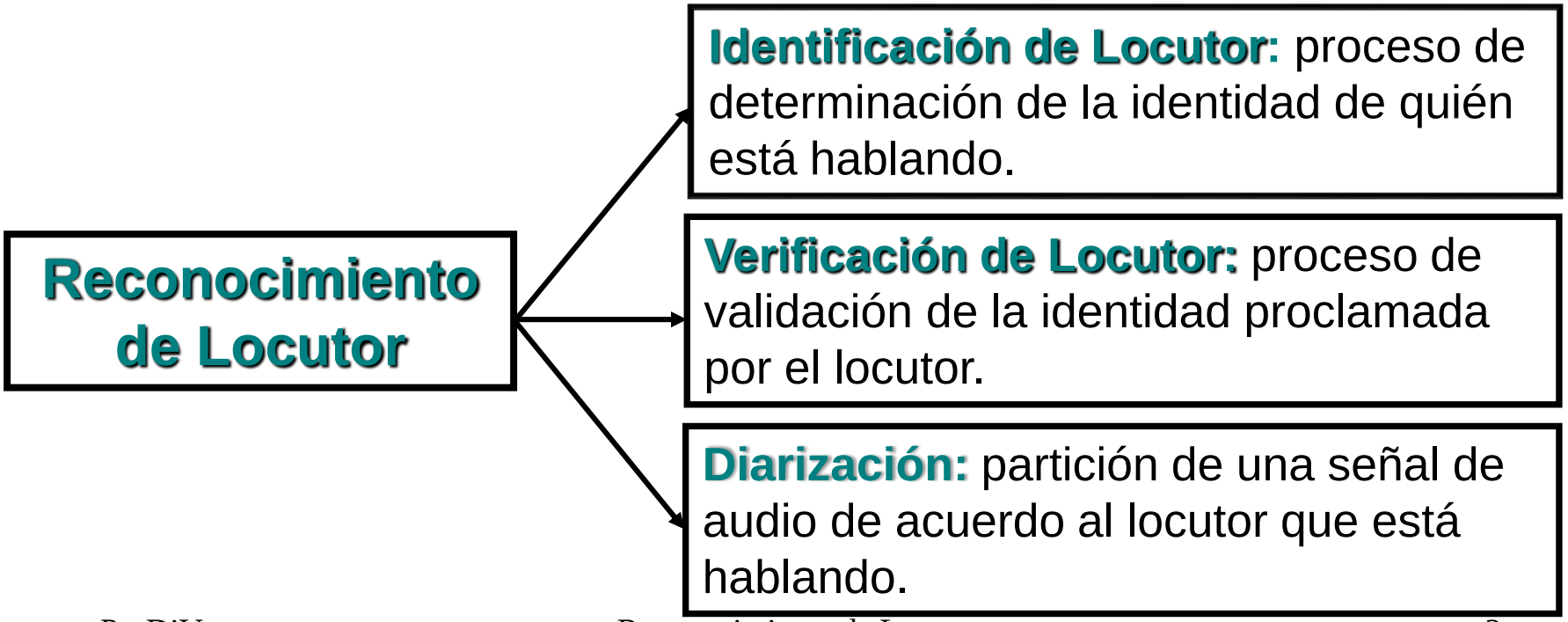
ProDiVoz

Reconocimiento de locutor

Introducción

Reconocimiento de locutor: Proceso de extracción automática de información relativa a la identidad de la persona a partir de muestras de voz. El proceso tiene dos etapas básicas: **entrenamiento** (recolección de muestras de voz de las personas a ser identificadas y generación de patrones o modelos asociados) y **reconocimiento** (comparación de los vectores característicos extraídos de las muestras del locutor desconocido con los patrones obtenidos en el entrenamiento, y toma de decisión).



