VOLUMEN MONOGRÁFICO (1999), 83-102

SISTEMAS DE RECONOCIMIENTO DE VOZ EN LAS TELECOMUNICACIONES

DANIEL TAPIAS MERINO
Telefónica Investigación y Desarrollo

RESUMEN. El *objetivo de este artículo es presentar el estado actual de la tecnología de reconocimiento de voz y los problemas asociados a la utilización de estas tecnologías en servicios reales ofrecidos a través de la línea telefónica. Para ello se hará una clasificación de los reconocedores de voz, se expondrán resumidamente sus principios de funcionamiento y se describirán y justificarán sus limitaciones actuales. Posteriormente se introducirá el concepto de sistema conversacional, que es la base de la nueva generación de productos de tecnología del habla, y se describirá el sistema conversacional ATOS (Automatic Telephone Operator Service) desarrollado en Telefónica Investigación y Desarrollo, S.A. (TI+D). Por último, se realizará una comparación de los sistemas más avanzados del mundo y se presentarán las tendencias actuales en esta área.*

PALABRAS CLAVE. *Procesamiento del habla, reconocimiento de voz, sistema conversacional.*

ABSTRACT. *This paper presents the state-of-the-art technology of voice recognition and the problems linked to its use through telephone line. Voice recognition devices will be classified according to their operation principles and their shortcomings described and accounted for. We then introduce the concept of conversational system, the basis of the new generation of speech technologies, and describe the conversational system ATOS (Automatic Telephone Operator Service), developed in Telefónica Investigación y Desarrollo, S.A. (TI+D). Finally, the most advanced systems in the world are compared and current trends introduced.*

KEYWORDS. *Speech processing, speech recognition, conversational system.*

1. INTRODUCCIÓN

En la actualidad, la capacidad de las máquinas para percibir su entorno es muy limitada, por lo que la interacción automática con el medio no es trivial salvo que las condiciones estén muy controladas y las señales sean fáciles de interpretar.

Sin embargo, la asombrosa facilidad y rapidez con que los ordenadores realizan operaciones matemáticas o resuelven algunas cuestiones tediosas como las de ordenación y búsqueda de información, hace que las expectativas sobre la capacidad de los mismos para la resolución de todo tipo de problemas hayan superado y superen, con mucho, a la realidad. Este hecho es responsable de un optimismo exagerado sobre la capacidad de los ordenadores que ha llevado, en muchos casos, a un análisis simplista y, por tanto, a la subestimación de problemas complejos como el reconocimiento de voz, la inteligencia artificial, etc.

En general, se podría decir que los ordenadores actuales automatizan satisfactoriamente muchas tareas que, para una persona, son difíciles, repetitivas o que requieren mucho tiempo, mientras que se muestran bastante ineficaces para resolver tareas aparentemente sencillas como leer caracteres escritos a mano, reconocer voz o identificar una imagen. Esto se debe, por ejemplo, a que para hacer reconocimiento de voz, hay que pasar de las tareas relativamente sencillas de detección o medición de señales, ordenación y búsqueda de información y realización de cálculos matemáticos a la labor de *interpretación* de los datos, que requiere de procesos complejos de razonamiento, de capacidad de aprendizaje y de bases de conocimiento.

Una dificultad añadida a la resolución de este tipo de problemas es que no se conocen los mecanismos que nos permiten percibir el entorno, a pesar de que la percepción es algo que experimentamos todos los seres vivos.

Si bien todo lo dicho en los párrafos anteriores es cierto, también lo es que los avances realizados en los más de cuarenta y cinco años de investigación en reconocimiento de voz junto con los avances en el campo de la informática, han hecho posible la resolución de muchas cuestiones que hace tan solo unos años pertenecían al mundo de la ciencia ficción. Este hecho ha desencadenado la proliferación de productos y servicios nuevos basados en tecnologías del habla que, si bien tienen aún muchas limitaciones, han alcanzado ya la madurez suficiente como para poderse emplear en múltiples aplicaciones.

En este artículo se intentará dar una visión general del estado actual de la tecnología de reconocimiento de voz particularizando para el caso de reconocimiento por línea telefónica. Se hará énfasis en los problemas que están por resolver y que, en este momento, son un reto no sólo para la división de Tecnología del Habla de TI+D sino también para el resto de los centros de investigación públicos y privados del mundo. Asimismo, se introducirá el concepto de sistema conversacional, se hablará de procesamiento del lenguaje natural y se presentará el sistema conversacional ATOS (Automatic Telephone Operator Service) desarrollado en TI+D. Por último, se compararán los sistemas más avanzados del mundo y se expondrán las tendencias de los principales centros de investigación.

2. OBJETIVO DEL RECONOCIMIENTO DE VOZ

El objetivo del reconocimiento de voz es convertir, fielmente, la voz en texto. Esto es: escribir lo que una persona esté diciendo sin cometer errores. Las limitaciones tecnológicas actuales impiden que los reconocedores alcancen totalmente este objetivo, por lo que siempre tienen asociada una tasa de error que se obtiene sumando tres conceptos:

- Sustituciones: palabras que han sido reconocidas erróneamente ("***p***asa" en lugar de "***t***asa"),
- Inserciones o sobregeneraciones: palabras que se han reconocido de más ("***el*** coche" en lugar de "coche"),
- Elisiones o infrageneraciones: palabras que se han reconocido de menos ("__ coche" en lugar de "***el*** coche").

Dado que siempre existe una tasa de error asociada a los reconocedores de voz, es importante que los servicios que utilizan sistemas de reconocimiento tengan mecanismos de recuperación frente a errores. Estos mecanismos dependerán del tipo de error que se haya producido. Por ejemplo, si el error afecta al significado de la frase, solo una estrategia eficiente y bien diseñada de gestión del diálogo permitirá que el sistema sea consciente del error y que tome la iniciativa para resolverlo por medio del diálogo correspondiente. Sin embargo los fallos que no afectan al significado de la frase pueden obviarse con relativa facilidad empleando analizadores semánticos que sean poco sensibles a errores de concordancia en género, número y/o persona, infrageneración o sobregeneración de nexos sintácticos (p. ej.: "a", "que"), etc.

3. DIAGRAMA DE BLOQUES DE UN RECONOCEDOR DE VOZ

En la figura 1 se presenta el esquema típico de un reconocedor de voz, al que se le han añadido los bloques de análisis semántico y de gestión del diálogo. He incluido estos dos últimos bloques porque mejoran las tasas de reconocimiento gracias a la información que tienen sobre el estado del diálogo, por lo que se puede considerar que forman parte del proceso de reconocimiento. Vamos a ilustrar esto con un ejemplo: supongamos que el usuario pronuncia una frase del tipo "*querría hacer una llamada*". El analizador semántico extraerá el significado de la frase, que en este caso se reduce a la acción que desea realizar el usuario ("*llamar*") y el gestor del diálogo pedirá al mismo que diga el número de teléfono o el nombre de persona o empresa a la que desea llamar. En estas circunstancias, lo lógico es que el usuario conteste con un número de teléfono o un nombre, por lo que si el gestor de diálogo proporciona esta información al reconocedor, se reduce la probabilidad de que éste falle.

