1

El Reconocimiento de Formas

1.1 Fundamentos

Si bien es relativamente sencillo el conseguir que una máquina (p.e.: un ordenador) capte lo que le rodea (basta dotarle de una adecuada serie de sensores y/o transductores: células fotoeléctricas, micrófonos,...), no lo es tanto el conseguir que sea capaz de interpretarlo y/o reconocerlo.

Ciertamente, es casi trivial conseguir que la máquina "clasifique" de manera elemental una serie de medidas recogidas por un sensor (basta una tabla) e identifique de esta manera una situación, un objeto o una variación de su entorno (p.e.: bajada de temperatura) (figura 1.1). Y efectivamente, ello equivale a una tarea perceptiva simple.

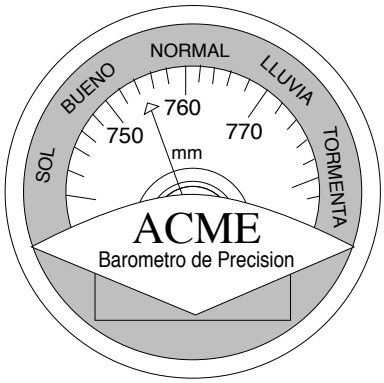

Figura 1.1 Una Máquina capaz de percibir (el estado del tiempo).

Pero el problema se presenta cuando la interpretación de lo que sucede en el entorno depende, no solo de un sensor, sino de varios; no solo de una medida de ese sensor, sino de muchas medidas recogidas a lo largo del tiempo. Y se complica mucho más cuando para poder comprender el entorno es necesario recurrir a una gran base de conocimientos

(experiencias) previamente almacenadas, y que debe ser consultada y comparada rápidamente con lo captado a través de los sensores. Con todo, la dificultad más grande aparece cuando se considera que las medidas de los sensores no son absolutamente fiables, y peor aún, que la descripción que se tiene del entorno tampoco lo es.

La complejidad involucrada es tal, que hoy en día, a pesar de largos estudios sobre la función perceptiva realizados por los psicólogos sobre personas y animales, y del ininterrumpido intento de duplicarla por parte de los técnicos (generalmente informáticos), la percepción, en su acepción más general, sigue siendo una tarea que, si bien todos los seres humanos llevamos a cabo rutinariamente, nadie sabe realmente cómo.

A pesar de todo ello, es enorme el interés que tiene el poder construir máquinas capaces de reconocer caracteres (leer libros), imágenes (p.e. identificar aviones), sonidos (p.e. entender cuando se les habla), etc..., por lo que nunca se ha cejado en el empeño. En la actualidad, principalmente gracias a la evolución de la informática, se han conseguido interesantes resultados prácticos, y lo que es más importante, grandes avances en la comprehensión de las dificultades y en la definición de los medios que deben ponerse en acción para superarlas. En el resto de este capítulo describimos (muy por encima) algunos de estos resultados, así como los medios que se utilizan para obtenerlos.

1.1.1 Las etapas de un reconocedor

La percepción de objetos por un sistema es el objetivo declarado de la disciplina del *Reconocimiento de Formas*, que se engloba en un conjunto de técnicas mucho más ambiciosas: la *Inteligencia Artificial* [Cohen,82]. El reconocimiento de formas, sin embargo, ha ido adquiriendo entidad de por sí mismo, hasta el punto de que en la actualidad constituye un campo de investigación que evoluciona con dinámica propia, a menudo independiente de la inteligencia artificial.

Reconocer un objeto consiste en asociarle (identificarlo con) un *mensaje semántico*, es decir un significado. En el caso más general, este mensaje semántico es simplemente un punto (también llamado *forma*) de un *universo semántico* (a veces llamado universo interno, en oposición al universo externo captado por los sensores). Normalmente, a una forma (p.e.: silla) le corresponden muchos posibles objetos (p.e.: todas las posibles sillas). Dicho de otro modo, una forma es un *conjunto de objetos* que se caracterizan por estar etiquetados por un mismo mensaje semántico.

Un sistema de reconocimiento de formas, en su versión más simple, estará constituído de la manera esquematizada en la figura 1.2, es decir, por dos módulos:

- Un módulo de *representación*, que obtiene, a partir de cada objeto, (usualmente) captado del mundo real mediante una serie de sensores, una representación conveniente para su utilización por el módulo de interpretación.

- Un módulo de *interpretación*, que proporciona, a partir del conjunto de formas, también representado convenientemente, y de la representación del objeto proporcionada por el módulo de representación, la forma a la que pertenece el objeto. Lleva a cabo pues un proceso de *comparación* ("pattern matching") entre el objeto y el conjunto de formas.

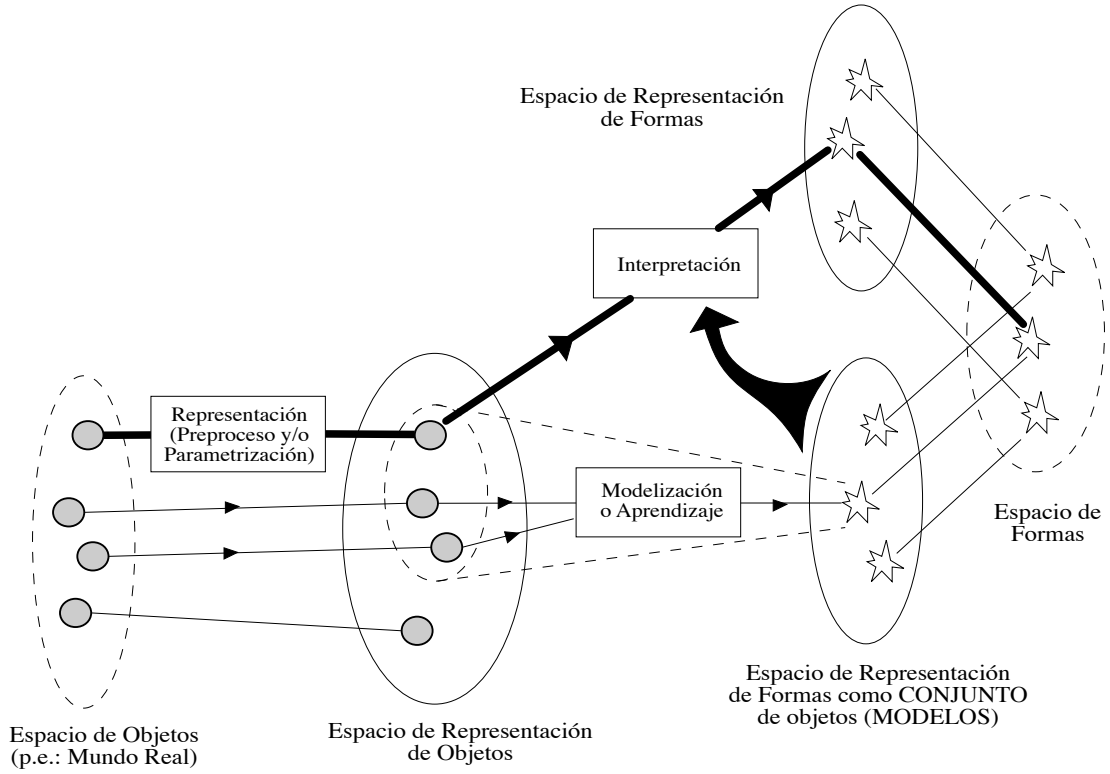

Figura 1.2 Composición de un reconocedor de formas. El trazo grueso muestra las sucesivas etapas de un proceso de reconocimiento.

