FUNDAMENTOS DEL RECONOCIMIENTO AUTOMÁTICO DE LA VOZ

"Concepto y clasificación de los reconocedores de voz"

Agustín Álvarez Marquina

Planteamiento del problema

- La forma mas natural de comunicación para la mayoría de la gente es el lenguaje hablado.
- Objetivo buscado: crear máquinas que reciban información hablada y actúen en consecuencia.
- Encontrar una máquina robusta, inteligente, que permita una conversación fluida es una meta lejana.
- Las aplicaciones desarrolladas hoy en día son aún muy limitadas.

Aplicaciones demandadas (I)

- **Todas aquellas donde es deseable poder reconocer preferentemente:**
  - Habla continua.
  - Múltiples hablantes con diversos acentos, formas de hablar, vocabularios y tendencias gramaticales.
  - Articulaciones pobres.
  - Ambientes ruidosos.
  - Sistemas que pudiesen aprender.

Aplicaciones demandadas (II)

- **Los sistemas actuales, con todas sus limitaciones son de utilidad en:**
  - Ambientes industriales (ej: tareas de ordenación).
  - Aplicaciones manos libres como por ejemplo.
    - Cirugía.
    - Asistencia a discapacitados.
    - Cabinas de pilotos.
  - Aplicaciones remotas.
  - Servicios telefónicos.

Definición de sistema reconocedor de voz (I)

- **El reconocimiento automático del habla es el procedimiento por el cual se convierte una señal acústica, capturada por un micrófono, en un conjunto de símbolos de un diccionario dado. Dichos símbolos están generalmente asociados con elementos semánticos de tipo palabra [COL97].**
  - R. A. Cole et al., *Survey of he State of the Art in Human Language Technology*, Cambridge University Press, 1997.

Definición de sistema reconocedor de voz (II)

- **Las palabras reconocidas pueden constituir el resultado final, como en aplicaciones de control, entrada de datos y preparación de documentos.**
- **También puede servir como entrada a otros módulos de procesamiento lingüístico con el fin de alcanzar la comprensión de la realización oral.**

Definición de sistema reconocedor de voz (III)

- **La forma de resolver el problema de reconocimiento suele adoptar una aproximación estadística basada en encaje de patrones al considerar que la señal de voz debe procesarse como si se tratase de un proceso estocástico.**
  - Con esta aproximación se considera, que inicialmente se dispone de un modelo de generación de la fuente/canal de voz que produce una cierta secuencia de palabras ***W***, es decir, la situación inversa a nuestro problema.

Definición de sistema reconocedor de voz (IV)

- Suponemos también, que este modelo se caracteriza por su falta de certeza y baja fiabilidad a la hora de convertir el conjunto de palabras anterior en la señal de voz ***S***.
- De esta manera, modelamos la conversión de ***W*** en la señal observada ***S*** como si de un canal ruidoso se tratara.

- **El reconocimiento se formula entonces como un problema de decodificación a posteriori en el cual debemos maximizar una determinada función objetivo [RAB96].**

Definición de sistema reconocedor de voz (V)

- Una forma de simplificar el problema es no trabajar con ***S*** directamente sino con una representación paramétrica de ***S***.
- La representación será una secuencia de vectores acústicos ***A***.

Definición de sistema reconocedor de voz (VI)

- **Si ahora empleamos la regla de Bayes para volver a formular el problema de decodificación, tenemos que:**

$$\arg\max_{W\in\Gamma} P(W \mid A) = \arg\max_{W\in\Gamma} P(A \mid W)\cdot P(W)$$

- donde Γ es el conjunto de todas las posibles secuencias de palabras, ***P(A\|W)*** es la probabilidad de la secuencia de vectores A, conocida la particular secuencia de palabras ***W*** y por último, ***P(W)*** es la probabilidad a priori de generar la secuencia ***W***.
- El primer término se conoce como el modelo acústico y al segundo ***P(W)*** como al modelo de lengua.

Definición de sistema reconocedor de voz (VII)

- Lo que pretendemos es modelar el sistema de producción de la voz como un canal ruidoso donde el ruido representa la variabilidad del hablante y del entorno.
- Teniendo en cuenta que no es posible en la práctica tener un conocimiento completo de las características de este canal, la aproximación estadística asume a menudo formas paramétricas particulares para $P_\theta(W|A)$ y $P_\omega(W)$, es decir, de acuerdo a modelos específicos.

