Reconocimiento de Locutor basado en Procesamiento de Voz

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 **reconocimiento** (comparación de las muestras del locutor desconocido con los datos de entrenamiento, y toma de decisión).

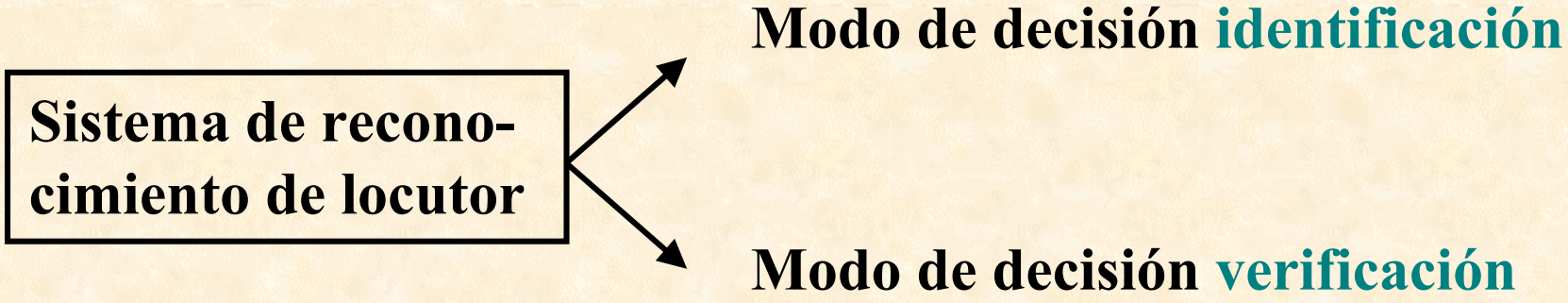
Identificación de Locutor: proceso de determinación de la identidad de quién está hablando.

Verificación de Locutor: proceso de validación de la identidad proclamada por el 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, 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, informacion y servicios reservados, correo por voz, acceso remoto a computadoras, etc.

Elementos básicos de un sistema de reconocimiento de locutor



Modo Identificación

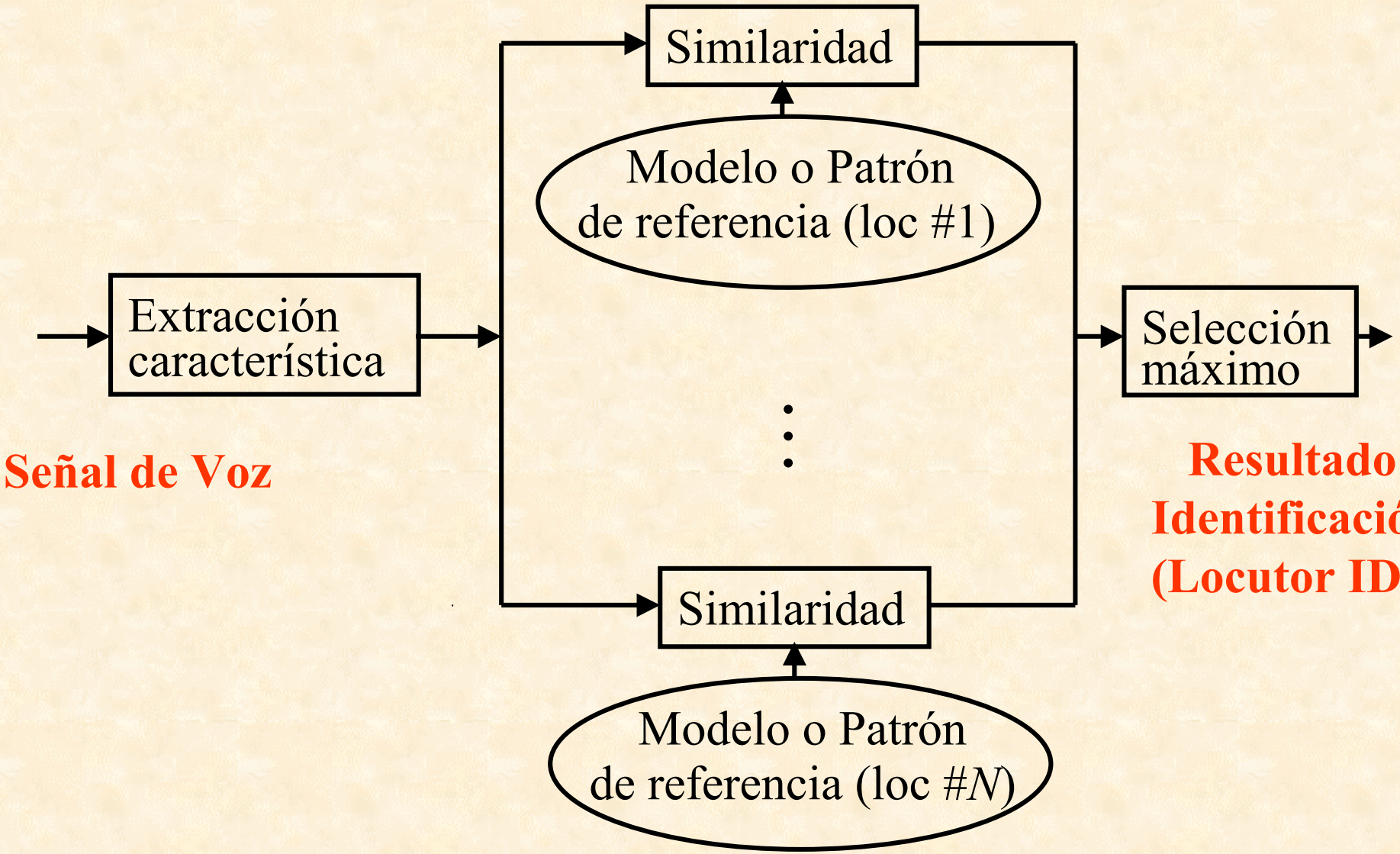
Una muestra de voz de un locutor desconocido es analizada y comparada con modelos de locutores conocidos. El locutor desconocido es identificado como aquel en el conjunto de locutores conocidos cuyo modelo mejor se ajusta a las muestras de voz de entrada.

