Reconocimiento Automático de Voz

Presentación basada en las siguientes **Referencias**:

[1] Rabiner, L. & Juang, B-H.. *Fundamentals of Speech Recognition,* Prentice Hall, N.J., 1993.

[2] Rabiner, L. & Juang, B-H.. *Speech Recognition by Machine*, Chap. 47 in *The Digital Signal Processing Handbook*, CRC Press, IEEE Press, 1998.

Caracterización de los Sistemas de Reconocimiento de Voz

Los sistemas de reconocimiento automático de voz se caracterizan teniendo en cuenta diferentes aspectos:

1. **La forma en que el usuario le habla a la máquina**. Existen básicamente tres formas:

- **Palabra Aislada:** el usuario habla palabras individuales (o frases) tomadas de un vocabulario determinado.
- **Palabras Conectadas:** el usuario habla en forma fluida una sucesión de palabras pertenecientes a un vocabulario restringido (*e.g.* dígitos telefónicos).
- **Habla continua:** el usuario habla fluidamente usando palabras de un vocabulario grande (usualmente ilimitado).

2. Tamaño del vocabulario de reconocimiento

- **Pequeño:** capaz de reconocer hasta 100 palabras.
- **Mediano:** entre 100 y 1000 palabras.
- **Grande:** más de 1000 palabras.

3. El conocimiento de los patrones de voz del usuario

- **Sistemas dependientes del locutor:** adaptados a locutores particulares.
- **Sistemas independientes de locutor:** trabajan con un población de locutores grande, la mayoría de los cuales son desconocidos para el sistema.
- **Sistemas adaptables:** se adaptan al locutor particular mientras el sistema está en uso.

4. Grado de conocimiento acústico-lingüístico usado por el sistema.

- **Sólo conocimiento acústico**. No usan conocimiento lingüístico.
- **Integración de conocimiento acústico y lingüístico.** El conocimiento lingüístico está usualmente representado por restricciones sintácticas y semánticas sobre la salida del sistema de reconocimiento.

5. Grado de diálogo entre el usuario y la máquina.

- **Unidireccional** (o pasivo). El usuario habla y la máquina realiza una acción como respuesta.
- **Sistema de diálogo activado por la máquina.** El sistema es el iniciador del diálogo, requiriendo información del usuario via una entrada verbal.
- **Sistema de diálogo natural.** La máquina "conversa" con el locutor, le solicita entradas, actúa en función de las entradas y trata de clarificar ambigüedades.

Fuentes de Variabilidad de las Señales de Voz

El reconocimiento automático de voz es una tarea inherentemente difícil debido a la variabilidad de las señales de voz. Algunas fuentes de variabilidad incluyen:

- ❑ Variabilidad en un locutor en mantener una pronunciación consistente y en el uso de palabras y frases.
- ❑ Variabilidad entre locutores debido a diferencias fisiológicos (*e.g.* diferente longitud del tracto vocal), acentos regionales, idiomas extranjeros, etc.
- ❑ Variabilidad entre transductores cuando se habla frente a diferentes micrófonos o aparatos telefónicos.
- ❑ Variabilidad introducida por el sistema de transmisión (redes de comunicación teléfonos celulares, etc.).
- ❑ Variabilidad en el ambiente, que incluyen conversaciones extrañas y eventos acústicos de fondo, como ruidos, etc.

Enfoques de Reconocimiento Automático de Voz

1. Enfoque Acústico-Fonético

Consiste en detectar sonidos elementales y asignarles determinados rótulos. La base de este enfoque es la hipótesis de que en el lenguaje hablado existe un número finito de unidades fonéticas distintas (**fonemas**) y que estas unidades pueden caracterizarse por un conjunto de propiedades acústicas que se manifiestan en la señal hablada en función del tiempo. Si bien las propiedades acústicas de los fonemas son altamente variables con el locutor y con los fonemas vecinos (co-articulación de sonidos), se asume que las reglas que gobiernan la variabilidad son simples y pueden ser aprendidas fácilmente por el sistema de reconocimiento.

El reconocimiento consiste básicamente de dos pasos:

- **Primer paso:** segmentación y rotulado. La señal es dividida en regiones acústicas a las que son asignados uno o más fonemas, resultando en una caracterización de la señal de voz mediante un reticulado de fonemas.