Definición de sistema reconocedor de voz (VIII)

- Todos los parámetros de los modelos estadísticos, como por ejemplo θ y ω, necesarios para la evaluación de la probabilidad acústica $P_\theta(W|A)$ y la probabilidad de lengua $P_\omega(W)$, se estiman a partir de largas colecciones, llamadas conjuntos de entrenamiento.
  - Estas colecciones están formadas por un gran número de realizaciones de habla, producidas en general por el conjunto de hablantes representativo de los usuarios del sistema final y por grandes cantidades de texto escrito.

Definición de sistema reconocedor de voz (IX)

- **Este proceso se conoce como entrenamiento de modelos o aprendizaje de modelos y constituye la base que garantiza el éxito del proceso de reconocimiento.**

Aplicación de los sistemas de reconocimiento (I)

- **Basándonos en el modelo específico de tareas, podemos encontrar cinco grandes categorías de aplicación de los sistemas de reconocimiento:**
  - Telecomunicaciones. Proporciona información o acceso a datos y servicios sobre línea telefónica.
    - Dos de las aplicaciones más extendidas incluyen el sistema de automatización para manejo de llamadas del servicio asistido por operador de AT&T o VRCP (*Voice Recognition Call Processing*) y el NTT ANSER para la realización de determinadas operaciones bancarias por teléfono.

Aplicación de los sistemas de reconocimiento (II)

- Oficina/escritorio. Capacidades de reconocimiento del habla para control de las agendas telefónicas y funciones del teléfono, dictado y rellenado de formularios.
- Negocios y manufactura, permitiendo la incorporación de la voz en proceso de control de calidad en cadenas de montaje, empaquetado, ordenación y distribución.
- Área médica y legal. Generación de informes, documentos y formularios con presencia de abundante vocabulario específico.
- Otras aplicaciones como control domótico, ayuda a discapacitados, juegos, etc.

Tecnología	Tarea	Modo	Vocabulario	Tasa de error
Palabras aisladas	Palabras equiprobables	Dependiente del locutor	10 dígitos	0%
			39 caracteres alfanuméricos	4,5%
			1.109 palabras básicas del inglés	4,3%
		Independiente del locutor	10 dígitos	0,1%
			39 caracteres alfanuméricos	7,0%
			1.218 nombres de ciudades	4,7%
Palabras conectadas	Dígitos con longitud conocida	Dependiente del locutor	10 dígitos	0,1%
		Independiente del locutor	11 dígitos	0,2%
	Jerga de líneas aéreas (complejidad= 4)	Dependiente del locutor	129 términos	0,1%
Habla fluida	Resource Management (RM) (complejidad= 60)	Independiente del locutor	991 palabras	3,0%
	Airline Travel Information System (ATIS) (complejidad= 25)	Independiente del locutor	1.800 palabras	3,0%
	Wall Street Journal (WSJ) (complejidad= 145)	Independiente del locutor	20.000 palabras	12,0%

Tabla 1.Tasas de error por palabra para diversos sistemas de reconocimiento evaluados en laboratorio [RAB96].

Clasificación de los sistemas de reconocimiento de voz (I)

- El reconocimiento de la voz es un problema difícil de resolver, entre otras cosas porque existen muchas fuentes de variabilidad asociadas con la señal de entrada al sistema.
- Las realizaciones acústicas de los diferentes fonemas, los alófonos, dependen mucho del contexto en el que aparecen.

Clasificación de los sistemas de reconocimiento de voz (II)

- El problema es que pequeñas variaciones en estas realizaciones pueden dar lugar a fuertes cambios en el significado del mensaje:
  - Los cambios en las condiciones del entorno de trabajo así como la calidad y posición de los micrófonos pueden ser origen también de variaciones acústicas.
  - Las diferencias en el habla para un mismo locutor fruto de cambios en el estado de ánimo, ritmo de producción, o calidad de la voz en determinadas circunstancias como por ejemplo afecciones del aparato respiratorio.

Clasificación de los sistemas de reconocimiento de voz (III)

- Las diferencias entre el habla de diversos locutores. En este caso las fuentes pueden ser el género, las diferencias sociolingüísticas, el dialectalismo e incluso el tamaño y forma del tracto vocal.