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

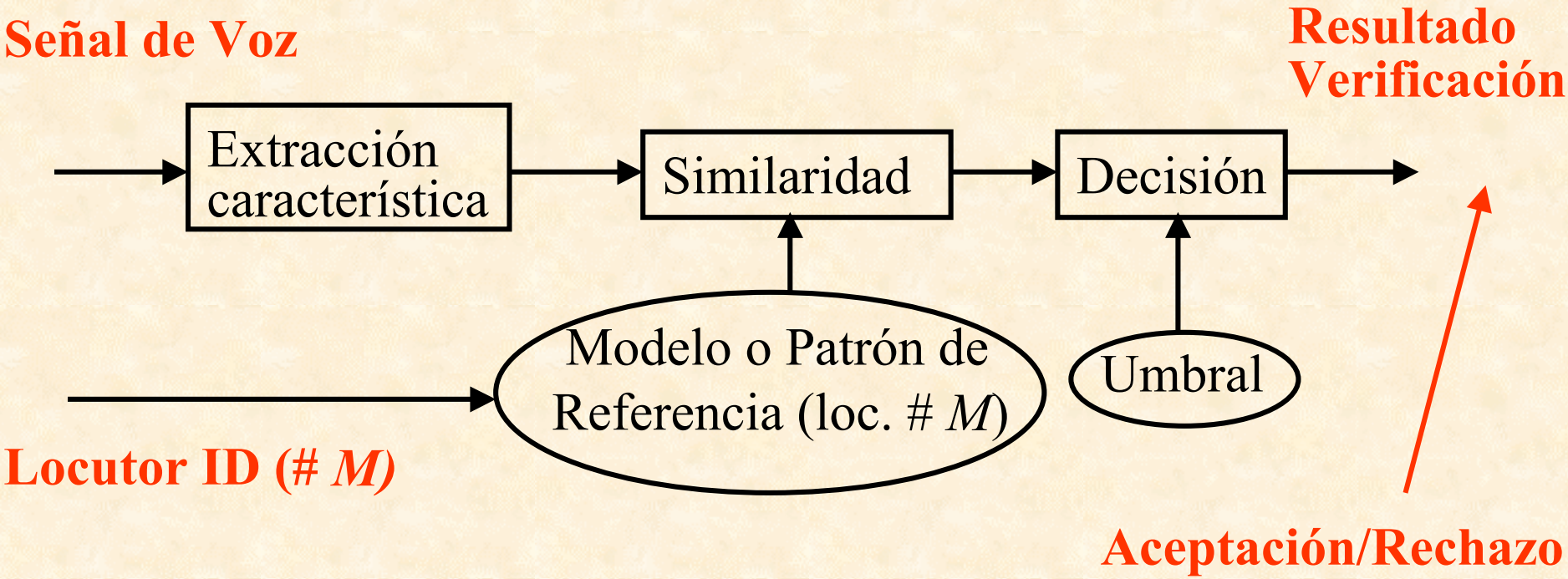
- **Conjunto cerrado:** el número de alternativas de decisión es igual a la dimensión de la población. El problema consiste en identificar un locutor dentro un conjunto de N locutores conocidos. Es identificado el locutor que obtiene el mejor puntaje (en la comparación) con la señal de ensayo.
- **Conjunto 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.