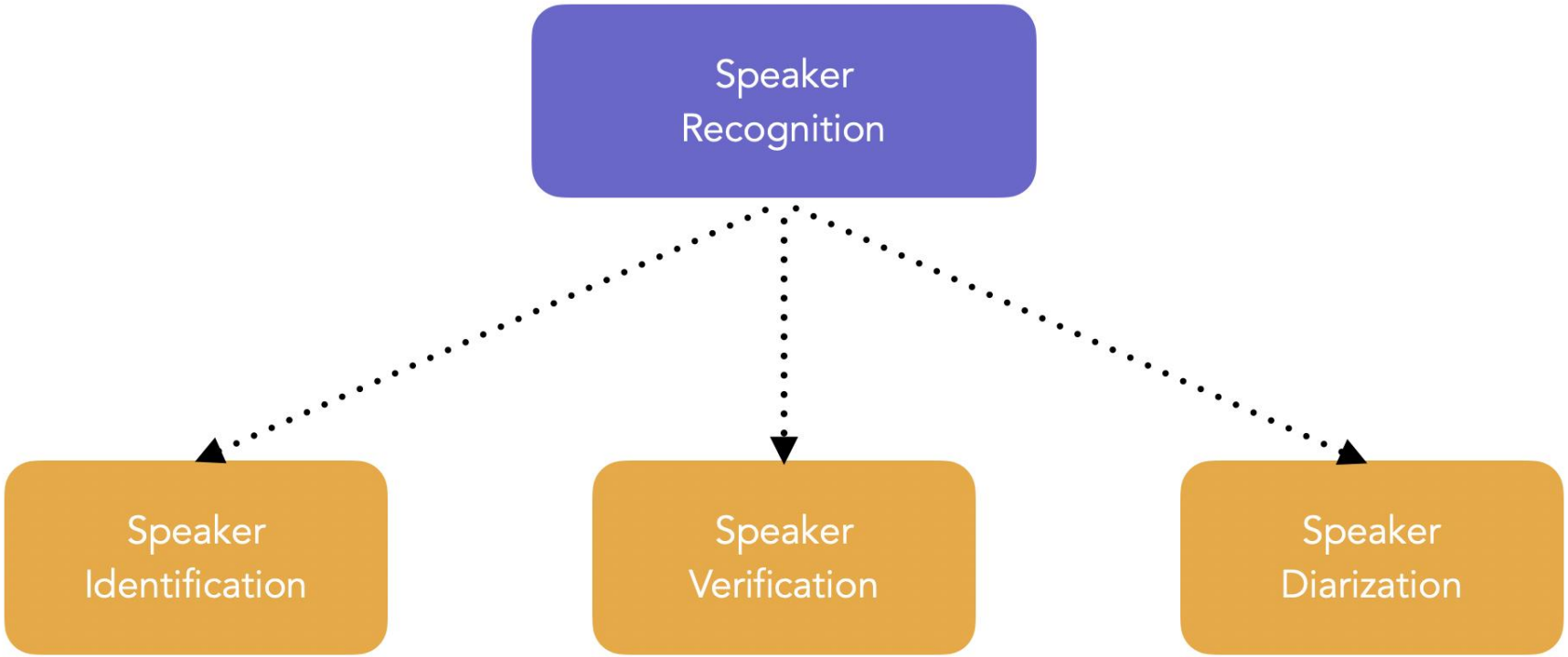
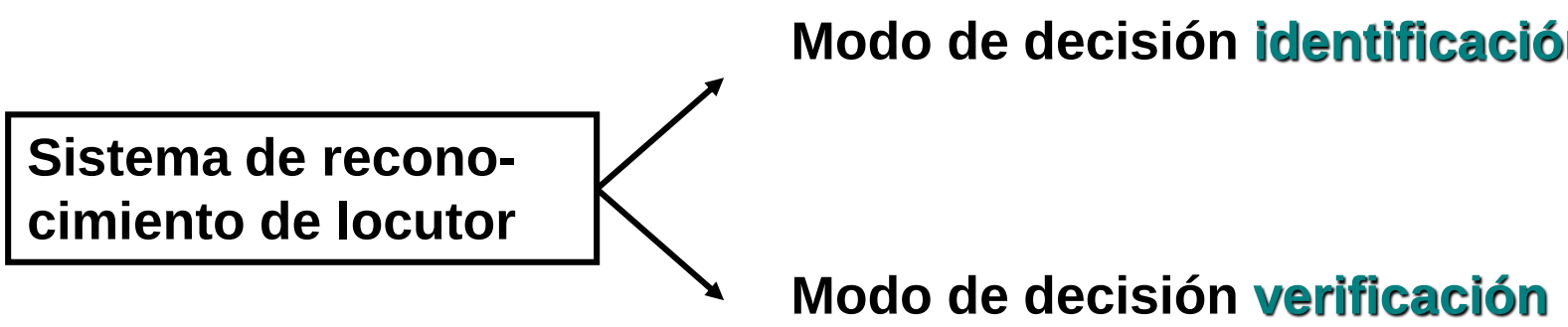


Fig.1 : Reconocimiento de Locutor

Numerosas Aplicaciones: acceso de seguridad controlado por voz, suministro de servicios o information particularizados activados por voz, rotulado de interlocutores en una conversación o diálogo grabados (**diarization**), investigaciones criminales y forenses que involucran muestras grabadas de voz, vigilancia, etc.