- **Debido a todo esto, un sistema de reconocimiento de voz puede caracterizarse por un conjunto de parámetros diverso. Entre los más importantes podemos citar:**

Clasificación de los sistemas de reconocimiento de voz (IV)

Parámetros	Rango
Tipo de discurso	[Palabras aisladas, habla continua]
Dependencia del locutor	[Dependiente del locutor, independiente del locutor]
Tamaño del vocabulario	[Pequeño (< 20 palabras), Grande (> 20000 palabras)]
Estilo de discurso	[Lectura, habla espontanea]
Modelo de lengua	[Contexto explícito, sensible al contexto]
Confusión	[Pequeña (<10), grande (> 100)]
Relación señal ruido	[Alta (> 30 dB), baja (< 10 dB)]
Tipo de transductor	[Micrófono de gradiente, teléfono]

Tabla 2. Parámetros típicos empleados en la caracterización de un reconocedor de habla [COL97].

Clasificación de los sistemas de reconocimiento de voz (V)

- **El rango expresa los límites de complejidad que tiene el sistema para un parámetro dado.**
  - La parte izquierda indica el caso más simple y la parte derecha la situación más compleja que se puede presentar.
- **Un sistema que reconozca palabras aisladas requiere que el hablante realice una pequeña pausa entre palabras, mientras que un sistema de reconocimiento continuo o de palabras conectadas no.**

Clasificación de los sistemas de reconocimiento de voz (VI)

- **Algunos sistemas necesitan realizar un proceso de entrenamiento para cada locutor.**
  - El usuario debe proveer algunas muestras de su voz antes de poder usar el sistema (dependencia del locutor).
- **Aunque depende de la tarea, en general, cuanto mayor es el tamaño del vocabulario mayores son las dificultades para el reconocedor.**
  - Este aspecto está fuertemente interrelacionado con las similitudes fonéticas que presenta el conjunto de las palabras del vocabulario.

Clasificación de los sistemas de reconocimiento de voz (VII)

- El habla espontanea, por su parte contiene o puede contener cambios bruscos y acusados en la prosodia.
- Las realizaciones de voz pueden ajustarse a secuencias determinadas de palabras con un contexto explícito (gramática muy reducida), o por el contrario acercarse a realizaciones de lenguaje natural.

Clasificación de los sistemas de reconocimiento de voz (VIII)

- Una medida usual de la dificultad de la tarea a realizar por un sistema y que combina el tamaño del vocabulario y el modelo de lengua es el que se conoce como confusión (*complexity*).
  - Depende de la media geométrica del número de palabras que pueden seguir a una dada, una vez que se ha aplicado el modelo de lengua correspondiente.

Clasificación de los sistemas de reconocimiento de voz (IX)

- **Por último, encontramos algunos factores del entorno como son el ruido presente en el ambiente de trabajo y la calidad de los mecanismos de captura de la voz.**
  - Desde micrófonos específicos que realizan parte del filtrado del ruido ambiente, hasta micrófonos de características variables según el usuario (teléfono).

Aspectos principales que condicionan su construcción (I)

- **Forma de producción del habla y tamaño de los vocabularios.**
  - Habla aislada, habla conectada o habla continua.
  - Vocabularios pequeños (<100 palabras), medianos (≈3000 palabras) o grandes (>10000 palabras).
  - Tipo y estilo de habla.
- **Conjunto de usuarios.**
  - Dependiente de locutor.
  - Independiente de locutor.

Aspectos principales que condicionan su construcción (II)

- **Entorno de funcionamiento.**
  - Robustez frente al ruido.
  - Protección frente a sonidos no lingüísticos y palabras de fuera del vocabulario.

Reconocedores actuales. Estado de la cuestión

- **Los sistemas que permiten lenguaje natural son experimentales (sistemas de diálogo).**
- **No son robustos al ruido normal del ambiente de funcionamiento.**
- **Reconocen mejor si se limitan a un solo hablante que entrene el sistema.**
- **Funcionan mal si el hablante no coopera activamente.**

Estructura de un sistema de reconocimiento del habla

- La siguiente figura muestra el esquema completo con los componentes, tanto acústicos como lingüísticos, que debe poseer un sistema de reconocimiento de voz.
- No obstante en la práctica, encontramos una estructura mucho más simple (Figura 2).