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.



Modo Identificación



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.

Extracción de característica

La extracción de característica se realiza generalmente mediante **análisis espectral de corta duración**, en particular se suele utilizar **análisis con banco de filtros**, o **análisis LPC**.

Análisis LPC

- El segmento de voz es dividido en cuadros (de entre 10 y 30 mseg) que usualmente tienen un solapamiento.
- Cada cuadro es pasado por una ventana para disminuir la distorsión.
- Para cada cuadro se determinan los coeficientes LPC cepstral (usualmente entre 10 y 20 coeficientes) que constituyen los vectores característicos ⇒ **reducción** de la tasa de información. Por ejemplo:

cuadro = 20 mseg, a 8 KHz, 8 bits por muestra ⇒ **1280 bits/cuadro**

14 coef. LPC cepstral por cuadro ⇒ **reducción 1280/(14*8)=11.4**

Medidas de Distorsión

• Una característica fundamental de los sistemas de reconocimiento de locutor es la forma en que los vectores característicos son combinados y comparados con los modelos de referencia.

• Para poder realizar estas operaciones es necesario definir una **medida de distorsión** o **distancia** entre vectores característicos.

• Algunas de las medidas de distorsión más utilizadas son las distancias o métricas inducidas por las normas en espacios L_p.

Por ejemplo, si f_i , $f_i^{'}$ con $i = 1,2,\cdots,D$ son las componentes de dos vectores característicos f y $f^{'}$, pueden definirse las siguientes normas L_p

$$d_1 = \sum_{i=1}^{D} \left| f_i - f_i^{'} \right| \qquad \textbf{distancia} \quad L_1$$

$$d_2 = \sqrt{\sum_{i=1}^{D} \left(f_i - f_i'\right)^2}$$ **distancia Euclidea (o L_2)**

• Una medida de distorsión muy utilizada para los coeficientes cepstral, que ha probado tener una muy buena performance tanto para reconocimiento de locutor como para reconocimiento de palabra, es la **distancia Euclidea ponderada**, definida como

$$d_{2w} = \sqrt{\sum_{i=1}^{D} \left(w_i \left(c_i - c_i'\right)\right)^2}$$

donde

$$w_i = \frac{1}{\sigma_i}$$

siendo σ_i^2 una estima de la varianza del i-ésimo coeficiente cepstral c_i .

• Una formulación más general, que tiene en cuenta la interacción entre coeficientes a traves de una matriz de covarianza es la denominada **distancia de Mahalanobis,** definida como

$$d_M = \sqrt{(\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu)}$$

donde

$$\mu \approx \frac{1}{N_e} \sum_{i=1}^{N_e} x_i^e$$

Media de los vectores de entrenamiento x_i^e

$$\Sigma \approx \frac{1}{N_e} \sum_{i=1}^{N_e} (x_i^e - \mu)(x_i^e - \mu)^T$$

Matriz de Covarianza de los vectores de **entrenamiento** x_i^e

$$\bar{x} = \frac{1}{N_r} \sum_{i=1}^{N_r} x_i^r$$

Media de los vectores de **reconocimiento** x_i^r

Métodos de Reconocimiento de Locutor

Enfoque the Ajuste de Plantillas (Template Matching)

Para distinguir entre distintos locutores se compara un promedio de los vectores característicos obtenidos en la etapa de reconocimiento usando un segmento relativamente largo de señal de voz, con una colección de promedios obtenidos durante la etapa de entrenamiento.

Para el caso de modo **texto-independiente** se necesita en general un segmento de varios segundos o minutos de señal de voz para asegurar que la voz del locutor pueda ser correctamente modelada por las características promedio de los vectores característicos sobre un rango amplio de sonidos, en vez de por sonidos particulares o fonemas.

Esta metodología es también conocida como **promediado estadístico de característica .** Es sensible a variaciones del canal de transmisión y a la presencia de ruído de fondo que modifican las características promedio de los vectores característicos.

Modelado Probabilístico de Locutores

• Los vectores característicos se consideran variables aleatorias caracterizados por una función de distribución de probabilidad en vez de por valores promedios (media y covarianza). La clasificación es basada en **medidas de verosimilitud** en vez de en medidas de distorsión o distancia entre vectores de reconocimiento y de referencia.