El módulo de representación puede no ser necesario, si se da la circunstancia de que los objetos ya vienen dados de manera conveniente para el módulo de interpretación. En muchos otros casos (como los presentados en la parte experimental de este trabajo) el módulo de representación está compuesto por varias etapas, algunas de las cuales efectúan un *preproceso* y otras una *parametrización*:

- Las etapas de *preproceso* efectuan en general transformaciones que, o bien no cambian el dominio de representación, o bien llevan a subconjuntos de éste (p.e.: filtrado de la señal, supresión de grises por umbral...).

- Las etapas de *parametrización* cambian el dominio de representación y a menudo reducen drásticamente la cantidad de información, suprimiendo aquella que resulta redundante y/o inútil. En este último caso, se puede considerar la operación como equivalente a extraer las características más significativas del objeto, para representarlo por un conjunto de *parámetros* o *descriptores* (p.e.: transformada discreta de Fourier, cadena de símbolos,...).

El conjunto de formas utilizado por el módulo de interpretación representa el conocimiento que tiene el sistema de su entorno. Muy a menudo la representación del mismo es simplemente un conjunto de representaciones de formas (p.e.: un conjunto de gramáticas), pero puede estar representado por una estructura única y compleja (p.e.: una única red o gramática). El proceso de obtención de este conjunto de formas se conoce como *modelización* o *aprendizaje* del entorno (de sus objetos) a reconocer y es efectuado por el *módulo de aprendizaje*.

1.1.2 Representación de objetos y formas

La representación elegida para los objetos se deriva usualmente del método de interpretación escogido. Estas representaciones pueden clasificarse en dos grandes tipos:

- *No estructuradas*: El objeto se representa por un conjunto de (sub)objetos sin relación ellos (o con una relación arbitraria). Ejemplo típico es un vector o matriz en los que no importa cómo se distribuyen los componentes (p.e.: un punto en un espacio n-dimensional).

- *Estructuradas*: El objeto se representa mediante un conjunto de (sub)objetos más un conjunto de relaciones entre ellos (una *estructura*). Normalmente las relaciones son de contigüedad y/o sucesión: cadenas, árboles, grafos ...

A su vez, los subobjetos de que están formados los objetos pueden estar representados de la misma manera que los objetos (es decir, estar formados de otros subobjetos de nivel inferiror), o ser simplemente un conjunto de parámetros continuos (números reales) o cuantizados (numeros enteros). Cuando son parámetros cuantizados con pocos niveles de cuantización (del orden de 100) se puede asignar a cada valor un nombre o *símbolo*.

Las posibles representaciones de las formas son mucho más complejas: no deben representar un único objeto sino un *conjunto de ellos*. Las representaciones de un conjunto de objetos se conocen también como *modelos*. Dos de los casos más simples (y más corrientes) se engloban en lo que se conoce como el reconocimiento *geométrico* o *estadístico* de formas y el reconocimiento *sintáctico* de formas (ver apartado 1.1.4).

Por otra parte, con el fin de comunicar el resultado del reconocimiento al experimentador o a otro posible módulo, es necesaria una representación alternativa de la forma, pues la representación como conjunto de objetos es sólo útil para el módulo de interpretación. Esta segunda representación de cada forma será similar a las utilizadas para los objetos, puesto que la forma puede considerarse un objeto (punto) del universo semántico. Si el número de formas (la talla del universo semántico) es pequeño esta representación puede consistir simplemente en un único valor cuantizado o *etiqueta*. El reconocedor de formas se podrá entonces considerar como un *clasificador*. En casos más complejos, podrá recurrirse a una representación estructurada de las formas (p.e.: cadena). Se llamará *traductor* a un reconocedor en los que tanto los objetos como las formas en salida estén representados de manera estructurada.

1.1.3 Composición de reconocedores

La posibilidad que tiene un reconocedor de comunicar la forma reconocida a otro módulo lleva inmediatamente a pensar en la *composición de reconocedores*, es decir, a utilizar otro reconocedor para analizar (identificar o reconocer) con más detalle lo que han reconocido otros reconocedores. Este procedimiento es de hecho muy utilizado siempre que los subobjetos de un objeto son estructurados: uno o varios reconocedores de nivel inferior se dedican a extraer los subobjetos. Otro caso más complejo se da en los *reconocedores multinivel*, formados por varios reconocedores idénticos, cuyos objetos están representados en el mismo dominio que sus formas en salida, y que están interconectados de manera más o menos regular, a menudo organizados en capas (etapas o niveles). Típico ejemplo de ello son los perceptrones multicapa [Lippmann,87].

En un reconocedor compuesto, las formas intermedias, obtenidas por los reconocedores de las capas inferiores, constituyen los objetos de las capas superiores. Estos objetos son *objetos abstractos*, puesto que no tienen relación directa con lo captado por los sensores, siendo su definición arbitraria y particular a un reconocedor determinado. En determinadas situaciones, estos objetos abstractos corresponderán a abstracciones más o menos intuitivas (fonemas, sílabas, polígonos, etc...), pero éste no es el caso más usual, principalmente debido a lo difícil que resulta definir estas abstracciones.

1.1.4 Reconocimiento Geométrico y Reconocimiento Estructural

Dos son actualmente las maneras más usuales de representar las formas en su aspecto de conjunto de objetos: aquellas que recurren a una visión *geométrica* de lo que es un conjunto, considerándolo como una (sub)región

de un determinado espacio y aquellas que prefieren verlo como formado por elementos que cumplen ciertas reglas *estructurales*.

1.1.4.1 Los modelos geométricos

Estos modelos asumen que los objetos son "de una pieza" (están representados de forma NO estructurada). Cada objeto se representa por una matriz (generalmente de una o dos dimensiones) de números o símbolos (los *parámetros*) y la forma se representa mediante uno o varios *objetos patrón* y una *medida de disimilitud* (que puede ser una distancia, métrica o no) entre ellos y los objetos a reconocer [Duda,73]. Un ejemplo en reconocimiento del habla se puede encontrar en [Rulot,85].

Cuando el espacio n-dimensional en el que se hallan representados los objetos es continuo, y se escogen adecuadamente los parámetros, a menudo se puede conseguir que los puntos (objetos) pertenecientes a la misma forma se agrupen en regiones. Estas regiones definen entonces las formas, y sus fronteras con otras (*fronteras de decisión*) pueden entonces describirse mediante *funciones discriminantes* [Duda,73] (figura 1.3).

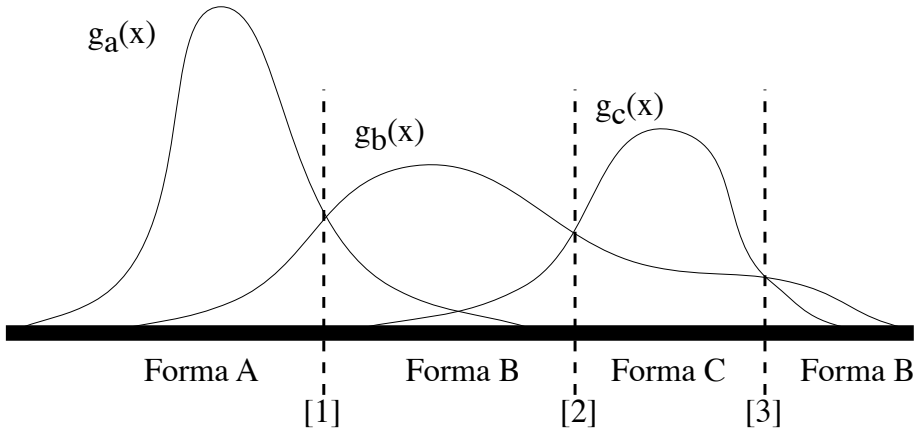

Figura 1.3 Formas (unidimensionales), sus funciones discriminantes (g(X)) y fronteras de decisión ([1][2][3])en un espacio unidimensional.