- **Segundo paso:** se trata de determinar una palabra (o conjunto de palabras) válida a partir de la secuencia de fonemas rotulados en el primer paso. Se introducen en esta etapa restricciones lingüísticas (vocabulario, sintaxis, y reglas semánticas)

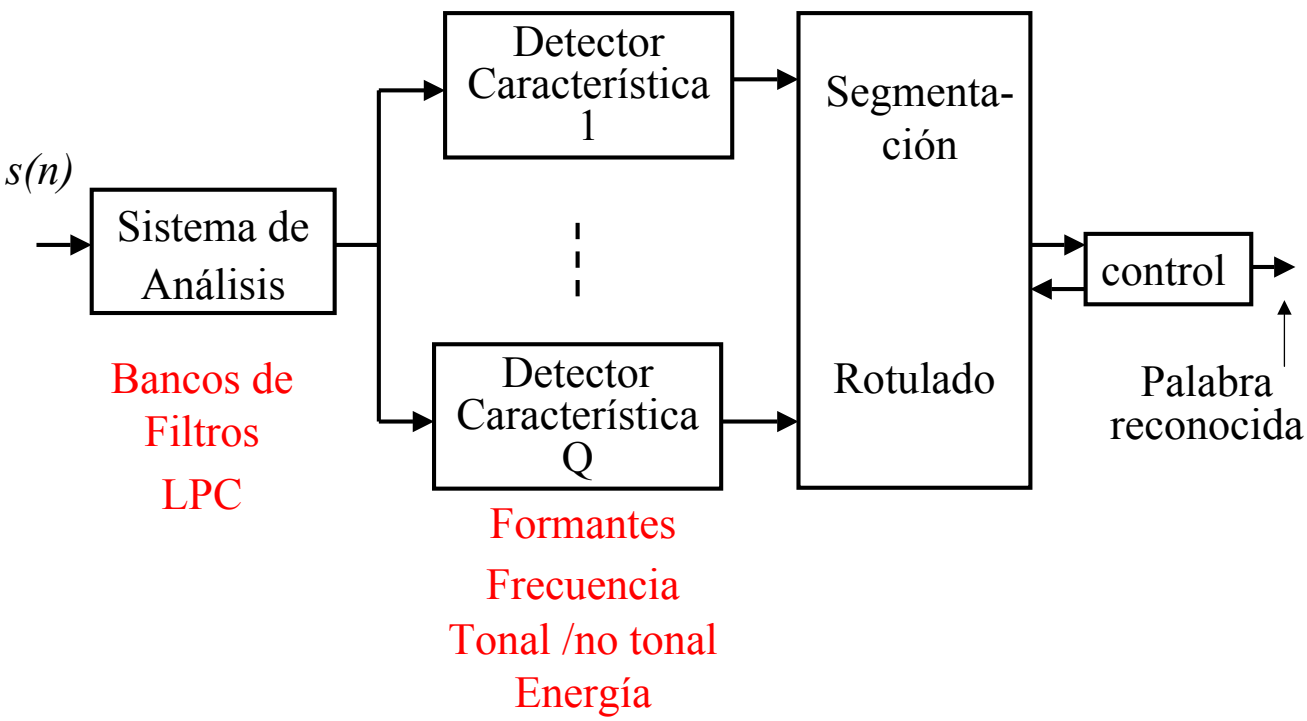

Fig. 1: Sistema de Reconocimiento de Voz basado en enfoque acústico-fonético.

La primera etapa en el procesamiento (que es común a todos los enfoques) es la etapa de **análisis de voz**, que provee una representación (espectral) de las características inestacionarias de la señal de voz. Los métodos más comunes en esta etapa son **análisis con banco de filtros** y **análisis LPC (Linear Predictive Coding)**.

En la siguiente etapa es la **extracción de característica** en donde se convierten las medidas espectrales en un conjunto de parámetros que describen la propiedades acústicas de las unidades fonéticas. Estos parámetros pueden ser: nasalidad (presencia o ausencia de resonancia nasal), fricación (presencia o ausencia de excitación aleatoria en la voz), ubicación de los formantes (frecuencias de las 3 primeras resonancias), clasificación entre sonidos tonales y no tonales, etc.

La tercer etapa del procesamiento es la etapa de **segmentación y rotulado** en donde el sistema trata de encontrar regiones estables donde las características cambian poco, que son rotuladas teniendo en cuenta cuan bien la característica en la región se ajusta a unidades fonéticas individuales. Esta es usualmente la etapa más difícil de llevar a cabo en forma confiable.

El resultado de la etapa de segmentación y rotulado es un reticulado de fonemas a partir del cual se determina la palabra (o secuencia de palabras) que mejor se ajusta, teniendo en cuenta restricciones lingüísticas (de vocabulario, de sintaxis, y semánticas).