Aplicación más difundida: **control de acceso** → discado por voz, transacciones bancarias desde la red telefónica, compras telefónicas, servicios de acceso a base de datos, información y servicios reservados, correo por voz, acceso remoto a computadoras, etc.

Elementos básicos de un sistema de reconocimiento de locutor



Modo Identificación

Una muestra de voz de un locutor desconocido es analizada y comparada con modelos de locutores conocidos. El locutor desconocido es identificado como aquel en el conjunto de locutores conocidos cuyo modelo mejor se ajusta a las muestras de voz de entrada.

Existen dos formas de operación para el modo de identificación: **conjunto cerrado**, y **conjunto abierto**.

- **Conjunto cerrado:** el número de alternativas de decisión es igual a la dimensión de la población. El problema consiste en identificar un locutor dentro un conjunto de *N* locutores conocidos. Es identificado el locutor que obtiene el mejor puntaje (en la comparación) con la señal de ensayo.
- **Conjunto abierto:** un modelo de referencia del locutor desconocido puede que no exista. El problema consiste en decidir si un locutor pertenece a un grupo de *N* locutores conocidos.

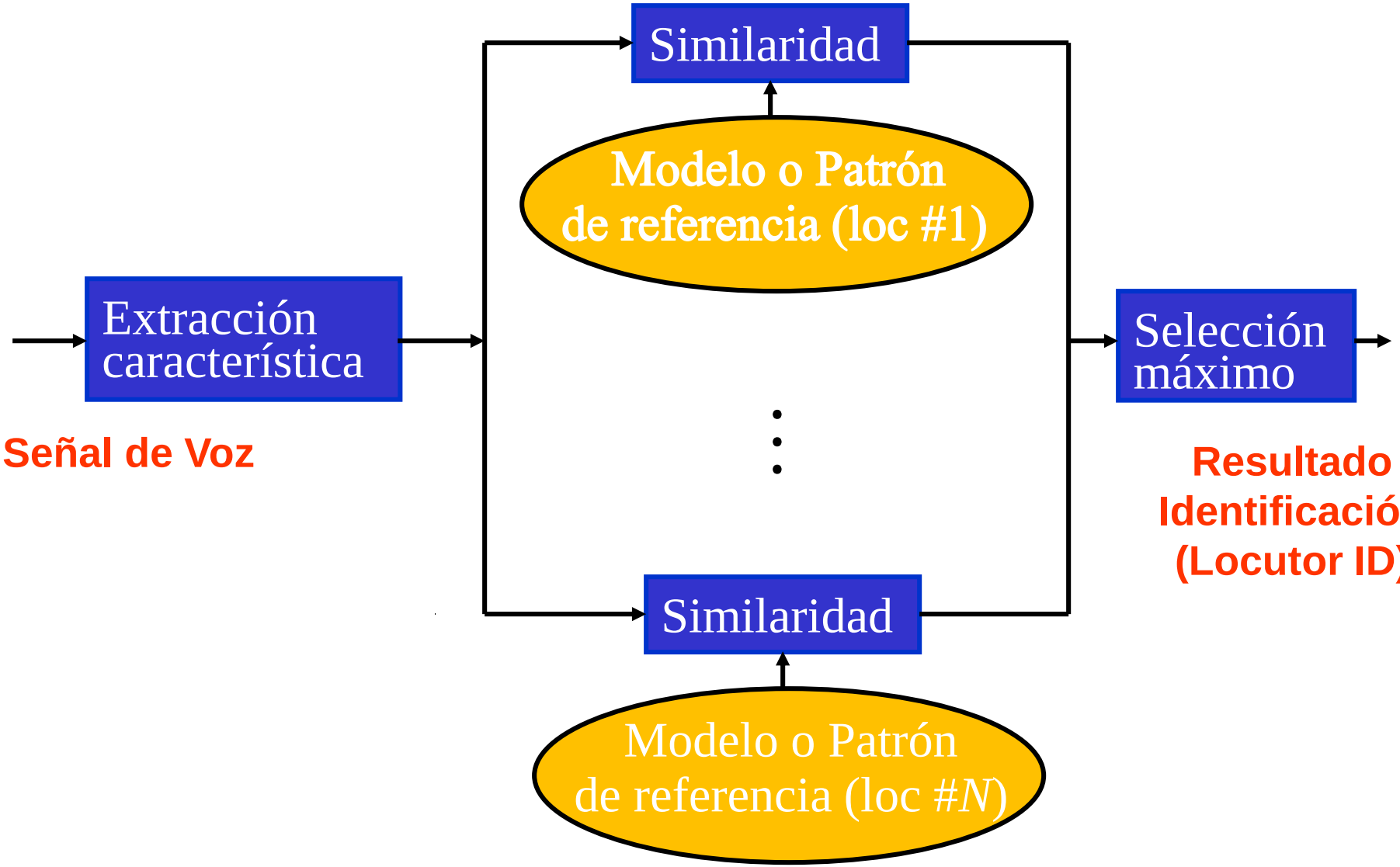


Fig. 2 : Reconocimiento de Locutor en Modo Identificación

