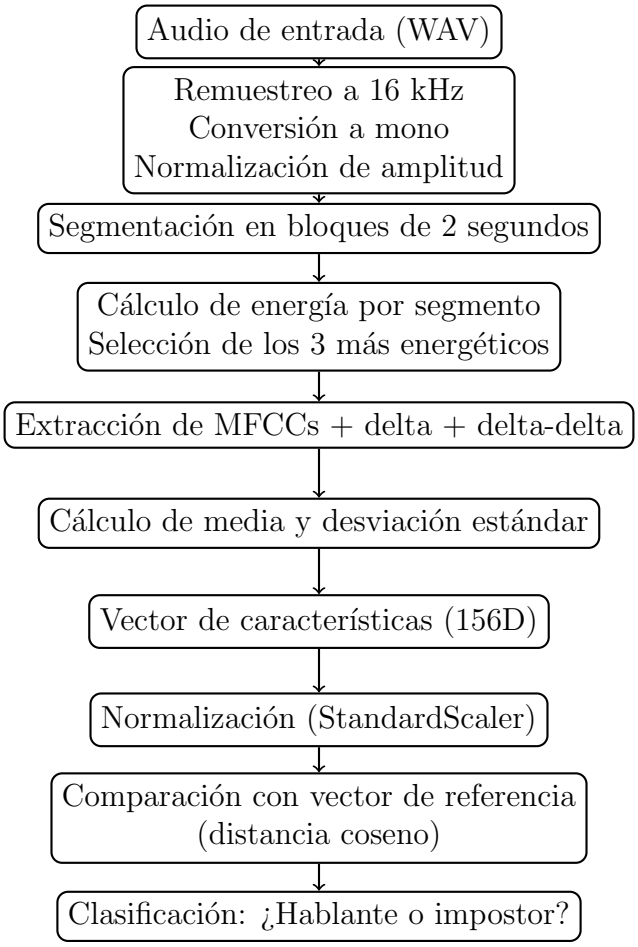
Metodología de Reconocimiento de Voz

Figura: Diagrama de la Metodología



Descripción de la Metodología

La metodología propuesta para el reconocimiento de voz se basa en representar cada audio mediante un vector de características espectro-temporales. Primero, cada archivo se convierte

a señal mono, se normaliza en amplitud y se remuestrea a 16kHz, una frecuencia estándar para señales de habla.

Luego, el audio se divide en segmentos de 2 segundos, y se calcula la energía promedio de cada uno. Se seleccionan los tres segmentos más energéticos, asumiendo que contienen contenido vocal más claro y menos silencio o ruido.

Sobre cada segmento seleccionado se extraen características usando coeficientes cepstrales en frecuencias Mel (MFCCs). Esta técnica se fundamenta en un banco de **128 filtros Mel** (triangulares y espaciados logarítmicamente), que imitan la percepción humana del sonido. Internamente, la señal se divide en ventanas aplicando una transformada rápida de Fourier (**FFT de tamaño 1024**) con un solapamiento del **50 %** (hop length = 512 para $fs = 16\,\text{kHz}$). En cada ventana se calculan **26 MFCCs** (`n_mfcc=26`), capturando la distribución energética en distintas bandas perceptuales.

Para capturar la dinámica temporal de la voz, se calculan también las derivadas de primer y segundo orden de los MFCCs, conocidas como **delta** y **delta-delta**, que describen cómo cambian los coeficientes en el tiempo. Así se obtiene una matriz de 78 características por ventana (26 MFCCs + 26 deltas + 26 delta-deltas).

Sobre esta matriz se aplican funciones estadísticas: se calcula la **media** y **desviación estándar** por cada coeficiente, generando un vector final de **156 dimensiones** por segmento. Promediando los vectores de los tres segmentos seleccionados se obtiene un vector representativo único por audio.

Antes de realizar comparaciones, todos los vectores se **normalizan** usando un *Standard-Scaler*, para eliminar efectos de escala y facilitar la discriminación.

Para la comparación, se utiliza la **distancia coseno**, que mide la orientación entre vectores independientemente de su magnitud. Se calcula esta distancia entre el audio de referencia y todos los audios del conjunto de prueba y de los impostores.

Se generan dos distribuciones de distancias: una para los audios válidos (test) y otra para los impostores. Estas distribuciones se suavizan mediante estimación de densidad (KDE), y se grafican. El **umbral de decisión óptimo** se encuentra como el punto de intersección entre ambas curvas suavizadas, usando el método de *Brent*. Este umbral separa mejor ambas clases, minimizando simultáneamente los errores de aceptación de impostores y rechazo de verdaderos positivos.