Ejemplo típico de función discriminante sería la inversa de la distancia a un centro (el patrón o *prototipo*). También lo sería la probabilida asociada a una distribución normal alrededor del mismo centro. En este último caso y en otros similares, se intenta aproximar una función discriminante mediante una función bien conocida, de la cual sólo resta determinar sus parámetros definitorios (aqui serían la media y desviación típica). Se habla entonces de reconocedor *paramétrico*. Muy a menudo estos modelos paramétricos son *estadísticos*, lo que justifica el nombre de *reconocimiento estadístico* de formas que también se le da al reconocimiento geométrico.

Otro ejemplo típico, en el que las funciones discriminantes son planos (o combinaciones de ellos), se encuentra en los clasificadores basados en redes neuronales tipo perceptrón [Duda,73] [Lippmann,87].

1.1.4.2 Los modelos estructurales

En estos modelos, se asume una representación estructurada de los objetos, estando usualmente los subobjetos representados de manera similar a la utilizada en los métodos globales. Las formas se representan mediante una serie de "reglas de composición" que deben cumplir los subobjetos para pertenecer al conjunto, existiendo un amplio abanico de posibilidades (árboles de decisión , expresiones lógicas, redes, modelos de Markov, reglas, gramáticas... [Hunt,75] [Gonzalez,78] [Fu,82] [Rabiner,83] [Cohen,82] [Valiant,84] [Quinlan,86] [Miclet,86]).

El reconocimiento debe basarse entonces, no sólo en identificar los subobjetos del objeto examinado aplicando para ello los métodos de reconocimiento geométrico de formas, sino también en analizar de alguna manera si la relación entre ellos es la buscada. Todo reconocedor estructural es por lo tanto, y por definición, un reconocedor compuesto, de por lo menos dos etapas: interpretación de subobjetos e interpretación de estructura.

Aquí el ejemplo serían los métodos gramaticales o sintácticos, en los que cada objeto es representado por un conjunto de símbolos cuya concatenación proporciona la estructura. Si esta estructura cumple una serie de reglas, especificadas mediante una gramática, es que pertenence a la forma representada por el lenguaje de esa gramática [Gonzalez,78] [Fu,82].

1.1.4.3 Métodos geométricos versus métodos estructurales

Los métodos geométricos de reconocimiento de formas son muy útiles para el reconocimiento de formas "simples", es decir, cuando se tiene un número pequeño de formas separables mediante funciones discriminantes no demasiado complejas. Estos métodos se utilizan por lo tanto muy frecuentemente para realizar clasificadores, han sido profusamente estudiados, y existe una gran variedad de algoritmos bien documentados y analizados que los utilizan y que funcionan perfectamente si los objetos tienen una distribución adecuada en el espacio de objetos [Duda,73]; lo cual desgraciadamente no es siempre cierto.

Por su parte, la aproximación estructural permite elaborar modelos más complejos, pues se dispone de una gran flexibilidad tanto para escoger el tipo de estructura y su configuración, como para elegir los subobjetos. Esta aproximación también permite introducir cómodamente conocimiento "a

priori", simplemente imponiendo restricciones a la estructura o difiniéndola "ad-hoc". Más importante aún, al contrario que en la aproximación geométrica, en la aproximación estructural es posible manejar espacios semánticos grandes e incluso infinitos (p.e.: para realizar traductores), así como obtener para cada objeto reconocido, no sólo la información de pertenencia a la forma, sino también una descripción en términos de subobjetos (la estructura) (p.e.: la secuencia de fonemas que forman la palabra reconocida).

En general pues, se emplea la aproximación estructural en los casos en que la aproximación geométrica es insuficiente, y en los que la información de estructura es de importancia relevante o fácil de utilizar (figura 1.4).

El que la aproximación estructural sea comparativamente más reciente que la geométrica, y el que su complejidad sea mayor, hacen que todavía quede mucho por hacer. En particular, una de las mayores dificultades, aún por resolver eficazmente, reside en la extracción de los subobjetos y su estructura a partir de un objeto complejo, operación que cae dentro del abierto problema de la *segmentación* (ver apartado siguiente). A pesar de ello, la ambición creciente de los diseñadores hace que la aproximación estructural se emplee cada vez más.

1.1.5 Segmentación

Uno de los problemas que más trabajos ha generado en el campo del reconocimiento de formas es el de la *segmentación*: la separación de los subobjetos de un objeto, o más específicamente, del fondo que lo rodea (en cuyo caso el problema es el de *detección de fronteras*).

La solución puede ser sencilla, cuando el objeto se halla sobre un fondo tiene una característica obvia que lo diferencia (mucha menor amplitud, color distinto, ...), o puede ser extremadamente difícil, cuando el objeto se mezcla con otros objetos, o forma parte integrante de un objeto más complejo (la pata de una silla, un fonema en una palabra) (figura 1.5). Ello explica el que generalmente, y en aras de la simplicidad, los sistemas no muy ambiciosos de reconocimiento de formas procuren evitar la segmentación. La mayoría de los experimentos que se encuentran en la literatura tratan de reconocer objetos *aislados* (en un fondo muy caracterizable), o como mucho objetos sumergidos en un ruido que aumenta su distorsión, pero no los mezcla con otros objetos. Incluso en el caso de reconocedores estructurales es posible obviar la segmentación utilizando como subobjetos particiones "naturales" de los parámetros que definen el objeto total (p.e., en reconocimiento de la palabra: un grupo de parámetros corresponde a un intervalo de tiempo, el siguiente grupo al siguiente intervalo, la estructura global es la secuencia de dichos grupos).