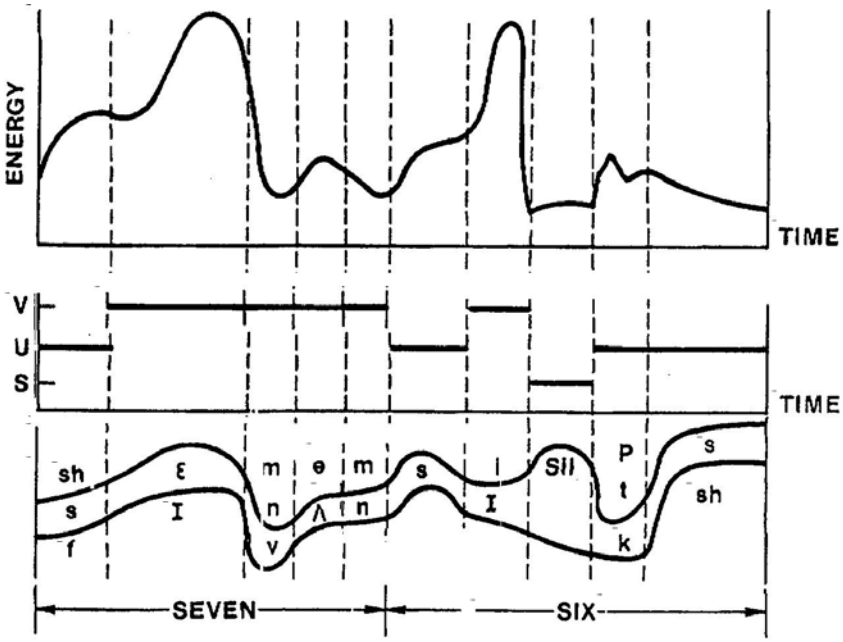

Fig. 2: Segmentación y rotulado de la secuencia de dígitos en inglés *seven-six* (tomado de [1]).

2. Enfoque de Reconocimiento de patrones

Consiste básicamente en dos pasos:

- **Primer Paso:** entrenamiento de patrones
- **Segundo Paso:** comparación de patrones

La característica principal de este enfoque es que usa un marco matemático bien definido y que establece representaciones consistentes de los patrones de voz que pueden usarse para comparaciones confiables a partir de un conjunto de muestras rotuladas, usando algoritmos de entrenamiento. La representación de los patrones de voz puede ser una **plantilla** (template), o un **modelo estadístico** (HMM: Hidden Markov Model), que puede aplicarse a un sonido (más pequeño que una palabra), una palabra, o una frase.

En la etapa de comparación de patrones se realiza una comparación directa entre la señal de voz desconocida (a reconocer) y todos los posibles patrones aprendidos en la etapa de entrenamiento, de manera de determinar el mejor ajuste de acuerdo a algún criterio.

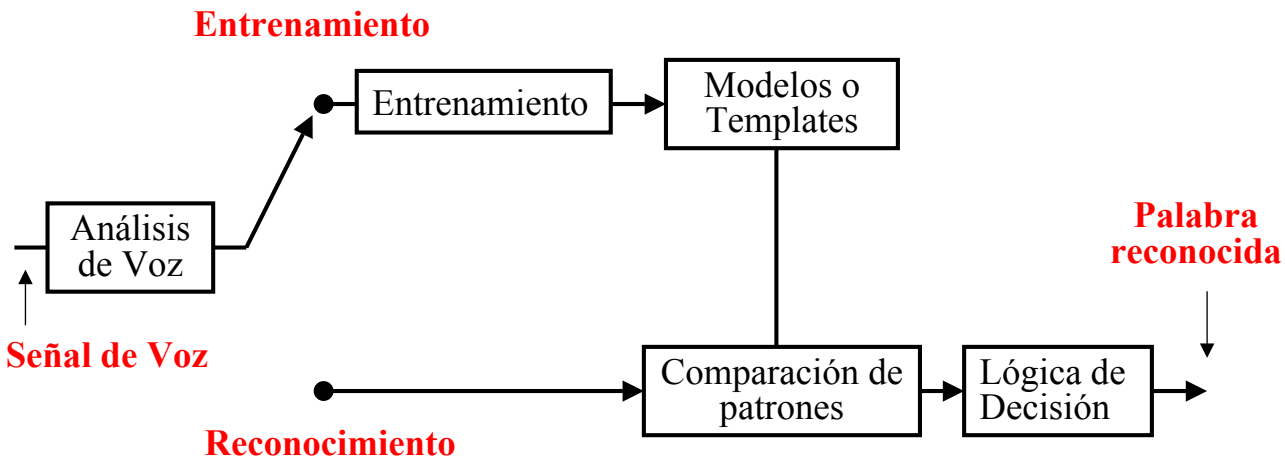

Fig. 3: Reconocimiento de Voz basado en reconocimiento de patrones.

3. Enfoque de Inteligencia Artificial

En este enfoque se intenta automatizar el procedimiento de reconocimiento de acuerdo a la forma en que una persona aplica su inteligencia en la visualización, análisis y caracterización de la voz basada en un conjunto de características acústicas. Algunas técnicas que se emplean son: sistemas expertos (redes neuronales) que integran conocimientos prácticos fonéticos, sintácticos, semánticos para la segmentación y el rotulado, y usan herramientas tales como rede neuronales artificiales para aprender las relaciones entre eventos fonéticos.