Modo Verificación

Un locutor desconocido proclama una identidad. Las muestras de voz del locutor desconocido son comparadas con el modelo correspondiente al locutor cuya identidad fue proclamada. Si el ajuste, medido por ejemplo con un ensayo de umbral, es suficientemente bueno se verifica la identidad proclamada. Aquí existen sólo dos alternativas de decisión: **aceptar** o **rechazar** la identidad proclamada, independientemente de la dimensión de la población.

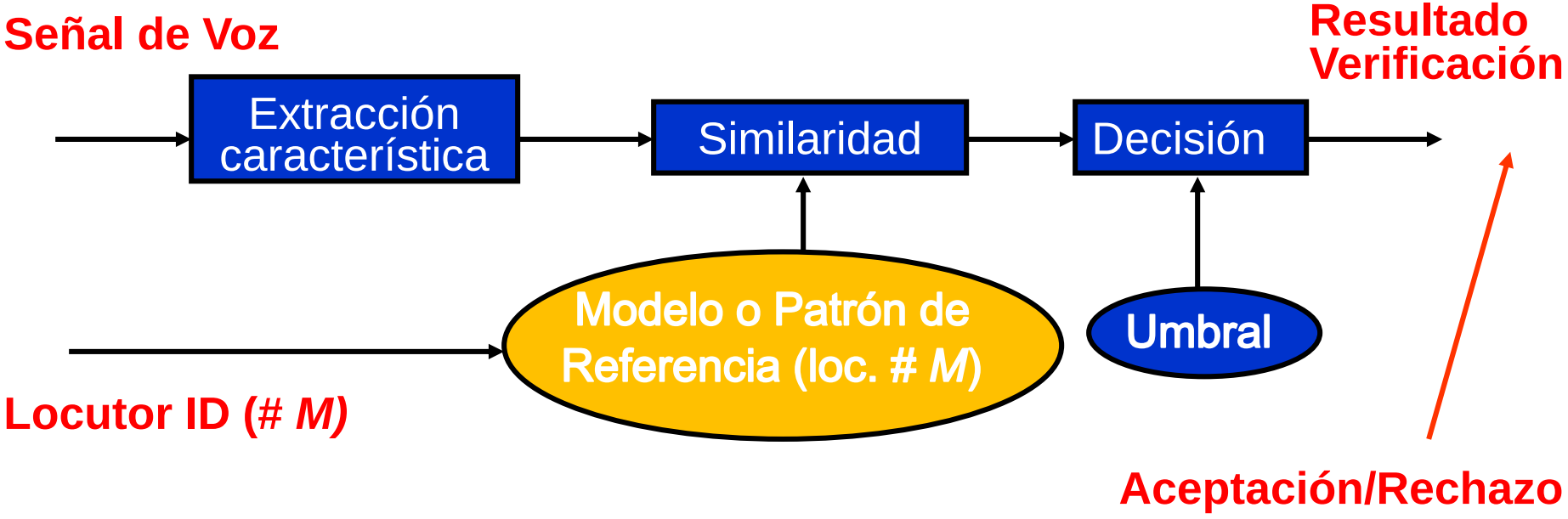


Fig. 3 : Reconocimiento de Locutor en Modo Verificación

Modos de Entrada

Típicamente los sistemas de reconocimiento de locutor operan en uno de los siguientes dos modos:

- **Texto-dependiente:** el locutor debe proveer muestras de voz correspondientes al mismo texto (**contraseña fija**) en las etapas de entrenamiento y de reconocimiento. La estructura de un sistema que usa una contraseña fija es relativamente simple. Las muestras se alinean temporalmente con los patrones o modelos de referencia creados en la etapa de entrenamiento correspondientes a la misma contraseña. Si las contraseñas son diferentes para diferentes locutores, esta diferencia puede usarse como información individual adicional para mejorar la performance del sistema.