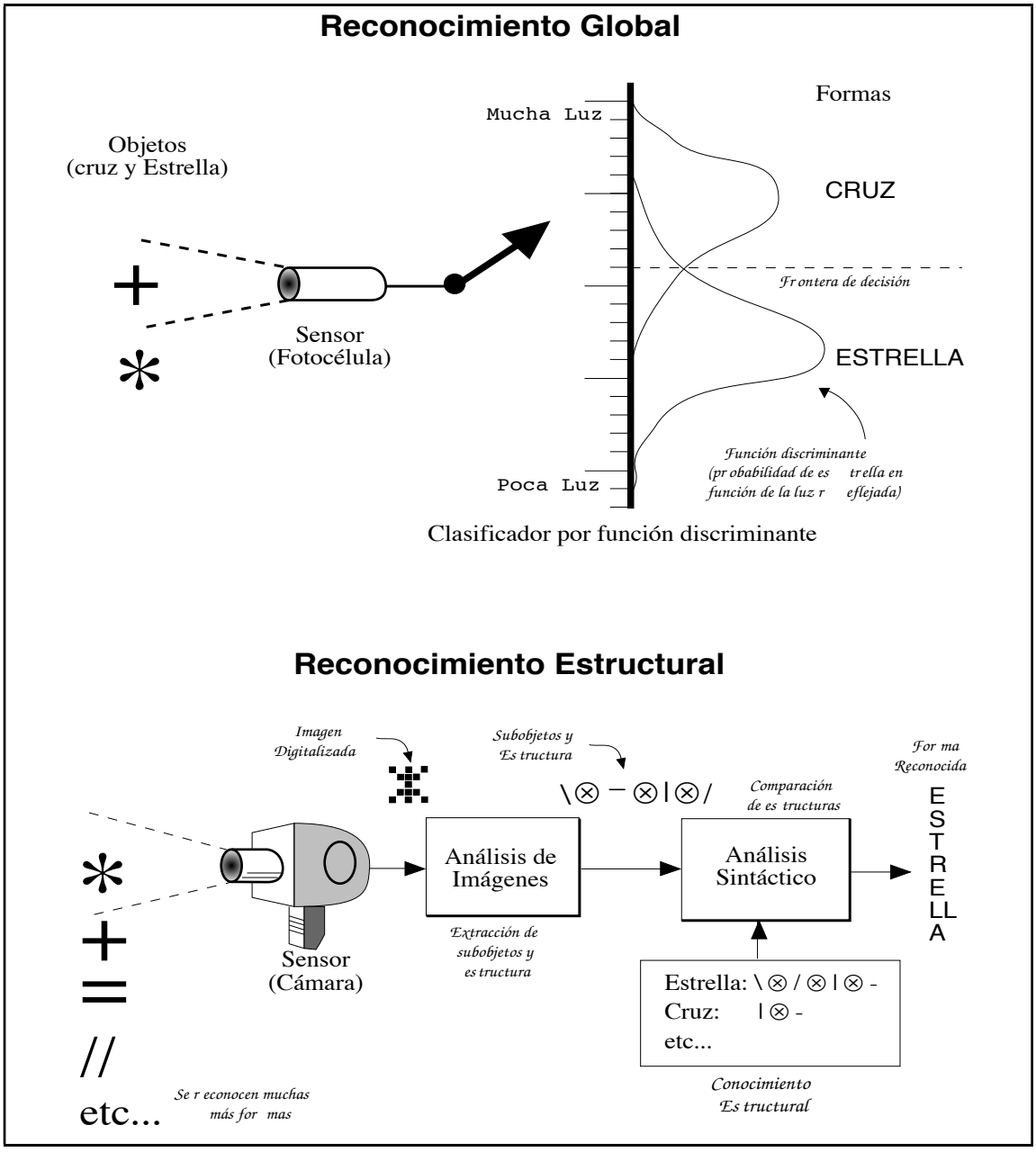

Figura 1.4 Comparación entre un método de reconocimiento global y otro sintáctico para la misma tarea. Obsérvese la diferencia de complejidad y potencia.

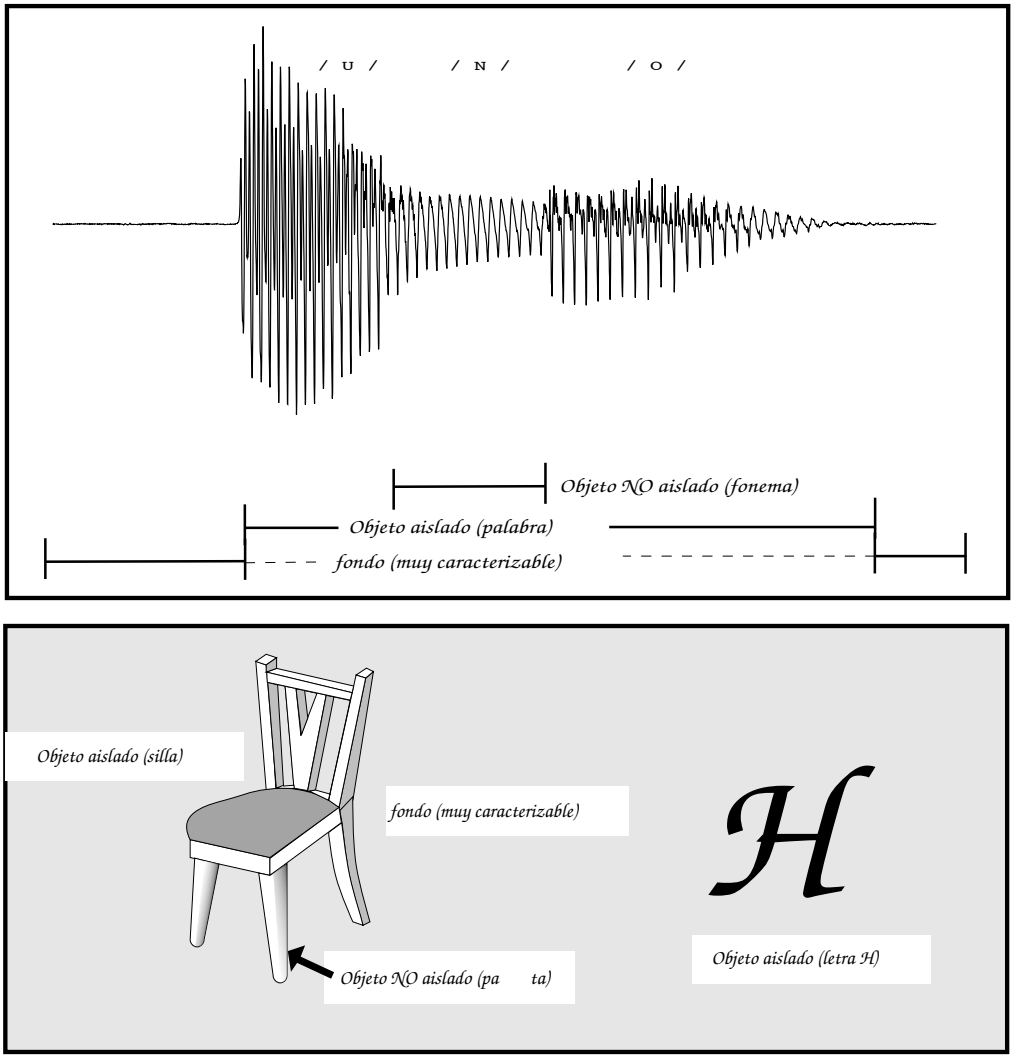

Figura 1.5 Formas aisladas y no aisladas. Segmentación.

Sólo una vez superado (en primera aproximación) el reconocimiento de objetos aislados, es cuando se aborda la segmentación. Esta representa aún un escollo importante para el logro de sistemas que operen en el mundo real (p.e.: el reconocimiento de la palabra continua, de la escritura enlazada), y aunque se puede demostrar que una solución automática (óptima o subóptima) es posible, los algoritmos conocidos hoy en día los siguen siendo muy costosos [Vidal,90].

En este trabajo sólo se tratará con objetos aislados, aunque no se descarta una extensión a objetos compuestos.

1.1.6 El tamaño de las subestructuras

Entre las muchas características relevantes de un objeto a la hora de reconocerlo, se encuentra una de especial interés para este trabajo: el tamaño (longitud) de las subestructuras.

El tamaño relativo (espacial o temporal) de los componentes de un objeto suele ser de carácter discriminativo, siendo por lo tanto a menudo parte integral de la descripción de una forma: no es posible distinguir un objeto de otro sin tener en cuenta lo grande o pequeño de las subformas que los componen.

De entre los muchos ejemplos posibles podemos mencionar: la acentuación de una sílaba (que a menudo sólo se percibe por su mayor duración), que distingue una palabra de otra; la duración de otra sílaba, que no cambia el significado de la palabra; la longitud de un brazo en la imagen de un cromosoma, que lo diferencia de otro; la longitud de un trazo en la imagen de una letra, que hace que sea esa y no otra... (figura 1.6).