En el proceso de reconocimiento se emplean cuatro tipos de información: los modelos acústicos, que permiten que el reconocedor identifique los sonidos pues proporcionan información sobre las propiedades y características de los mismos. El diccionario, que indica qué conjunto de sonidos forma cada palabra del vocabulario. El modelo del lenguaje, que tiene información de cómo se deben combinar las palabras para formar frases y, por último, en los sistemas de diálogo o conversacionales, el reconocedor suele disponer de predicciones sobre el contenido de la siguiente frase que pronunciará el locutor.

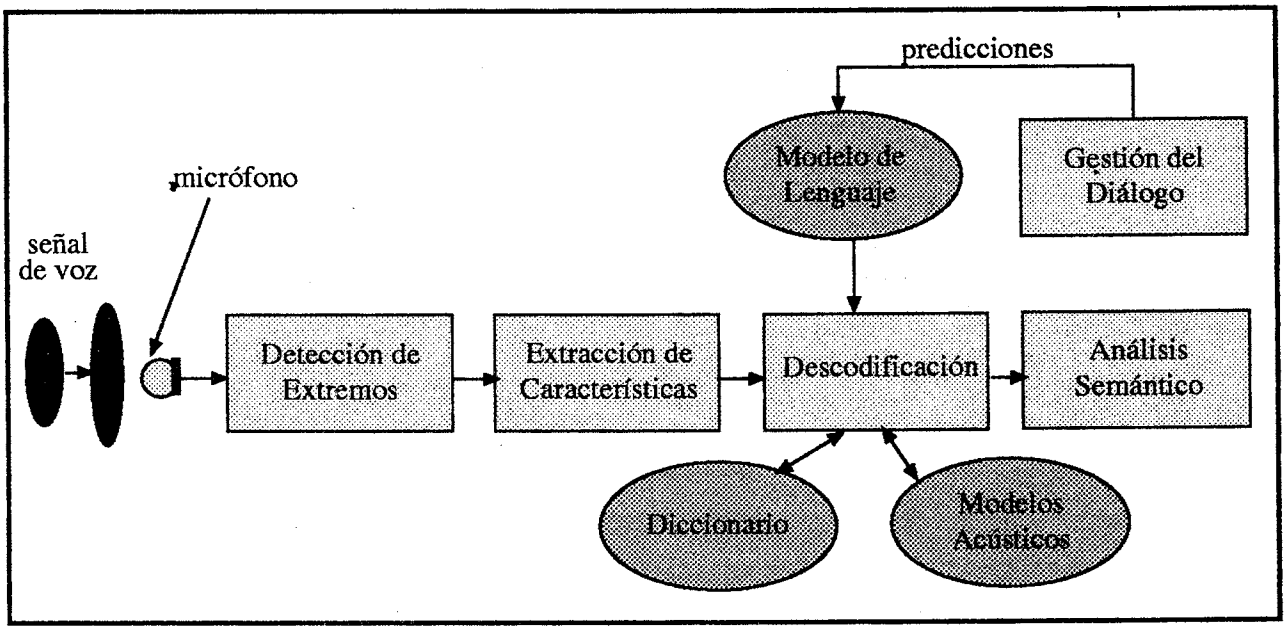

Figura 1. *Diagrama*

Así pues, el funcionamiento del reconocedor completo es el siguiente: La señal de voz entra por el micrófono y se convierte en una señal eléctrica analógica que es posteriormente digitalizada (convertida en secuencias de unos y ceros). Esta señal pasa al detector de extremos, que es el encargado de detectar la presencia de voz y de pasar dicha voz al siguiente bloque del reconocedor. El extractor de características calcula una serie de parámetros de la señal de voz que tienen información relevante para el proceso de reconocimiento. Estos parámetros se pasan al descodificador, el cual se apoya en los modelos acústicos, los modelos del lenguaje, las predicciones y el diccionario para generar la frase reconocida. Posteriormente, el analizador semántico extraerá el significado de la frase, que será utilizado por el gestor del diálogo para, en función del estado de la conversación, tomar la decisión más adecuada y hacer una predicción sobre la siguiente interacción con el usuario.

Como el problema general del reconocimiento de voz no está totalmente resuelto, existen muchos tipos de reconocedores especializados en resolver problemas concretos. Por este motivo, no se pueden comparar dos reconocedores si no están

especializados en la misma tarea y, aún en este caso, habrá que asegurar que las condiciones de la prueba son idénticas para ambos, antes de pronunciarse sobre la calidad de cada uno de ellos: comparar dos reconocedores sin tener en cuenta su especialización es como comparar un coche de carreras con uno familiar. No hay uno mejor que otro, simplemente son distintos. Como éste es un punto importante que da lugar a mucha confusión, a continuación se presenta una clasificación de los sistemas de reconocimiento atendiendo a varios criterios:

- *Según el número de locutores que pueden reconocer:*
  - Dependientes del locutor: Sólo reconocen a la persona para la que han sido entrenados.
  - Multilocutor: reconocen a un conjunto pequeño de personas.
  - Independientes del locutor: reconocen a cualquier persona.

- *Según el tamaño del vocabulario que reconocen:*
  - Reconocedores de vocabularios pequeños: hasta 40 palabras.
  - Reconocedores de vocabularios medios: hasta 400 palabras.
  - Reconocedores de vocabularios grandes: hasta 4.000 palabras.
  - Reconocedores de vocabularios muy grandes: hasta 40.000 palabras.
  - Reconocedores de vocabularios ilimitados: más de 40.000 palabras.

- *Según el canal:*
  - Reconocedores a través de micrófono.
  - Reconocedores para la red telefónica (fija, móvil analógica o móvil digital).

- *Según el tiempo de respuesta:*
  - Reconocedores de tiempo real: son reconocedores que dan la respuesta lo suficientemente deprisa como para que un usuario pueda interaccionar con ellos.
  - Resto: reconocedores en los que el tiempo de respuesta no es un factor importante (por ejemplo, sistemas de reconocimiento empleados para la transcripción de informes).