- **Texto-Independiente:** el locutor no está restringido a proveer un texto específico en la etapa de reconocimiento.

Ambos modos de operación tienen potencialmente el problema de que alguien que reproduzca una grabación de las palabras clave de un locutor autorizado puede ser aceptado por el sistema. Una solución a esto son los sistemas que presentan al usuario contraseñas aleatorias (por ejemplo un conjunto de dígitos) que cambian cada vez que el sistema es usado.

Verificación de locutor usando Gaussian Mixtures Models (GMM)

- Dado un segmento de voz *Y,* y una hipótesis de locutor *S*, la verificación de locutor consiste en determinar si *Y* fue generado por *S.*
- Una hipótesis subyacente es que *Y* contiene habla de un sólo locutor. Si esto no se verifica la tarea se denomina verificación multi-locutor.
- La verificación de un sólo locutor puede reformularse como un **Test de Hipótesis** entre

H_0 : *Y* corresponde al locutor *S*

H_1 : *Y* **no** corresponde al locutor *S*

- El test óptimo corresponde al test de cociente de verosimilitudes (**likelihood ratio**)

$$\frac{p(Y/H_0)}{p(Y/H_1)} \begin{cases} \geq \theta & \text{aceptar } H_0 \\ < \theta & \text{rechazar } H_0 \end{cases}$$

donde $p(Y/H_i)$ con $i = 0, 1$ es la función de densidad de probabilidad de la hipótesis H_i evaluada en el segmento de voz Y, también denominada la **verosimilitud (likelihood)** de H_i dado el segmento de voz Y. El umbral de decisión para aceptar o rechazar la hipótesis H_0 es θ.

Se suele tomar el logaritmo de este cociente, lo que se conoce como cociente de verosimilitud logarítmica (**log-likelihood ratio**)

$$\Lambda(Y) = \log\left(p(Y/H_0)\right) - \log\left(p(Y/H_1)\right)$$

El aspecto fundamental del sistema de verificación de locutor es la forma en que se computan las verosimilitudes $p(Y/H_0)$ y $p(Y/H_1)$.

Las componentes básicas de un sistema de verificación de locutor basado en log-likelihood ratios se muestra en la Fig. 4.