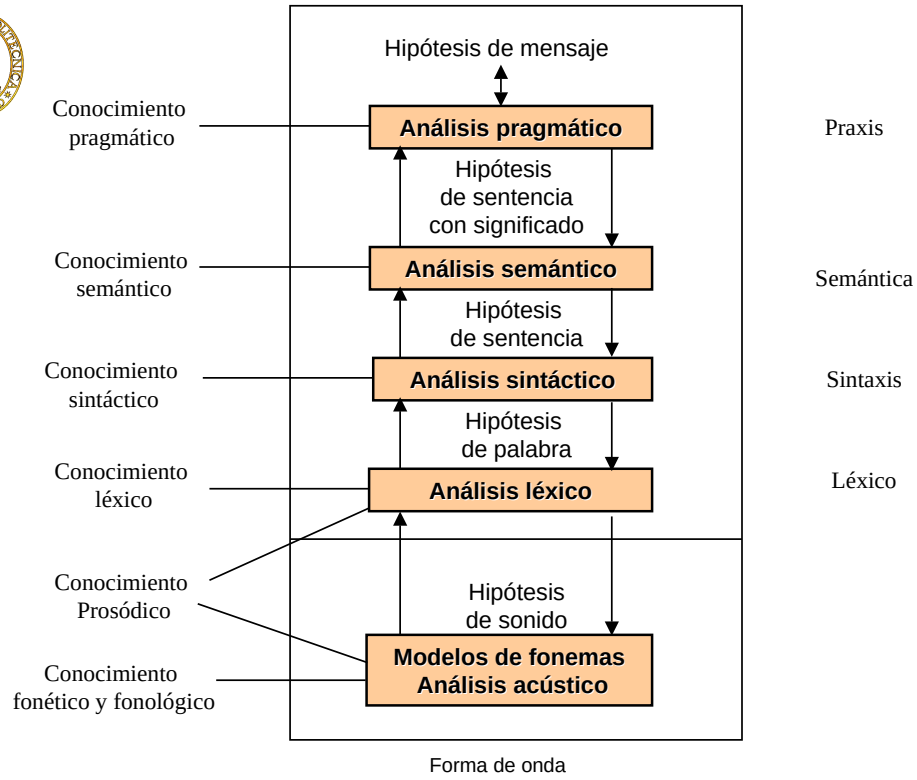

Figura 1. Esquema de un reconocedor con procesadores acústicos y lingüísticos [DEL93].

Esquema de un reconocedor de voz

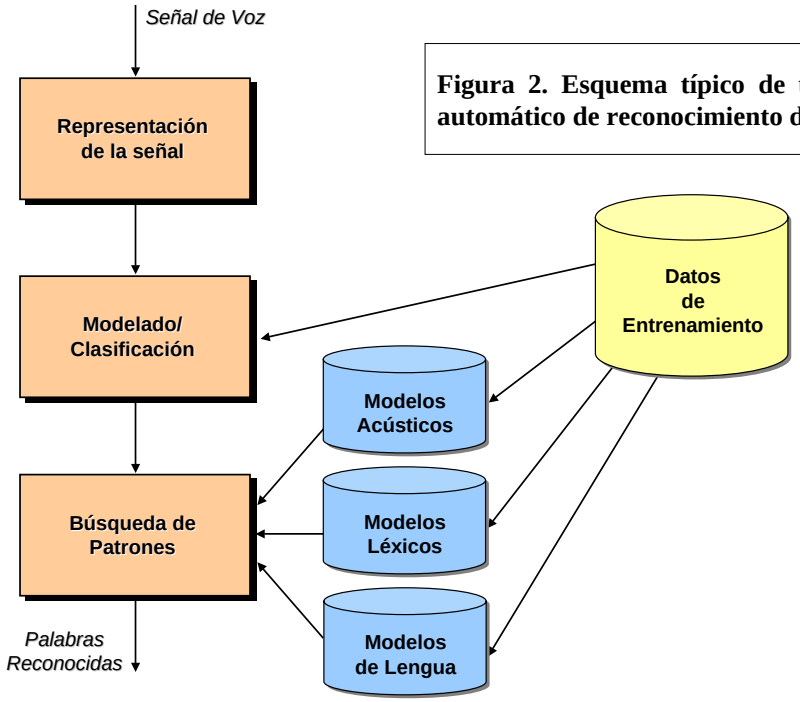

Figura 2. Esquema típico de un sistema automático de reconocimiento de voz.

Representación de la señal (I)

- **Tras obtener los datos de entrada al sistema por medio de la digitalización de la señal de voz, las muestras del paso anterior se transforman en un conjunto de medidas o características útiles.**
- **Esta es la fase que llamamos representación de la señal.**
  - La realización de esta tarea se lleva a cabo a un ritmo constante, usualmente de entre 10 y 20 ms.
  - Como producto de salida se obtienen una serie de vectores que de alguna manera contienen en esencia la señal original pero con un tamaño mucho menor.