4. PROBLEMAS ASOCIADOS AL RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Sin duda alguna, el principal problema asociado a la utilización de técnicas de reconocimiento en servicios reales por línea telefónica es la propia limitación de la tecnología: si las máquinas hicieran reconocimiento de voz con tasas de error similares a las de los seres humanos, las dificultades serían menores y rotundamente

distintas a las que existen en la actualidad. Desafortunadamente, la tecnología no ha evolucionado lo suficiente y hay todavía muchas áreas en las que los centros de investigación están empezando a trabajar, por lo que habrá que esperar bastantes años antes de que todos los problemas estén totalmente resueltos.

Sin embargo, ya se puede sacar provecho de la tecnología existente, si se tienen en cuenta dos factores muy importantes a la hora de diseñar un servicio:

- *Un servicio nunca funciona a la primera* aunque se tenga una larga experiencia en el desarrollo de los mismos: El proceso de diseño y realización de un servicio es iterativo y como tal debe considerarse a la hora de su planificación. Cada aplicación tiene sus propias peculiaridades, ya sea por el tipo de usuarios al que va dirigido, por las propias particularidades de la aplicación o por ambas. Siempre hay que tener presente el factor humano, ya que el comportamiento de los usuarios es difícilmente previsible. Es siempre aconsejable hacer unas pruebas de campo con el prototipo y, a partir de ellas, hacer las modificaciones oportunas para adaptar el sistema a las necesidades del usuario y corregir los posibles errores de diálogo, de temporización, etc. Por ejemplo: si el reconocedor que ponemos en servicio sólo reconoce números, el diálogo deberá dirigir al usuario para que sólo diga números, porque en caso contrario el sistema fallará estrepitosamente.
- *Los usuarios no comprenden las limitaciones de la tecnología actual:* Los errores de reconocimiento causan sentimientos de frustración en el usuario, que no puede aceptar que una máquina no entienda algo que ha dicho correctamente y que él ha entendido perfectamente. Estos sentimientos de frustración desembocan en un rechazo a los servicios que emplean estas tecnologías. Por esto es muy importante valorar cómo afectan las limitaciones de la tecnología al tipo de aplicación en el que se van a utilizar antes de decidirse a su utilización.

A continuación voy a presentar los problemas que están todavía sin resolver y que constituyen un reto para los centros de investigación de todo el mundo.

4.1. *Adaptación al locutor*

Es asombrosa la facilidad con que los seres humanos percibimos diferencias en las voces de distintas personas. Esta capacidad nos permite identificar a una persona por su voz o incluso, con un poco de entrenamiento, conocer datos de la persona como la región en la que vive habitualmente o en la que aprendió a hablar, su nivel cultural, su sexo, su edad, rasgos de su forma de ser, su estado de ánimo, etc.

Por tanto, el conjunto de sonidos emitidos al hablar no sólo lleva la información del mensaje contenido en la frase pronunciada, sino también información sobre el interlocutor.

Toda esta información complementaria al propio mensaje, lejos de dificultar la comprensión, nos ayuda a mejorar el proceso de comunicación. Sin embargo, en el caso de los sistemas de reconocimiento, las diferencias entre distintas voces tienen efectos negativos en la tasa de error. En la figura se ilustra este hecho, y se muestran las distintas tasas de error de palabra (TEP) y de acierto de palabra (TAP) para diez locutores distintos tras emplear la primera versión del reconocedor del sistema conversacional ATOS en junio de 1996.

El reconocedor fue evaluado, sin que los usuarios supieran el vocabulario que el sistema era capaz de reconocer, empleando frases de habla natural (sin restricciones) para solicitar los casi 30 servicios distintos que proporcionaba esta primera versión del ATOS. El tamaño del vocabulario que manejaba el sistema era de unas 2.000 palabras. Véase cómo la tasa de error para el locutor "I" es prácticamente despreciable mientras que la tasa de error del locutor "G" es del 25%, por lo que mientras el locutor "I" estará muy satisfecho con el servicio, el segundo (con un error por cada cuatro palabras) tendrá la sensación de que el sistema funciona mal.

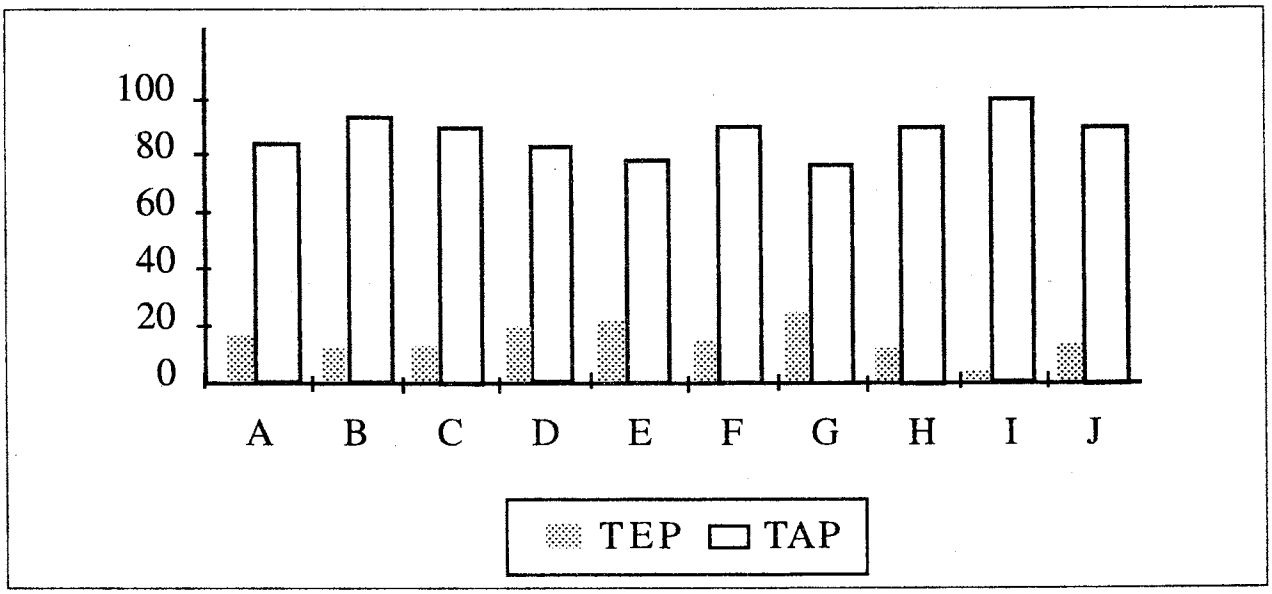

Figura 2. *Tasas de acierto (TAP) y de error (TEP) por locutor*

Las diferencias en las tasas de error se deben a que los modelos acústicos que maneja un reconocedor para poder realizar el proceso de reconocimiento se obtienen, en el proceso de entrenamiento, a partir de un conjunto finito de voces de muchas personas, de modo que:

- El sistema no funcionará bien con aquellas voces que sean distintas de las empleadas en el proceso de entrenamiento.
- Los modelos acústicos modelan una voz "promedio" resultado de procesar las voces del conjunto empleado para entrenar el sistema, por lo que tampoco las

voces parecidas a las de dicho conjunto se reconocen con las tasas de error que se tendrían si el sistema hubiera sido entrenado para una sola persona.

