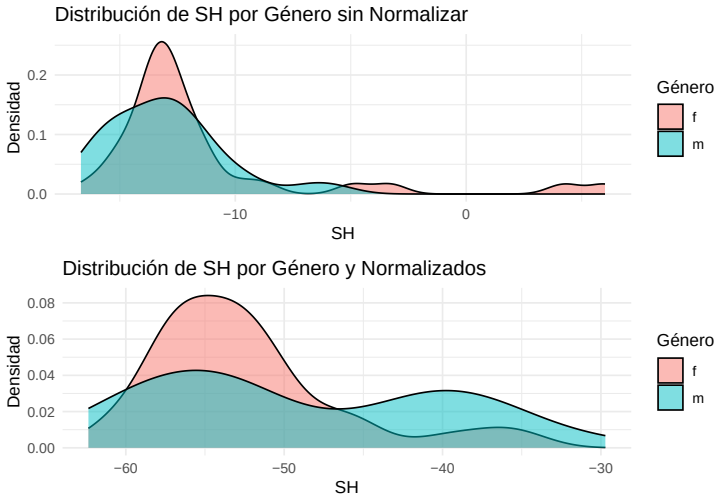
Notamos un comportamiento muy semejante al de la característica Harmonic-to-Noise Ratio (HNR): los CE de las voces masculinas se encuentran mucho más concentradas en torno a una media cercana a 175 mientras que las voces femeninas se reparten de forma más homogénea a lo largo del rango total 100-260, teniendo dos máximos locales en 120 y 230.

En cuánto al análisis por locutor, podemos observar considerable variabilidad de centroides espectrales dependiendo del audio (salvo en el caso del locutor 01), lo que puede dificultar el uso de esta medida para la identificación concreta del locutor. Sin embargo, al igual que en características anteriores la diferencia de distribución de densidad entre géneros puede aportar información relevante a la hora de clasificarlos.

2.11. Shimmer

La característica Shimmer mide la variabilidad de ciclo a ciclo de la amplitud en una señal de voz. Esta variabilidad se ve afectada por la tensión de las cuerdas vocales y factores fisiológicos.

Para capturar esta variabilidad en primer lugar diferenciamos la señal y normalizamos estas diferencias según las amplitudes de la señal original. Despues realizamos la media de estas diferencias y por ultimo transformamos el resultado de decibels.



Podemos observar en el grafico de densidad diferencias en la distribución por genero. La distribucion de SH para mujeres parece estar mas concentrada, tanto en los datos que se han normalizado y los que no, mientras que la distribucion para hombres parece estar más dispersa. Estas diferencias podrían reflejar características fisiológicas de las cuerdas vocales.

Aunque se observan diferencias claras entre géneros en términos de rango y dispersión de SH, también es evidente que hay un solapamiento y variabilidad entre locutores del mismo género.

Esto indica que SH puede ser una característica útil para clasificación de locutores dentro del mismo género o entre géneros, pero debería combinarse con otras métricas para obtener resultados más precisos.

2.12. Perceptual Linear Prediction

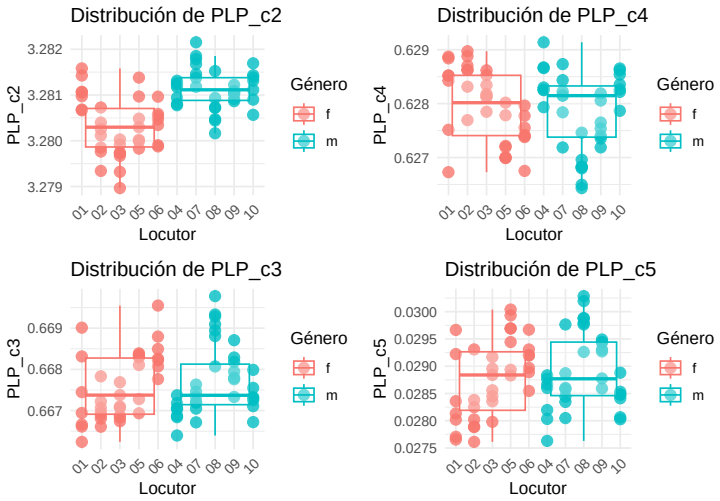
El Perceptual Linear Prediction (PLP) es una técnica de extracción de características basada en principios psicoacústicos que incorpora el modelo de predicción lineal (LPC) para representar señales acústicas de forma robusta y compacta. Aunque comparte similitudes con el cálculo tradicional de coeficientes LPC, PLP introduce modificaciones inspiradas en la percepción humana del sonido, ajustando el espectro de la señal para reflejar cómo el oído humano procesa y percibe las frecuencias.

Para obtener los coeficientes LPCC seguiremos los siguientes pasos:

1. Calculamos el espectrograma utilizando una ventana Hamming.
2. Extraer la densidad espectral de potencia (PSD).
3. Banco de filtros Bark: Filtramos la señal mediante un banco de filtros basado en las bandas críticas percibidas por el oído humano.
4. Aplicamos un filtro de pre-énfasis para ajustar las amplitudes de frecuencias para reflejar cómo el oído humano percibe las distintas intensidades a diferentes frecuencias.