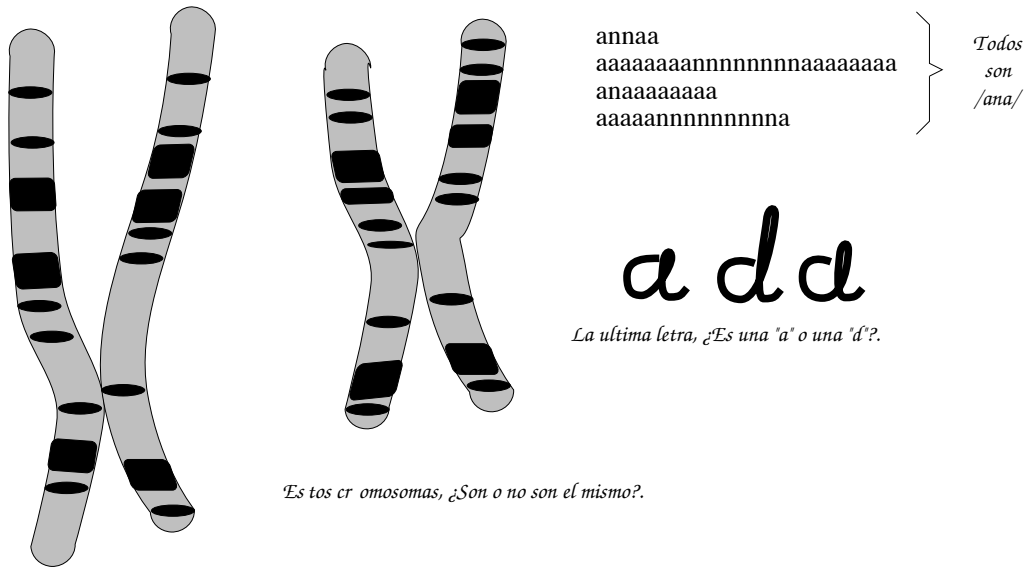

Figura 1.6 La importancia del tamaño para el reconocimiento de formas.

A pesar de la importancia de esta característica, a menudo no se la ha tenido explícitamente en cuenta a la hora de estudiar y diseñar una determinada representación de objetos o formas, lo que conduce a que muchos modelos ampliamente utilizados hoy en día tengan una fuerte disfunción en este campo. Actualmente se tiende a introducir de manera más o menos explícita la representación de la longitud de las subestructuras, ya sea complicando modelos ya existentes o, como es el caso en el presente trabajo, desarrollando nuevos modelos que tengan implícita a priori dicha representación [Rulot,87].

1.1.7 El Aprendizaje Automático

Sea cual sea el modelo escogido para representar las formas, será necesario proporcionar el conjunto de éstas al sistema antes de que sea posible utilizar el módulo de interpretación. Desgraciadamente, en general **no se conoce** la descripción exacta de una forma en los términos del modelo elegido (y normalmente en ningún otro), lo que obliga a dotar al sistema de la capacidad de *aprender* (inferir) [Angluin,83] (el conjunto de) las formas.

1.1.7.1 Taxonomía del aprendizaje

El proceso de aprender una forma se puede llevar a cabo mediante instrucción directa por parte del "maestro", es decir mediante transferencia sin más del procedimiento para reconocerla (p.e: mediante programación, introducción de reglas). Ello se conoce como *aprendizaje deductivo*. Por el contrario, en el *aprendizaje inductivo*, el sistema debe llevar a cabo algún proceso de abstracción (generar formas) por sí mismo. Ello hace imprescindible la utilización de ejemplos, que el sistema analiza y clasifica sin ayuda (aprendizaje *no supervisado*) o con ayuda del maestro (aprendizaje *supervisado*) (figura 1.7).

En el caso del aprendizaje supervisado *activo o informante*, el sistema estudia los ejemplos y sugiere otros nuevos; el maestro le dice cómo clasificarlos. Si el aprendizaje es *pasivo o textual* el sistema no genera nuevos ejemplos y la labor del maestro se limita a clasificar los ejemplos iniciales. Los ejemplos en el aprendizaje supervisado pueden ser sólo *positivos* (son de esa forma) o positivos y *negativos* (no son de esa forma).

En general en el aprendizaje inductivo las únicas abstracciones que se le piden al sistema se refieren a la generación de nuevas formas. El resto de la información (representación más adecuada, tipo de estructura, método de identificación apropiado, etc...) se le proporciona de manera deductiva. Nada impide sin embargo que el sistema sea capaz de abstraer (generar hipótesis) en este campo también, aunque usualmente no es necesario (ni fácil).

1.1.7.2 Factibilidad del aprendizaje Inductivo

El método usual de aprendizaje inductivo es la *presentación de ejemplos*. Se debe disponer de una serie de objetos del mundo real, etiquetados o no con la forma a la que pertenecen. Estos objetos se muestran al sistema, y éste, mediante un proceso de *inducción*, debe inferir la descripción de la forma acorde con el modelo de representación e identificación que esté empleando.

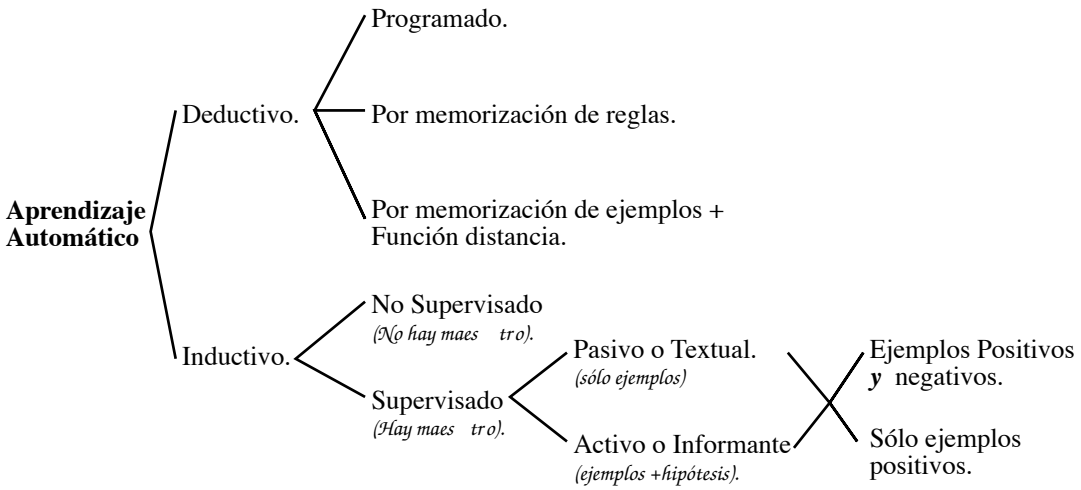

Figura 1.7 Taxonomía del aprendizaje.

Es evidente que la *convergencia* de este proceso (¿el sistema terminará por aprender al cabo de un número suficiente de ejemplos?) está lejos de ser obvia, y depende fuertemente del tipo de modelo de representación e identificación utilizado. Determinados resultados teóricos [Gold,67] afirman incluso que en el caso más general la convergencia es indemostrable.