Adicionalmente, en reconocimiento de habla continua, el modelo del lenguaje y el diccionario afectan también a las tasas de reconocimiento: si la persona que utiliza el sistema pronuncia frases que están bien modeladas por el modelo del lenguaje y cuyas palabras están incluidas en el diccionario, el reconocedor funcionará mejor que en el caso contrario.

Por tanto, es necesario dotar a los reconocedores de técnicas de adaptación al locutor que permitan que el sistema pueda modificar dinámicamente los modelos acústicos, el diccionario y el modelo del lenguaje a fin de que la tasa de error sea la mínima posible e igual para todos los posibles usuarios.

El problema de la adaptación al locutor no es en absoluto trivial dado que hay múltiples causas por las que dos voces son distintas y la forma en que cada una de estas causas afecta a la señal de voz y, por tanto, a sus parámetros no es fácil de predecir de una forma determinista en la mayoría de los casos. Igualmente, el problema de adaptar el modelo del lenguaje y de modificar dinámica y automáticamente el diccionario es complejo.

A continuación se presenta una clasificación de las causas de la variabilidad de la voz que sin pretender ser completa ni totalmente precisa, sí da una idea de la complejidad del problema con que nos enfrentamos. Las causas de la variabilidad se han dividido en tres grandes grupos: diferencias culturales, fisiológicas y del entorno:

4.1.1. *Diferencias culturales*

Se clasifican como diferencias culturales a todas aquellas que han sido aprendidas por el individuo. En este sentido, aparecen las siguientes:

- Amplitud de los sonidos (volumen).
- Conjunto de sonidos empleado.
- Duración de los sonidos.
- Entonación.
- Forma de construir las frases (¿Comiste? en lugar de ¿Has comido?).
- Velocidad del habla.
- Vocabulario empleado, que está directamente relacionado con el nivel cultural, el dialecto y el contexto (filosofía, medicina, ingeniería, arte, deportes, etc.)

Cabe destacar que, en algunas ocasiones, hay ciertos sonidos pronunciados por un locutor que pueden ayudar a predecir la forma en la que va a pronunciar otras palabras. Por ejemplo, en la zona de Madrid es típico pronunciar "es que" como /e j k e/, por lo que si un locutor emplea esa pronunciación, hará lo mismo con otras

palabras en que el sonido /s/ preceda al sonido /k/: "escapar" se pronunciará como /e j k a p a r/.

La combinación de los factores anteriores da lugar a distintos estilos del habla para un mismo locutor (Llisterri 1992, Eskénazi 1993):

- Claro.
- Articulado.
- Formal.
- Desenfadado.
- Espontáneo.
- Culto.
- Vulgar.
- Leído.
- Dictado.
- Íntimo.
- Dialecto/acento (Alvar 1996a, 1996b).

4.1.2. *Diferencias fisiológicas*

En este apartado se sitúan las diferencias inherentes a la constitución física del locutor o a su estado físico o de salud:

- Cansancio.
- Congestión nasal.
- Forma y tamaño del tracto vocal.
- Frecuencia fundamental de las cuerdas vocales.
- Forma del pulso glotal: Hay diferencias importantes entre el pulso glotal de las mujeres y el de los hombres (Junqua y Haton 1996).

Así, por ejemplo, la voz suele ser más lenta y forzada cuando el locutor está cansado. El tracto vocal de las mujeres al ser, en general, más corto que el de los hombres genera formantes más separados y de frecuencias más altas. La congestión nasal afecta a la pronunciación de las nasales, etc.

4.1.3. *Diferencias en el entorno*

El entorno en el que está inmerso el locutor afecta de forma definitiva a las características de la voz:

- El ruido de fondo hace que el esfuerzo realizado por el aparato fonador sea mayor, lo que modifica el proceso de producción de la voz. Este fenómeno se conoce con el nombre de efecto Lombard.
- Los factores mecánicos como las vibraciones o aceleraciones.

- Los estados emocionales del individuo (miedo, ira, enfado, sorpresa, nerviosismo, estado de ánimo, etc.)
- Tipo de entorno o de canal desde el punto de vista acústico: reverberante, anecoico, distorsionador, etc.

4.2. *Resistencia frente al entorno*

Tanto en los experimentos publicados por Lippmann (1997) como en los realizados recientemente en Telefónica I+D, se demuestra que el ruido de fondo afecta relativamente poco a las tasas de reconocimiento de los seres humanos y, por tanto, al proceso de comunicación. Esto se debe fundamentalmente a tres factores:

- Las personas tenemos dos oídos, lo que nos permite la identificación de las fuentes de sonido y su separación, gracias al procesado realizado posteriormente en el cerebro.
- La capacidad de predicción del cerebro, apoyándose en una serie de fuentes de conocimiento como el propio conocimiento del lenguaje, de la persona o personas con las que se está hablando, el contexto y tema de la conversación, etc.
- La capacidad de adaptación al ruido y de cancelación de éste.

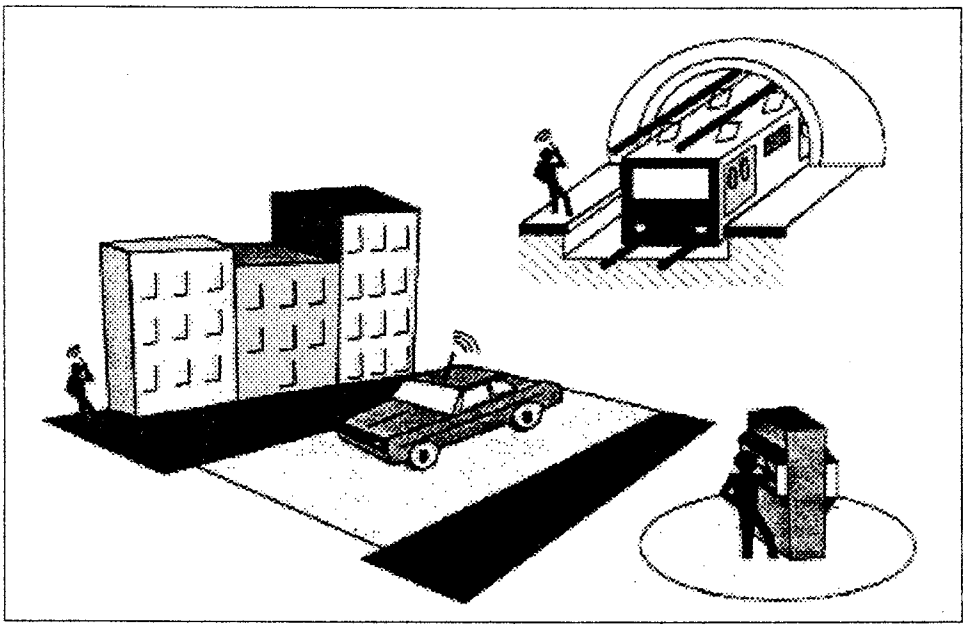

Figura 3

Estos factores hacen que los seres humanos seamos claramente superiores a los reconocedores de voz, que hacen su trabajo basándose en conocimiento acústico y en un modelo pobre del lenguaje. Está claro, por tanto, que los sistemas de reconocimiento deberían emplear más información para realizar su tarea, aunque no se sepa a ciencia cierta qué información ni cómo utilizarla.