Representación de la señal (II)

- Un vector de características deberá contener toda la información relevante para realizar el proceso de reconocimiento, eliminando el resto de contenidos redundantes o que no sean útiles para eñ sistema.
- Durante los momentos en los que no haya presencia de voz se deberá desactivar el proceso de extracción de rasgos.
  - Problema de la detección de principio y fin.

Búsqueda de patrones

- Posteriormente, las medidas de la etapa anterior se emplean para encontrar la palabra candidato que proporcione el mejor encaje.
- Para ello se tiene en cuenta las restricciones impuestas por los modelos acústico, léxico y de lengua (gramática) disponibles.
- Ésta es la fase búsqueda de patrones.

Modelado/Clasificación (I)

- **Los sistemas de reconocimiento intentan estimar y configurar las fuentes de variación descritas anteriormente de diversas formas en los que constituye la *fase de modelado/clasificación*.**
  - En el nivel de representación de la señal, los desarrolladores emplean representaciones que enfaticen las características importantes desde un punto de vista perceptual.
  - Por otra parte se intenta desenfatizar aquellas otras que son dependientes de determinados hablantes.

Modelado/Clasificación (II)

- **La adaptación a un determinado hablante o a las condiciones del entorno de funcionamiento del sistema se realiza a través de un conjunto de procesos de normalización.**
- **La normalización tiene como finalidad restaurar las características del vector, es decir, conseguir medidas con valores lo más próximos posibles a los que se obtendrían en condiciones neutras o a las presentes cuando se entrenó el reconocedor.**

Proceso de entrenamiento (I)

- **El entrenamiento, constituye una etapa previa a la entrada en funcionamiento del conjunto y tiene como finalidad establecer los diferentes conjuntos de modelos empleados durante el proceso de búsqueda:**
  - **Modelos acústicos**, que recojan realizaciones dependientes del género del locutor, entonaciones, variantes dialectales, etc.
  - **Modelos léxicos**, que contemplen pronunciaciones alternativas de las palabras, con objeto de permitir que los algoritmos de búsqueda encuentren diferentes caminos.

Proceso de entrenamiento (II)

  - **Modelos de lengua**, que permiten estimar la frecuencia de ocurrencia de determinadas secuencias de palabras.

- **Además durante el entrenamiento se puede obtener también la información necesaria para ajustar los procesos de *modelado* y *normalización*.**

Etapas de entrenamiento y puesta en marcha de un reconocedor

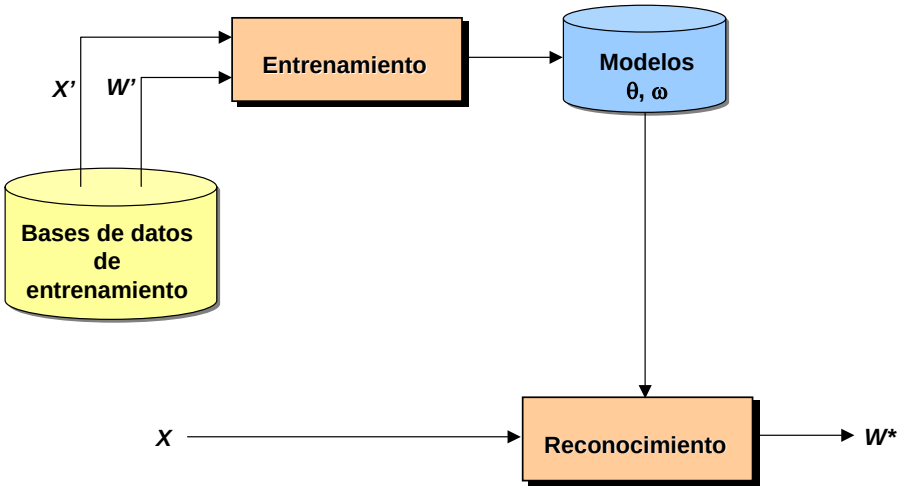

Figura 3. Esquema simplificado de las etapas de desarrollo de un reconocedor.

Bibliografía

[COL97] R. A. Cole et al., *Survey of he State of the Art in Human Language Technology*, Cambridge University Press, 1997.

[DEL93] J. R. Deller, J. G. Proakis and J. H. L. Hansen, Discrete-Time Processing of Speech Signals, Mac Millan, N. Y., 1993.

[RAB96] L. R. Rabiner, B. H. Juang and C. H. Lee, "An Overview of Automatic Speech Recognition", *Automatic Speech and Speaker Recognition: Advanced Topics*, C. H. Lee, F. K. Soong and K. K. Paliwal editores, Kluwer Academic Publisher, 1996, pp. 1-30.