• Asumiendo que los locutores tienen una distribución conocida, con funciones de densidad de probabilidad continuas p_i , entonces la probabilidad de que un vector característico x haya sido generado por el i-ésimo locutor es $p_i(x)$. Usando la **Regla de Bayes**, la probabilidad de que dado un x, el locutor sea el i-ésimo es

$$P(\text{locutor} = i \mid x) = \frac{p_i(x) P_i}{p(x)}$$

donde P_i es la probabilidad *a priori* de que el vector haya sido generado por el *i*-ésimo locutor, y *p(x)* es la probabilidad de que el vector haya sido generado por cualquier locutor, es decir, es el promedio de las densidades de los locutores

$$p(x) = \sum_{i=1}^{N} p_i(x) P_i$$

donde N es el número de locutores. Típicamente $P_i = P_j \quad , \forall \; i, j$, por lo que resulta

$$P(\text{locutor} = i \mid x) = \frac{p_i(x) P_i}{p(x)} = \frac{p_i(x)}{\sum_{j=1}^{N} p_j(x)}$$

- El problema es cómo estimar las densidades de probabilidad $p_i(x)$

Estimación de densidad de probabilidad

- **Método del "vecino más cercano"**

La idea es estimar la función de densidad de probabilidad en un punto x , a partir de un conjunto de muestras $R = \{r_i\}$. Para ello se mide la distancia entre x y aquel punto en R que sea el más próximo a x, es decir **el vecino más cercano a *x*** ,

$$d_{NN}(x, R) = \min_{r_j \in R} |x - r_j|$$

Intuitivamente, cuanto más chica sea la distancia al vecino más cercano, mayor será la densidad de probabilidad. Específicamente, una estima de la densidad de probabilidad en x, puede calcularse como

$$\hat{p}(x) = \frac{1}{V_n(d_{NN}(x, R))}$$

donde $V_n(\rho)$ es el volumen de la esfera de radio ρ en un espacio *n*-dimensional, donde *n* es la dimensión del vector *x*.

Tomando logaritmos, y considerando que el volumen de una esfera de radio ρ es proporcional a ρ^n, resulta

$$\ln \hat{p}(x) = -n \ln(d_{NN}(x, R))$$

Sean $U = \{u_i\}$ y $R = \{r_i\}$ conjuntos de vectores característicos correspondientes a las etapas de reconocimiento y entrenamiento respectivamente. El conjunto *R* será usado para estimar la densidad de probabilidad del locutor (en realidad se tiene un conjunto *R* por cada locutor perteneciente al conjunto de la población). Se tiene

$$\ln \hat{p}(u_i) = -n \ln(d_{NN}(u_i, R))$$

y la verosimilitud (logarítmica) del locutor resulta

$$\hat{\ell}(U) = -\sum_{u_i \in U} \ln\left(d_{NN}(u_i, R)\right)$$

Es identificado aquel locutor que tenga la más grande verosimilitud, es decir aquel locutor cuyo modelo de referencia *R* está más próximo al conjunto de reconocimiento *U*.

El **Método del Vecino más Cercano** es un método **no paramétrico.**

Una variante propuesta por Higgens *et al.* define una distancia simétrica entre los conjuntos de reconocimiento *U* y de referencia *R*.

$$D(U,R) = \frac{1}{|U|}\sum_{u_i \in U} d_{NN}^2(u_i, R) + \frac{1}{|R|}\sum_{r_j \in R} d_{NN}^2(r_j, U) -$$

$$\frac{1}{|U|}\sum_{u_i \in U} d_{NN}^2(u_i, U - \{u_i\}) - \frac{1}{|R|}\sum_{r_j \in R} d_{NN}^2(r_j, R - \{r_j\})$$

• El primer término es similar a la verosimilitud logarítmica, excepto que en vez de distancias logarítmicas, las distancias están elevadas al cuadrado y sumadas, lo que da mejores en la práctica.

• El segundo término es para simetrizar la distancia haciendo

$$D(U,R) = D(R,U)$$

• Los últimos dos términos son medidas de dispersión de los vectores en los conjuntos de reconocimiento U, y de referencia R, respectivamente. Estos términos están presentes para evitar que aquellos vectores que tienen varianzas grandes no sean indebidamente penalizados.

• Finalmente los términos están normalizados por los factores $1/|U|$ y $1/|R|$, para compenzar por las diferentes cardinalidades de los conjuntos U, y R.