Desde el punto de vista del reconocimiento de voz, ruido es cualquier señal acústica o eléctrica que contamine la señal de voz que queremos reconocer. Atendiendo a esta definición, el ruido se puede clasificar en tres grandes grupos:

- *Ruido estacionario:* dentro de este grupo englobaríamos todos aquellos ruidos cuyas propiedades se mantienen constantes a lo largo del tiempo o al menos en un periodo suficientemente largo de tiempo. La característica fundamental de este grupo es que conocido el ruido que hay en un periodo de tiempo, podemos predecir con bastante exactitud el ruido que tendremos en un instante de tiempo posterior. Como ejemplos de este tipo de ruido tendríamos el ruido producido por los tubos fluorescentes, por los ventiladores de los ordenadores, por un motor al ralentí, etc.
- *Ruido no estacionario:* agruparía todos los ruidos que no entran en el apartado anterior. Esto es: ruidos impredecibles, que están variando continuamente o que aparecen de forma intermitente sin ninguna periodicidad. Desafortunadamente, la mayoría de los ruidos que existen en el mundo real son de este tipo. Así, por ejemplo, el ruido del tráfico o el de una fábrica pertenecen a este grupo.
- *Voces de fondo:* pertenecerían estrictamente al segundo grupo, pero he preferido separarlo del mismo por su efecto especialmente dañino en los sistemas de reconocimiento. Las voces de fondo no sólo degradan las tasas de reconocimiento por contaminar la voz que el reconocedor está procesando, sino que, en un momento dado, un servicio podría fallar por reconocer la voz de fondo producida por un interlocutor distinto del usuario del servicio (aun habiendo reconocido correctamente la frase pronunciada por dicho interlocutor).

En el caso de reconocimiento por línea telefónica, el problema del ruido se agrava por la influencia del canal telefónico, entendiéndose por canal el conjunto de todos los elementos que hay entre el locutor y el reconocedor: El micrófono, la propia línea telefónica y los diversos circuitos eléctricos y electrónicos por los que pasa la voz antes de llegar al sistema de reconocimiento. En estos momentos, la calidad de las líneas telefónicas españolas es bastante buena y sigue mejorando de forma continuada, debido a su creciente grado de digitalización. Por contra, la también creciente variedad de teléfonos que se conectan a la red telefónica produce diversos efectos en la señal de voz que, en general, no son nada beneficiosos para el buen funcionamiento del reconocedor.

Existen diversas técnicas para atacar el problema del ruido estacionario y ciertas distorsiones debidas a la variabilidad del canal. Sin embargo, no se puede decir lo mismo del caso del ruido no estacionario y de las voces de fondo, que siguen siendo problemas sin resolver para el caso de reconocimiento por línea telefónica.

En la figura 3 se observan diversos entornos desde los cuales se pueden realizar llamadas telefónicas en la actualidad. Hay que resaltar que, con la introducción de los teléfonos móviles digitales (GSM) y debido a su creciente utilización, han surgido dos nuevas dificultades que no se han mencionado explícitamente: La primera estaría incluida en lo que llamamos anteriormente canal telefónico y consiste en el efecto nocivo de la codificación GSM y del canal de transmisión en las tasas de reconocimiento. La segunda se refiere a todos los nuevos entornos desde los que se puede realizar una llamada telefónica con esta tecnología, que en general son ruidosos y por tanto incrementarán las tasas de error.

En la figura 4 se observa el efecto del ruido estacionario en las tasas de reconocimiento de un reconocedor de palabras aisladas por línea telefónica cuyo vocabulario es de unas 500 palabras. En la realización del experimento no se empleó ninguna técnica de compensación de ruido. El eje de abscisas (horizontal) muestra la relación señal a ruido (RSR). Esto es, el cociente entre la potencia de señal de voz y la de ruido en decibelios (dB). El eje va de 10 dB a 30 dB (10 dB indica que la potencia de señal de voz es 10 veces superior a la de ruido y 30 dB indica que la potencia de voz es 1000 veces superior a la de ruido). En el eje de ordenadas (vertical) se presenta la tasa de acierto de palabras (TAP). Como se puede observar, la tasa de error aumenta considerablemente cuando la relación señal a ruido disminuye.