En efecto, considerado formalmente, el aprendizaje o *inferencia* inductivos es un proceso de obtención de *reglas* a partir de *ejemplos*. De la manera más general posible, una regla es una *función recursiva parcial* **f** y un ejemplo es un par (**x**,**f**(**x**)), donde **x** pertenece al dominio de **f**. Un algoritmo de inferencia sintetiza el algoritmo de **f** a partir de sus ejemplos. Gold afirma que el inferir una función recursiva en general es un problema indecidible (!!). Afortunadamente el mismo Gold demuestra que si nos limitamos a las funciones recursivas *primitivas* sí que se las puede identificar en el límite a partir de ejemplos.

1.2 El Reconocimiento Automático del Habla

Como dice su nombre, el "Reconocimiento Automático del Habla" (RAH) pretende conseguir que las máquinas (usualmente ordenadores) sean capaces, tanto de obedecer órdenes expresadas oralmente, como de dialogar interactivamente con seres humanos mediante el uso de la palabra hablada.

El problema de la *síntesis del habla*, implícito en esta última y mayor pretensión, se puede dar hoy en día como resuelto, aunque en la práctica queden por pulir algunos "detalles" (calidad de la voz, entonación, ...) para conseguir una voz indistinguible de la de un ser humano [Casacuberta,87].

En RAH sin embargo, el panorama no es tan esperanzador. Las dificultades son debidas a una larga lista de factores, los cuales en su mayor parte dan cuenta de la mayor o menor variabilidad presente en los objetos a reconocer, y por lo tanto de la cantidad de información que necesitará el sistema para poder conseguir resultados aceptables. Entre estos factores cabe destacar:

- El *vocabulario*: número de palabras y/o frases a reconocer.
- La cantidad de locutores, utilizándose los términos: *Monolocutor* (el sistema reconoce con un único locutor), *Multilocutor* (sólo funciona con los locutores que conoce) e *Independiente del Locutor* (reconoce con cualquier locutor).
- El *sexo* de los locutores: como es de común conocimiento, la voz femenina es distinta de la masculina: no sólo cambia el tono fundamental de un sexo a otro, sino que la variación de formantes no es proporcional a la variación de dicho fundamental.
- La *calidad* de la señal vocal: la señal está afectada por *ruido*, procedente del ambiente (fábrica, cabina de avión,...) o de una mala transmisión (líneas telefónicas).

Los sistemas de RAH se pueden subdividir actualmente en dos grupos bien diferenciados por la complejidad del problema que intentan resolver: los que se dedican únicamente al reconocimiento de *palabras aisladas* (separadas unas de otras por silencios) y los que se dedican al reconocimiento de la *habla continua* (frases más o menos largas sin separación entre palabras).

En el habla continua, la cantidad de combinaciones de palabras (número de frases) posibles es extremadamente elevada, y es infactible el disponer de muestras de todas las combinaciones para realizar un aprendizaje fiable de cada una. Algo similar ocurre con el aprendizaje de modelos de palabras aisladas cuando el vocabulario crece. Todo ello obliga a recurrir a la descomposición en subunidades subléxicas (fonemas, palabras,...) y a utilizar algún método de representación estructural, que limite el número de posibilidades (p.e.: el número de palabras que pueden venir a continuación de una dada: la *perplejidad* del lenguaje). Como anteriormente se ha mencionado, esto lleva inevitablemente a enfrentarse con el problema de la segmentación, agravado en este caso por la *coarticulación* que existe entre las distintas unidades, que altera a éstas fuertemente en los puntos de unión.

Es sin embargo en estos sistemas en los que se centra actualmente la investigación, pues sólo mediante ellos se podrá resolver el problema del RAH.

1.2.1 Adquisición, parametrización y etiquetado en RAH

En reconocimiento del habla, los objetos son señales sonoras (ondas de presión en el aire) que representan palabras o frases. Es pues necesario utilizar una etapa de representación, que transforme estas señales en algo más conveniente para su utilización por el módulo de interpretación. En la actualidad, la etapa de representación de objetos en reconocimiento del habla está típicamente compuesta por tres subniveles: el preproceso, la parametrización y el etiquetado (que puede considerarse un nivel superior de parametrización) (figura 1.8).

- El primer subnivel, el del *preproceso* está formado por el conjunto mecánico - eléctrico - electrónico constituído por el micrófono (que transforma la onda sonora de presión en onda eléctrica), el filtro (que suprime componentes indeseables de la onda eléctrica) y el conversor analógico/digital (AD) (que transforma la onda eléctrica en una serie de medidas de amplitud).

- El segundo subnivel, el de *parametrización*, tiene como objetivo típico el de reducir la enorme cantidad de información proveniente del nivel anterior (≈120000 bits/sg.) en algo más manejable (≈5000 bits/sg.) .

  La parametrización efectúa usualmente un cambio de espacio de representación, pasando de un espacio de una dimensión (tiempo) a otro de dos dimensiones (típicamente tiempo/frecuencia), siendo por lo tanto una transformación de tipo estrictamente matemático, que transforma una serie de medidas de amplitud en una serie de vectores de parámetros [Casacuberta,87].

  El tipo de parametrización varía de un sistema a otro, los más utilizados son: el banco de filtros, los coeficientes de predicción lineal, y los coeficientes cepstrales [Makhoul,75] [Rabiner,78] [Benedí,89]. El autor, en un trabajo anterior, propuso y estudió para este fin la los valores de la función de autocorrelación de la señal muestreada a un bit [Rulot,85].

- El tercer subnivel, el *etiquetado* o *cuantificación vectorial*, no siempre se halla presente, pero es muy utilizado en los sistemas que utilizan la aproximación estructural en RAH, puesto que permite una reducción aún mayor de la cantidad de información (≈300 bits/sg.) y proporciona una representación extremadamente adecuada para aplicar los métodos estructurales (gramáticas, Modelos de markov,...).

  El etiquetado se lleva a cabo normalmente mediante algún tipo de análisis estadístico que permite clasificar los vectores de parámetros del nivel anterior en una serie reducida de clases, cuyos nombres o

símbolos son los que se utilizan en lugar del vector original [Duda,73] [Gray,84].