5. Aplicamos una transformación para simular la potencia percibida por el oído humano. Cada elemento espectral se eleva a la potencia de 0.33. Este enfoque se utiliza en lugar de una transformación logarítmica porque el logaritmo tiende a comprimir de forma más agresiva los valores altos, mientras que la raíz cúbica proporciona una representación más fiel de la percepción auditiva, especialmente en señales con variaciones amplias de intensidad.
6. Cálculo de coeficientes LPC: A partir de cada segmento de la señal(usaremos la función lpc)
7. Calcular los coeficientes cepstrales (LPCCs): Transformamos los coeficientes LPC en coeficientes cepstrales LPCCS.

Por lo tanto, la dimensión de salida de nuestras características dependerá del orden seleccionado para los modelos lineales, así como del tamaño de la ventana y el solapamiento utilizado. En nuestro caso, siguiendo recomendaciones generales, obtenemos una salida aproximadamente de 14X6,247. Para compactar aún más la información, calcularemos la media de cada coeficiente, lo que nos dará un vector de longitud 14. A la hora de aplicar modelos, podríamos explorar tanto el uso de la matriz completa como el vector compacto que resume la información. Sin embargo, para la visualización de los valores de los coeficientes según el locutor, emplearemos el vector compacto.



Las distribuciones muestran una clara diferencia entre locutores, aunque no se observa una distinción significativa entre géneros al analizar un solo coeficiente en particular. Esto sugiere que el coeficiente individual podría ser útil para distinguir entre locutores. Sin embargo, al considerar los 14 coeficientes en conjunto, es probable que se puedan obtener mejores resultados para la clasificación por género.

3. Selección de características

En el siguiente apartado vamos a probar distintas formas de seleccionar características. Con el objetivo de clasificar mejor sin la necesidad de utilizar todos los recursos. Buscando obtener la máxima relevancia con el objetivo y la menor redundancia.

3.1. Importancias de las características usando RandomForest

En este apartado lanzaremos 1000 random forest, de los cuales obtendremos la importancia de las características. Quedándonos con las 10 mejores características. Siendo estas:

HNR_carac, lpcc_3_sd, lpcc_3_mean, lpcc_2_sd, ZCR_carac, lpcc_2_mean, lpcc_4_mean, MEL_COE4_6, ps_pow, SH_carac_norm.

3.2. Técnicas recursivas de eliminación de características (RFE)

Sin embargo, utilizaremos el RFE. Este consistirá en ejecutar algoritmos de selección de características que irá eliminando recursivamente las menos importantes. Probaremos a obtener las 10 mejores características.

El método devuelve que el número ideal de características es 8, siendo las variables más relevantes CE_carac, lpcc_3_sd, lpcc_2_sd, lpcc_4_sd, lpcc_2_mean, lpcc_3_mean, lpcc_4_mean, MEL_COE7_12.

3.3. Correlaciones

De las características calculadas previamente vamos a reducirlas a la mitad. Para ello utilizaremos las menos correlacionadas entre sí. Siendo estas: MEL_COE7_12, CE_carac, lpcc_2_mean, lpcc_4_mean.

4. Algoritmos de ML

El primer paso antes de hacer ningún algoritmo de ML es añadir a nuestro df de características las posibles etiquetas de estos algoritmos. En este caso las etiquetas se encuentran en el df meta_audios, siendo estas el locutor y el género.

De los algoritmos empleados hemos decidido utilizar un modelo más simple, que en este caso es el Random Forest, y uno más complejo; el SVM. Para evaluar como se compartan a partir de las características extraídas.

4.1. Métodos empleando como etiqueta el género

Primeramente, preparamos los datos separándolos en conjuntos de train y test. Es fundamental agrupar por género para garantizar un balance adecuado, evitando que un conjunto (por ejemplo, train) tenga una mayoría desproporcionada de un género, como “M”, y el otro conjunto, como test, tenga una mayoría de “F”. Este enfoque asegura una representación equitativa de ambos géneros en los datos de entrenamiento y prueba.

Se han empleado las características obtenidas de la selección.

Ahora separamos las etiquetas del conjunto de características de train

4.1.1. Random forest

Para el cual obtenemos la siguiente matriz de confusión:

```
##           Reference
## Prediction f m
##          f 9 0
##          m 0 9
```

4.1.2. SVM

Para el cual obtenemos la siguiente matriz de confusión:

```
##           Reference
## Prediction f m
##          f 9 0
##          m 0 9
```

4.1.3. Conclusiones

A partir de las matrices de confusión podemos observar lo siguiente:

1. Obtenemos un 100% de acierto con las características escogidas.
2. No es necesario emplear un modelo complejo para la clasificación por género. Ya que los resultados del Random Forest son los mismo que los del SVM.

4.2. Métodos empleando como etiqueta el locutor

Realizamos el mismo proceso anterior para dividir el conjunto de entrenamiento de forma equitativa entre los locutores. Esta vez, vamos a emplear todas las características. De los resultados mencionar que ya no obtenemos resultados perfectos. Obteniendo peores con el Random Forest que utilizando el SVM.

Obteniendo un Accuracy de 0.75 para el RF y Accuracy de 0.95 para el SVM.