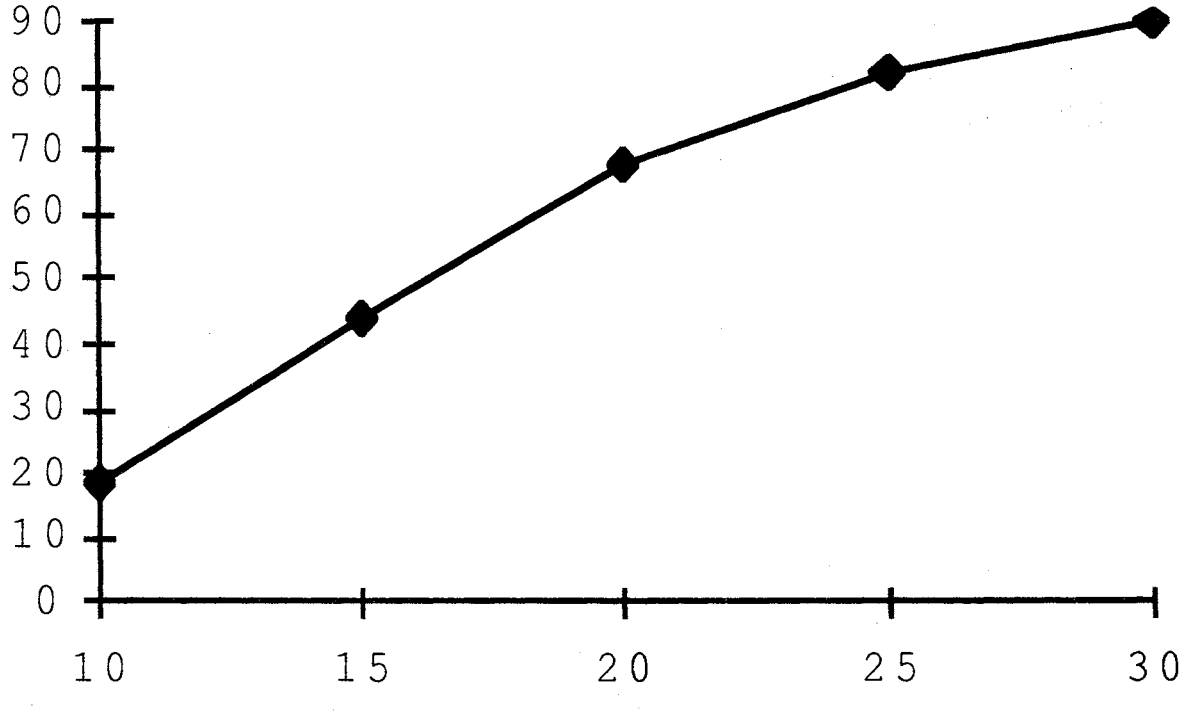

Figura 4. *Efecto del ruido estacionario en las tasas de reconocimiento*

4.3. *Independencia del dominio semántico*

De todo lo visto hasta ahora, se desprende que el reconocimiento de voz por línea telefónica se enfrenta con una serie de dificultades que limitan seriamente las posibilidades de utilización de estas tecnologías en servicios reales: por un lado el problema de la independencia del locutor, que, como se ha mostrado en el apartado 4.1, no está totalmente resuelto. Por otro lado, está el problema del canal y del ruido, que degrada las tasas de reconocimiento muy seriamente, tal y como se ha visto en el apartado anterior.

Dado que ninguno de estos problemas está resuelto en la actualidad, la única manera que existe de utilizar estas tecnologías con garantía de éxito es:

- reducir el vocabulario que maneja el reconocedor de voz, a fin de disminuir la probabilidad de que confunda unas palabras con otras y eliminar ambigüedades (ya que una misma palabra puede tener significados distintos en dos contextos distintos).
- diseñar diálogos que guíen al usuario para reducir al máximo la libertad del mismo a la hora de interaccionar con el sistema y así conseguir que diga las cosas que el reconocedor es capaz de reconocer.

La reducción del vocabulario conlleva la especialización del mismo en la tarea en la que se va a utilizar el reconocedor y, consecuentemente, la especialización del modelo del lenguaje, de los modelos acústicos, del diccionario y del módulo de procesamiento del lenguaje natural (PLN): el modelo del lenguaje estará diseñado para que favorezca al máximo combinaciones de palabras con sentido dentro del dominio de la tarea, los modelos acústicos se habrán obtenido probablemente a partir de frases relacionadas con dicho dominio, el diccionario estará compuesto únicamente por las palabras del vocabulario y el módulo de PLN estará configurado por reglas que incluso podrán estar metidas en el propio código y que estarán adaptadas al dominio en cuestión. Así pues, si quisiéramos emplear el mismo reconocedor en una tarea distinta, habría que cambiar el vocabulario y, con ello, los modelos del lenguaje, el diccionario, los modelos acústicos y el módulo de PLN.

En la mayoría de los casos, cuando se va a utilizar un sistema de reconocimiento para dar un nuevo servicio por línea telefónica, no se conoce a priori el comportamiento de los usuarios al encontrarse con un sistema automático, ni se sabe qué vocabulario ni qué tipo de frases van a emplear para comunicarse con el mismo. Por tanto, en la creación de un nuevo servicio, hay que establecer una hipótesis de partida para poder configurar un sistema inicial que, tras sucesivas pruebas de campo, va completándose y adaptándose a la tarea para la que fue diseñado.

El problema es que mientras que el diccionario y el módulo de PLN se pueden crear a partir de unas hipótesis de trabajo, el modelo del lenguaje necesita de miles de frases para poder generar un modelo que esté adaptado a la tarea. Esto es, que no

podremos disponer del mismo hasta que no se haya probado el sistema lo suficiente como para disponer de dichas frases. En cuanto a los modelos acústicos, el problema es menor, dado que existen técnicas que permiten predecir modelos a partir de los que ya están generados, por lo que se puede disponer de unos modelos iniciales que funcionen de una forma bastante satisfactoria. Adicionalmente, cada vez hay más bases de datos de voz disponibles, por lo que es relativamente sencillo generar modelos acústicos genéricos, que tengan un buen comportamiento en la mayoría de las tareas.

La solución al problema del modelo del lenguaje es la de crear aplicaciones Mago de Oz, que consisten en hacer creer al usuario que está empleando un servicio automático, cuando en realidad es una operadora la que está dando el servicio. De esta forma, se puede hacer un estudio del vocabulario empleado y recoger las frases y voz necesarias para entrenar una primera versión del sistema. El principal inconveniente de este método, además del coste, es el tiempo necesario para recoger todos los datos que se necesitan. Es necesario, por tanto, disponer de técnicas que permitan predecir el modelo del lenguaje de una tarea concreta a partir de modelos más generalistas para poder hacer servicios nuevos de una forma más rápida y económica.

Ésta es un área en el que se está haciendo un esfuerzo de investigación importante y en la que Telefónica I+D ha desarrollado ya algunas estrategias con el objetivo de mitigar, en lo posible, el efecto de la falta de datos para entrenar apropiadamente los modelos del lenguaje. En la figura 5 se pueden ver las tasas de error de un reconocedor de habla continua con un vocabulario de 5.000 palabras en tres experimentos distintos: evaluación del reconocedor en la misma tarea para la que se había entrenado, evaluación realizada en una tarea distinta y evaluación en la tarea distinta, tras aplicar un método de adaptación del modelo del lenguaje desarrollado en la empresa. Como se puede comprobar, la técnica de adaptación reduce en un 27% la tasa de error empleando tan solo unos pocos cientos de frases (Crespo 1997).