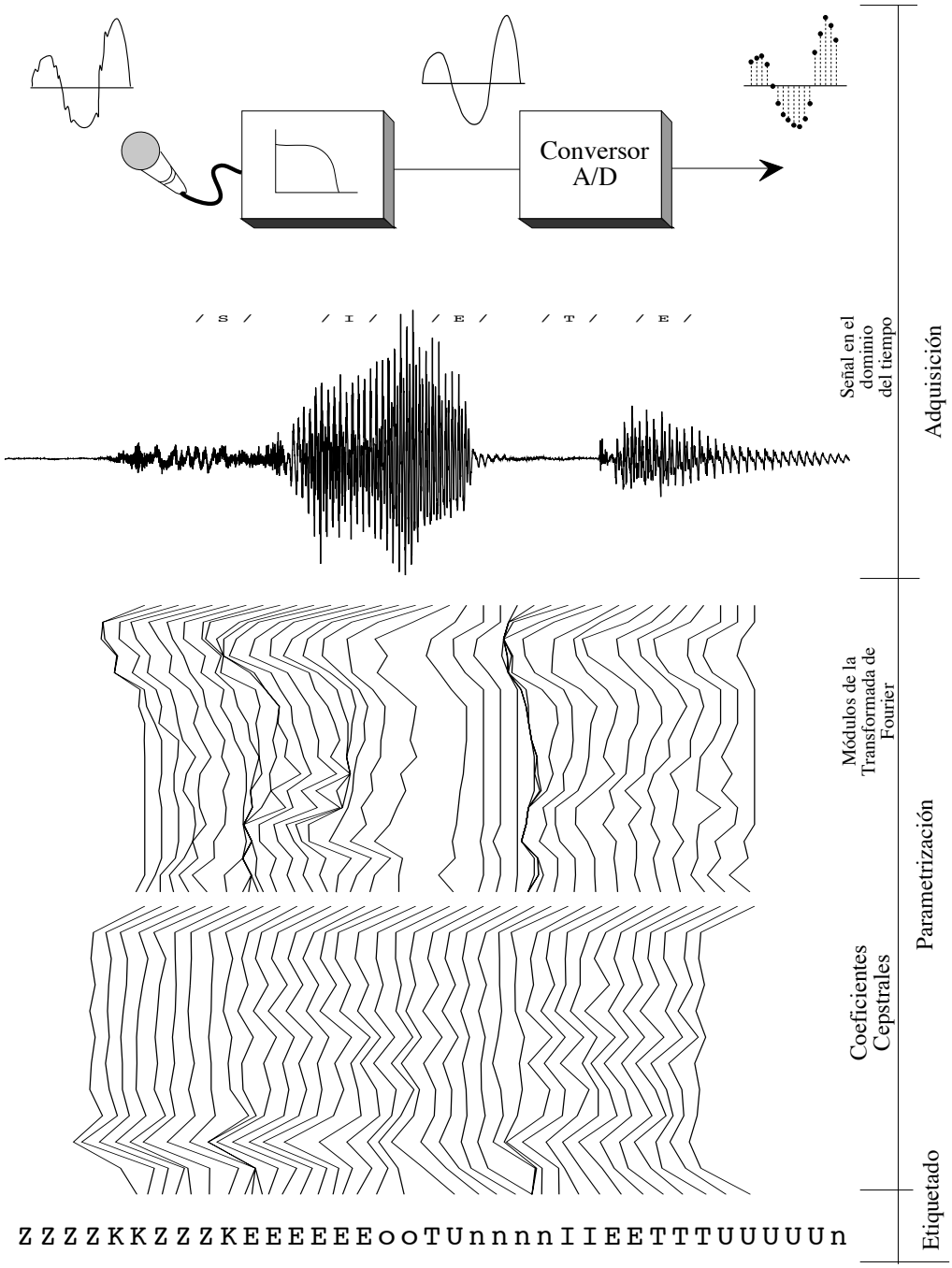

Figura 1.8 Adquisición, parametrización y etiquetado en RAH.

1.2.2 Estado del Arte

Actualmente existen sistemas capaces de reconocer con porcentajes de aciertos superiores al **98%** vocabularios *sencillos* (palabras bien distintas fonéticamente) y *pequeños* (del orden del centenar de palabras) de palabras aisladas, todo ello en entornos independientes del locutor y con locutores de ambos sexos. Estos resultados se obtienen incluso en ambientes ruidosos [Rabiner,87] [Loeb,87] [Watanabe,88], pero aunque son suficientes para determinadas aplicaciones puntuales, no permiten una comunicación hombre-máquina realmente fiable. En el estado actual de la técnica, conseguir más del **99%** de aciertos es posible únicamente imponiendo condiciones más restrictivas, por ejemplo suprimiendo el ruido (este trabajo) o restringiéndose a locutores de un único sexo [Lippmann,87].

La mayoría de estos sistemas se basan en métodos globales, aunque la tendencia actual es utilizar un modelo estructural proveniente de la teoría de la información: los *HMM* ("Hidden Markov Models": Modelos Ocultos de Markov) [Rabiner,83]. Empleando estos modelos, con fonemas o unidades subfonéticas como subformas, es posible manejar grandes vocabularios y obtener resultados esperanzadores. Por ejemplo, [Jouvet,86] con 1000 palabras obtiene hasta un **95%** de reconocimiento mientras que [D'Orta,87] con 3000 palabras consigue hasta un **84%** pero con un sistema dependiente del locutor.

Como se observa, la tasa de reconocimiento baja rápidamente en función de la complejidad de la tarea y se halla aún muy lejos de la máxima alcanzable, dada por la obtenida por seres humanos. Por ejemplo, considérese un vocabulario especialmente difícil: el *e-set*, formado por las letras que se pronuncian con /e/ (/be/, /ce/, /de/, /ge/, /pe/,...). Con él, los seres humanos consiguen un 98% de palabras reconocidas [Bahl,87]. Los sistemas actuales, a pesar de ser sólo 9 palabras, consiguen en el mejor de los casos, un **92%** [Bahl,87] (multilocutor). Otro ejemplo, en el que la dificultad se debe a la enormidad del vocabulario, lo constituye el *Tangora* [Averbuch,87] [Cerf-Danon,91]. Este sistema, con un vocabulario de 20000 palabras, reconoce en promedio un **97%** de las mismas, eso sí, es un sistema dependiente del locutor, considera solo palabras aisladas y restringe fuertemente la variabilidad del lenguaje mediante estadísticas lingüisticas. Los resultados de Tangora no varían sensiblemente cuando se le aplica a distintas lenguas europeas (ha sido probado en inglés, francés, alemán, italiano y español).

En el reconocimiento del habla continua, los resultados de reconocimiento a nivel de palabra se reducen drásticamente, haciéndose imposible obtener más de un 80% de reconocimiento si no se imponen restricciones lingüísticas. En este campo, los participantes en el proyecto *SPICOS* han desarrollado un sistema de reconocimiento que con un vocabulario de 1000 palabras reconoce un 75% de ellas a partir de frases, y ello independientemente del locutor [Ney,91]. El mismo sistema, cuando se

le impone un modelo de lenguaje que limita la perplejidad media a 100 obtiene incluso un 91% de tasa de reconocimiento de palabras. Por su parte, *Dragon Systems*, con un sistema que se adapta al locutor y con vocabularios de 5000 palabras de perplejidad media 140, obtiene un 94.2% de reconocimientos de palabras [Baker,91]; mientras que con otro vocabulario de 3400 palabras, pero de perplejidad mucho mayor (430), logra un más de un 85%. La misma compañía afirma que tras una adaptación de con alrededor de 2000 palabras, se reconocen textos formados a partir de un vocabulario de 25000 palabras con un porcentaje de aciertos de palabras del orden de 87%. Experimentos multilengua con este sistema (se han probado las mismas del *Tangora* más el holandés) han proporcionado resultados medios de 85%, con muy poca variación de una lengua a otra, en un vocabulario de 2000 a 4000 palabras [Bamberg,91]. También de gran interés es el sistema SPHINX [Lee,88], que con un vocabulario de 1000 palabras obtiene un 70,6% de aciertos en palabras a partir de frases (independientemente del locutor), consiguiendo incluso hasta un 96% si, con el mismo vocabulario, se restrige la perplejidad (20) mediante un modelo de lenguaje.

1.3 El Reconocimiento Automático de Caracteres

El reconocimiento de imágenes puede decirse que tuvo su comienzo en 1870, cuando Carey inventó el Scanner Retina, un sistema de transmisión de imágenes que usaba un mosaico de fotocélulas. Sin embargo, el verdadero impulso se dió con el invento del Scanner Secuencial por Nipkow, que más tarde daría lugar a la televisión, y por la aparición de los ordenadores al final de la década de los 40.

De entre los múltiples sistemas que desde entonces han tratado imágenes para su reconocimiento (fotos de satélites, imágenes de entorno para los sistemas de visión, piezas a clasificar, etc...), pronto destacaron los *reconocedores ópticos de caracteres* (OCR), por la cantidad de aplicaciones prácticas inmediatas que permitían vislumbar (ayuda a ciegos, comprobación de firmas, pero sobre todo ofimática y correo) y su relativa sencillez (imágenes planas, sin sombras, número limitado de formas).