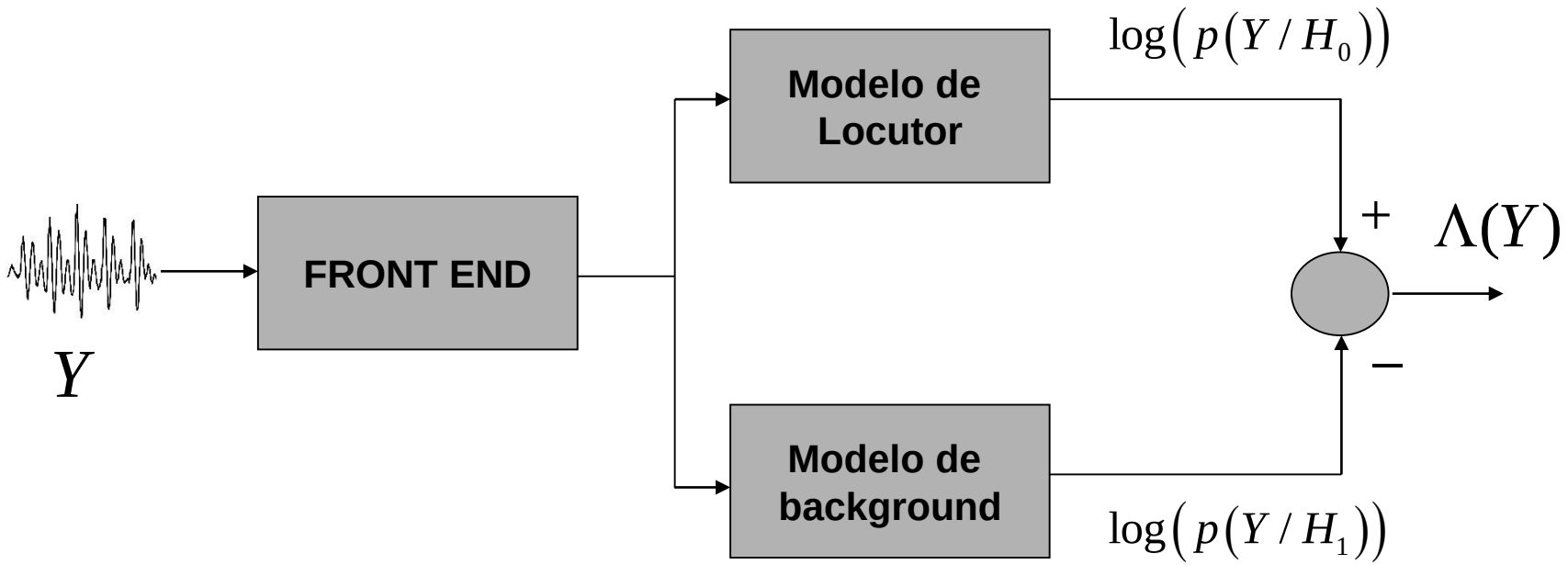


Fig. 4: Sistema de verificación de locutor basado en loglikelihood ratio

El bloque FRONTEND permite extraer las características de la señal de voz que contienen información que es propia del locutor particular. La salida de esta etapa es típicamente una secuencia de vectores característicos que representan al segmento de voz bajo análisis $X = \{x_1, x_2, \ldots, x_T\}$, donde x_t es el vector característico indexado en el tiempo discreto $t \in [1, 2, \ldots, T]$.

Hay dos enfoques para la construcción del Modelo de background:

1. Se utiliza un conjunto de modelos de otros locutores (**cohorte**) para cubrir el espacio de la hipótesis alternativa. Dado un conjunto de N modelos de locutores de background $\{\lambda_1, \lambda_2, \cdots, \lambda_N\}$, la hipótesis alternativa viene dada por una función

$$f\left(p(X \mid \lambda_1), \cdots, p(X \mid \lambda_N)\right)$$

que es usualmente el máximo o el promedio de las likelihoods.
Desventaja: requiere un modelo por cada locutor alternativo

2. Se utilizan muestras de un número grande de locutores para entrenar un **único** modelo que representa la hipótesis alternativa. Este modelo es denominado **Universal Background Model (UBM).**

El enfoque más exitoso para el cálculo de las probabilidades $p(X \mid \lambda)$ ha sido el uso de **Gaussian Mixtures Models (GMM)** , donde las densidades de probabilidad son representadas por la combinación lineal de M densidades Gaussianas unimodales $p_i(X)$, es decir

$$p(X \mid \lambda) = \sum_{i=1}^{M} w_i \, p_i(X)$$

donde cada $p_i(X)$ está caracterizada por su media $\mu_i \in \mathbb{R}^D$ y su matriz de covarianza $\sum_i$, y donde w_i son los pesos en la combinación lineal.

$$p_i(X) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i\right|^{1/2}} e^{-\frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i)}$$

Los GMM permiten representar densidades de probabilidad de forma arbitraria con reducida carga computacional y requerimientos de almacenamiento bajos. Pueden pensarse como un HMM con un único estado, con densidades de observación de mezclas Gaussianas, y con probabilidades de transición iguales.

Típicamente, para el **UBM** se utiliza un GMM con 512 a 2048 mezclas, y vectores característicos de dimensión 24, con matrices de covarianza diagonales (en contraposición a matrices llenas).
Bajo la hipótesis de independencia de los vectores característicos, la loglikelihood de que una secuencia $X = \{x_1, x_2, \cdots, x_T\}$ haya sido generada por el modelo es el promedio sobre todos los vectores característicos, es decir

$$\log\ p(X \mid \lambda) = \frac{1}{T}\sum_{t=1}^{T} \log\ p\left(x_t \mid \lambda\right)$$

Para el **Modelo de Locutor** se entrena un GMM con muestras de cada locutor. Existen dos enfoques para el Modelo de Locutor:

1. Entrenar un GMM de baja dimensión (de 64 a 256 mezclas) dependiendo de la cantidad de datos disponibles.
2. Adaptar el UBM GMM al locutor usando MAP (Maximum A Posteriori) Adaptation. Usando el algoritmo EM se adaptan solo las medias de las GM.(ya que adaptar las covarianzas no mejora la performance), según:

$$\mu_k^{MAP} = \alpha_k \mu_k + (1-\alpha_k)\mu_k^{UBM}$$

donde

$$\alpha_k = \frac{n_k}{n_k + \tau_k}$$

τ_k factor de relevancia, entre 8 y 32.