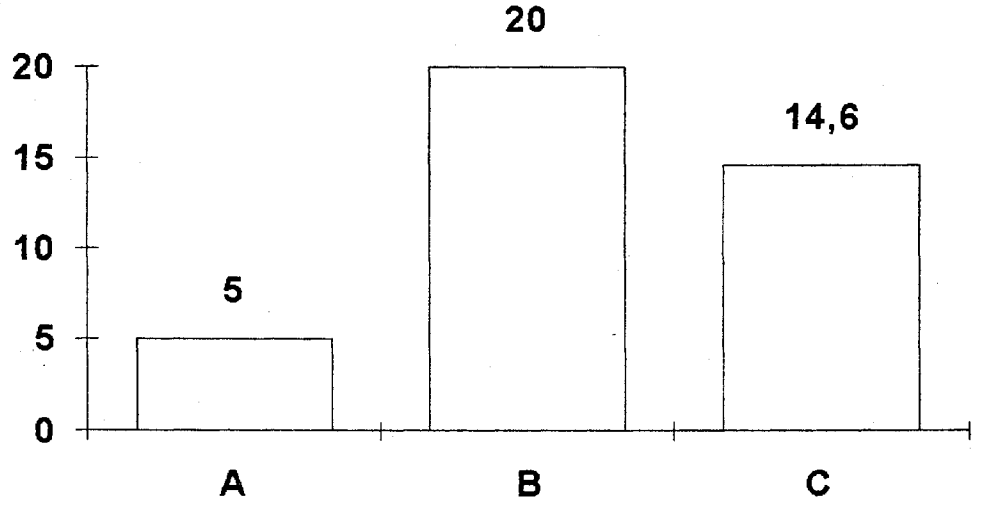

Figura 5. *Tasas de error asociadas al cambio de dominio*

4.4. *Habla espontánea*

Ésta es otra de las grandes cuestiones con las que se enfrenta el reconocimiento de habla continua en sistemas reales. El habla espontánea o natural es el tipo de habla que empleamos los seres humanos al comunicarnos entre nosotros y presenta una serie de problemas con respecto a otros estilos de habla como el habla leída, ya que no sigue completamente la normativa. Este estilo de habla presenta las siguientes características:

- Los sonidos aparecen pobremente articulados con relativa frecuencia.
- Se coarticulan muchos sonidos.
- La velocidad del habla es variable.
- Suelen aparecer correcciones, falsos comienzos de frase y sonidos guturales.
- No sigue estrictamente las reglas del lenguaje.
- Un mismo idioma puede tener multitud de acentos/dialectos.

Los efectos de todas estas características en la señal de voz son importantes, ya que existen muchos sonidos que desaparecen o que pierden parte de sus propiedades acústicas, hay sonidos que no pertenecen al idioma "normativo" pero sí pertenecen a un determinado dialecto del mismo, los modelos del lenguaje funcionan peor que con estilo de habla leído ya que las correcciones, falsos comienzos y sonidos guturales, junto con la falta de gramaticalidad de algunas frases, reducen la capacidad de corrección del modelo del lenguaje, etc.

Los cuatro problemas descritos constituyen los principales retos que tiene la tecnología de reconocimiento de voz en nuestros días. Las soluciones de estas cuestiones, lejos de ser responsabilidad de una única disciplina de la ciencia, se alcanzarán como resultado de la sinergia de múltiples áreas de conocimiento y precisarán de ordenadores mucho más potentes que los actuales. Quedan todavía muchos experimentos por hacer y muchas cosas por comprender antes de llegar a soluciones definitivas, pero también es cierto que los medios con que contamos actualmente, junto con el elevado número de personas que están investigando en este tema, hacen que podamos mirar hacia el futuro con optimismo. Ya existen muchos servicios que pueden automatizarse con la tecnología existente y las mejoras incrementales que se producen de forma continuada cada año van permitiendo el desarrollo de sistemas más complejos que hacen posible la puesta en marcha de servicios cada vez más sofisticados.

Como muestra de esta evolución, en el siguiente apartado se describe el sistema conversacional ATOS, desarrollado en TI+D en 1996. En este momento se está finalizando la segunda versión del mismo, que pasará a ser producto comercial en breve.

5. EL SISTEMA CONVERSACIONAL ATOS

El sistema conversacional ATOS está concebido para que funcione en tiempo real y dé servicio por línea telefónica. Con el término "sistema conversacional" se indica que el sistema podrá aceptar ordenes habladas y mantener una conversación "inteligente" sobre la tarea para la que ha sido configurado.

ATOS está constituido por cuatro elementos: Un reconocedor de habla continua que maneja un vocabulario de 2.000 palabras, un módulo de procesamiento del lenguaje natural que extrae el significado de la frase reconocida, toma la decisión sobre la acción que se va a seguir y genera la frase con que interaccionará con el usuario, un conversor de texto en voz para pronunciar las frases con que da o solicita información al usuario y un gestor de bases de datos con el que accede a la información solicitada.

El módulo de procesamiento del lenguaje natural está formado por tres módulos: El analizador semántico, que es el encargado de extraer el significado de la frase reconocida, el gestor de diálogo, que toma las decisiones sobre las siguientes acciones que se deben realizar, y el generador de lenguaje natural, que genera las frases con que el sistema se comunicará con el usuario.

Con todos estos componentes, el sistema ATOS se ha configurado para que dé los siguientes servicios:

- Servicios de PABX: en la actualidad, las centralitas telefónicas o PABX ofrecen una gran variedad de servicios que pueden ser utilizados desde cualquier terminal telefónico que esté conectado a ellas. La cantidad de servicios a los que se tiene acceso desde un teléfono depende, por un lado del propio teléfono y, por otro, de la categoría asignada al mismo en la PABX. Así, por ejemplo, la capacidad para realizar llamadas internacionales, locales, etc. depende de la categoría asignada al terminal, mientras que la posibilidad de ver el número de teléfono desde el que se origina la llamada entrante depende de si el teléfono tiene un "display" (pantalla pequeña) o no. En cualquier caso, aunque el número de funciones disponible es variable, hay un gran número de ellas que pueden ser utilizadas desde cualquier terminal.
  Estos servicios presentan el inconveniente de que para poder utilizarlos, hay que memorizar una serie de códigos o consultar el manual de usuario cada vez que se deseen usar, lo que desanima a muchos usuarios que finalmente no emplean estas facilidades o solo utilizan un pequeño conjunto de las mismas.
  El sistema ATOS pretende solventar este problema, ya que permite al usuario explicar con su propia voz el tipo de servicio que quiere emplear. Los servicios de PABX proporcionados por el sistema ATOS son llamada en espera, multiconferencia, desvío de llamadas, modo no molesten, contestador, etc.

- Otros servicios: Adicionalmente, ATOS proporciona el servicio de directorio telefónico, que permite realizar llamadas de teléfono diciendo el nombre y el apellido de la persona o simplemente su número de teléfono. También ofrece el servicio de agenda personal, que permite que el usuario cree su propia agenda de teléfonos.