No obstante los primeros logros en reconocimiento de caracteres los consiguiera Tyurin en 1900, y hubieran otros intentos memorables como el Optófono de Fourier d'Albe (1912) y el sistema táctil de Thomas (1926), las aplicaciones comerciales del reconocimiento de caracteres sólo tuvieron lugar a mediados de los 40, con el desarrollo de los primeros reconocedores de imágenes y la aparición de pioneros como David Shepard, el fundador de "Intelligent Machine Research Co." [Mantas,86].

Las posibles tareas a las que se puede asignar un OCR se pueden clasificar [Mantas,86] (figura 1.9), por orden de dificultad, según el reconocimiento sea de caracteres...

- **Impresos** (*Fixed Font* y *Multi Font* Character Recognition (CR)).
- **Trazados** (en tableta gráfica o similar), en los que además de la imagen del carácter se dispone de la información temporal de su trazado (*On-Line* CR).
- **Manuscritos**, caracteres también escritos a mano, pero separados y no caligráficos (*Handwritten* CR).

Por otra parte, si los caracteres no están aislados, sino que se encadenan unos con otros para formar palabras, se habla de reconocimiento de **Escritura** (*Script* CR). Aquí la dificultad es máxima, al presentarse en toda su magnitud el problema de la segmentación.

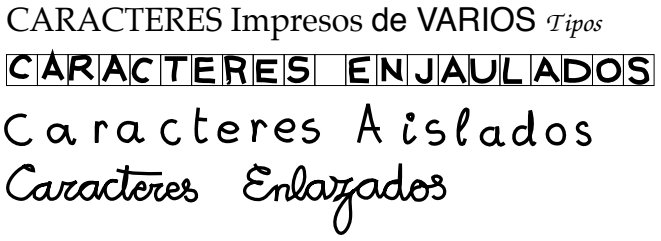

Figura 1.9 Distintos grados de dificultad en reconocimiento de caracteres

Aparte de la manera como están escritos, otras posibles fuentes de dificultad en el reconocimiento de caracteres, si nos restringimos a alfabetos occidentales (ni que decir tiene que el problema de los símbolos chinos, árabes, etc... presenta complicaciones adicionales), son debidas a lo parecido de ciertas letras y/o dígitos (U-V, C-L, a-d, n-h, I-1, Z-2, S-5, G-6), que incluso en algunos casos pueden escribirse igual (O-0, l-1) o casi indistinguiblemente de sus correspondientes en mayúsculas (O-o, K-k, C-c), lo que a menudo obliga a recurrir al contexto o incluso a su posición con respecto a la línea base (P-p, Y-y) para diferenciarlas.

1.3.1 Adquisición, reducción de ruido y parametrización en RAC

Ya que trata directamente con objetos del mundo real, en RAC también es necesario obtener una representación del objeto externo a reconocer, que sea adecuada para el módulo de interpretación. En el caso de RAC, el módulo de representación comporta usualmente 3 subniveles: *preproceso*, *reducción de ruido* (que en realidad es un preproceso de nivel superior) y *parametrización*.

- El nivel de *preproceso* consta normalmente (en reconocimiento óptico) de un dispositivo mecánico-electrónico (una cámara o un scanner óptico) que *barre* la imagen horizontal y verticalmente con

el fin de transformarla en una serie de medidas de intensidad luminosa; y de un digitalizador, que transforma esas medidas (expresadas normalmente en voltajes eléctricos) en números binarios que transfiere al ordenador.

Alternativamente (reconocimiento de caracteres trazados), lo que se obtiene y posteriormente se digitaliza son las coordenadas sucesivas de la posición del lápiz, junto con información de si éste toca la superficie o no (tableta gráfica o pantalla sensible).

- Durante la etapa de *reducción de ruido*, se procesa la matriz bidimensional obtenida en la etapa anterior mediante variados algoritmos de tratamiento de imágenes, con el fin de suprimir y/o realzar determinadas características: suavizado, realze de bordes, realze de contrastes, supresión de grises mediante umbral, submuestreo, supresión de puntos aislados, normalización de tamaño, alineamiento con línea base, normalización de orientación, esqueletización, etc... En todos los casos este proceso, fundamentalmente matemático, proporciona una matriz (imagen) similar a la inicial pero a menudo con mucha menos información.

  Muchos de estos tratamientos son innecesarios cuando se tiene la secuencia de posiciones obtenida de un carácter trazado [Tappert,90], aunque de igual manera se aplican suavizados, filtrados normalizaciones, etc...

- La *parametrización* permite a continuación obtener una descripción adecuada de la imagen en función de lo que se quiere reconocer y reducir aún más la cantidad de información. En el caso de los caracteres puede prescindirse de este paso (se reconoce simplemente la imagen mediante procedimientos globales), o bien pueden utilizarse transformaciones globales (de Karnhunen-Loeve, de Fourier, cálculo de momentos,...). También pueden extraerse propiedades locales (puntos extremos, ángulos, uniones en T y cruces,...) y/o geométricas (segmentos, curvaturas,...), o pueden buscarse representaciones estructurales (conversión a un grafo, a una secuencia de direcciones, descripción topológica, descripción en base a un conjunto de figuras elementales,...).

1.3.2 Estado del Arte

Existen en la actualidad gran cantidad de programas de OCR comercializados. Están principalmente destinados al reconocimiento de caracteres impresos (Multi-Font CR), y funcionan en su mayoría en ordenadores personales tipo PC-IBM Compatible o Macintosh. Su tasa de reconocimiento normalmente se halla entre **80**% y **95**%, obteniendo desde luego los mejores resultados cuando funcionan con tipos de letra para los

que han sido "afinados" o entrenados [Robinson,90]. Estos sistemas incluyen a menudo "reconocedores de composición", siendo capaces de separar columnas y bloques de texto y de distinguir a éstos de las figuras.

La literatura científica hoy en día se halla más centrada en el reconocimiento de caracteres manuscritos, que en cierta manera puede considerarse un "superconjunto" del de los caracteres impresos, y que desde luego involucra una mucho mayor dificultad dada la mucha mayor variabilidad (¿quién no se ha tropezado con una letra "de médico"?). En este campo, y restringiéndose al reconocimiento de caracteres aislados, se obtienen actualmente unas tasas de reconocimiento de **98.3** a **99**% en dígitos aislados [Kurosawa,86] [Shridar,86] [Baptista,88].

En condiciones poco favorables (enorme número de escritores, condiciones de escritura incontroladas) como es el caso cuando se quieren reconocer los códigos postales (ZIP codes), los resultados obviamente empeoran, lográndose un promedio de 92% de reconocimiento, variando entre **85**% y **97**% [Lam,88] [Nadal,90] [Kimura,91]. Lo mismo ocurre si se intenta reconocer un conjunto de formas mayor que el de los dígitos, como las 26 letras inglesas: **88**% [Brown,88] o el kanji: **86.7**% (con símbolos difíciles) [Sekita,88].

Cuando se intenta reconocer escritura *enlazada*, se puede, a pesar de todo, conseguir resultados muy buenos aprovechando la información contextual que supone la palabra: **92,5**% de aciertos en palabras [Kundu,89], o limitando el reconocimiento a caracteres trazados: **89.9**% en caracteres (letras inglesas mayúsculas y minúsculas) [Fuyisaki,90].