La primera versión del sistema conversacional ATOS fue evaluada en junio de 1996 (Álvarez, Tapias, Crespo, Cortázar y Martínez 1997) para determinar los puntos débiles del sistema y así poder resolverlos en la segunda versión. Tras la evaluación se vió que los usuarios habían pronunciado un 3.1% de palabras que no estaban en el vocabulario del sistema y que sólo el 1.4% de las frases pronunciadas habían tenido falsos comienzos. En la figura 6 se presentan los resultados obtenidos en esta evaluación. En la primera columna aparece el error global del sistema. Esto es, el porcentaje de frases que el sistema no pudo entender (22.1%). En la segunda columna se representa el error del módulo de procesamiento del lenguaje natural cuando se le pasaron las frases reconocidas (7.9%). La tercera columna representa el mismo error cuando las frases de entrada no tienen errores de reconocimiento (6.5%), que como se ve es muy parecido al anterior, por lo que muchos de los errores de reconocimiento no afectan seriamente al funcionamiento del módulo de procesamiento del lenguaje natural. La cuarta columna muestra la tasa de error del reconocedor, que es del 25%, cinco veces superior a la tasa de error del mismo cuando las frases de entrada fueron leídas (quinta columna).

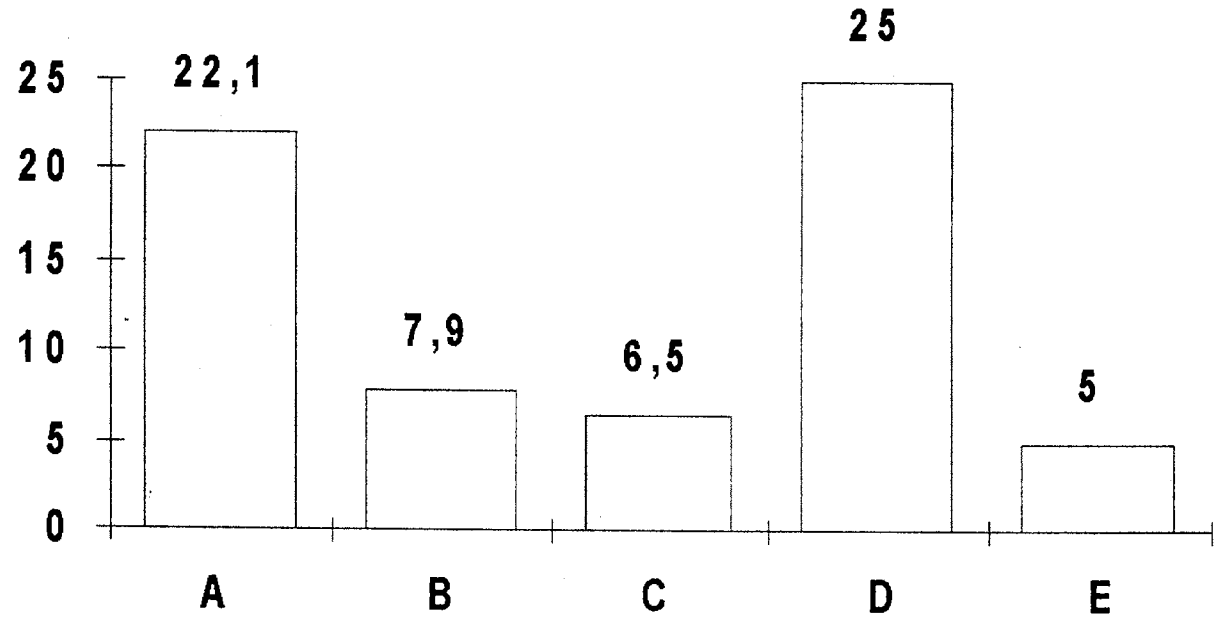

Figura 6. *Tasas de error del sistema conversacional ATOS (Junio de 1996)*

De todo este estudio se dedujo que había que mejorar tanto el reconocedor como la capacidad de recuperación frente a errores del sistema. Así mismo, se vio que ciertos tipos de anáforas y elipsis deberían ser resueltos por el sistema para permitir un diálogo más natural y evitar ciertos tipos de error.

6. SITUACIÓN ACTUAL DEL RECONOCIMIENTO DE HABLA CONTINUA

Para hablar de los sistemas más avanzados del mundo en reconocimiento del habla hay que referirse a los sistemas que participan en las evaluaciones organizadas anualmente por DARPA (Defense Advance Research Projects Agency) en Estados Unidos.

En los últimos años, las evaluaciones se han centrado en la tarea de reconocer programas de radio. Esta es una tarea muy compleja porque el tamaño del vocabulario es ilimitado, en el sentido de que se puede escuchar cualquier palabra y además, el entorno puede cambiar y de hecho cambia continuamente (entrevistas en la calle, música de fondo, voz de locutorio, llamadas telefónicas, etc.). En 1995 los programas de radio se reconocían directamente, sin proporcionar al reconocedor información sobre el tipo de entorno en que tenía que reconocer, lo que dió lugar a tasas de error elevadas porque la detección del entorno se hacía de forma automática. En las evaluaciones de 1996 se proporcionó esta información a priori, de forma que se evaluasen solamente los reconocedores y no los reconocedores junto con los clasificadores de entorno. Finalmente, en 1997, se obtuvo la información del entorno por medio de un clasificador automático realizado por la Universidad de Carnegie Mellon, aunque la utilización de esta información fue opcional. En la figura 7 se muestran los resultados de 1996 y en la figura 8 los resultados de 1997. No presento los resultados de las evaluaciones de 1998 dado que se están realizando en este momento y todavía no se dispone de los resultados.

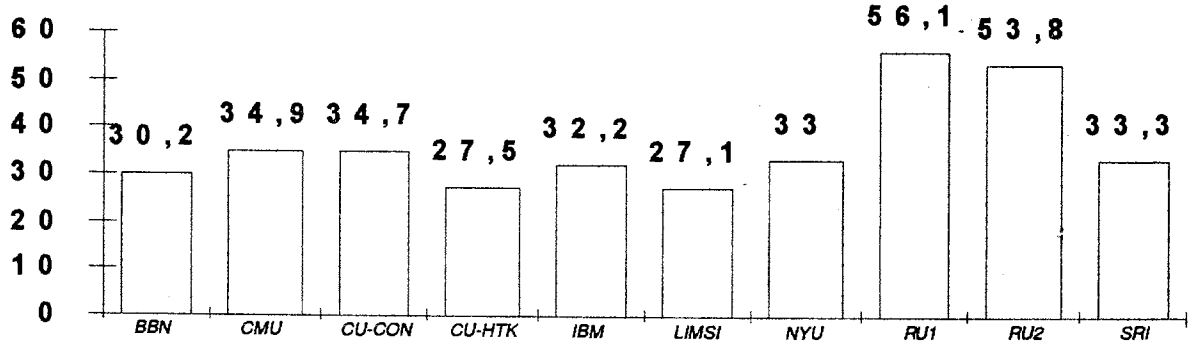

Figura 7. *Resultados de las evaluaciones de DARPA de 1996*

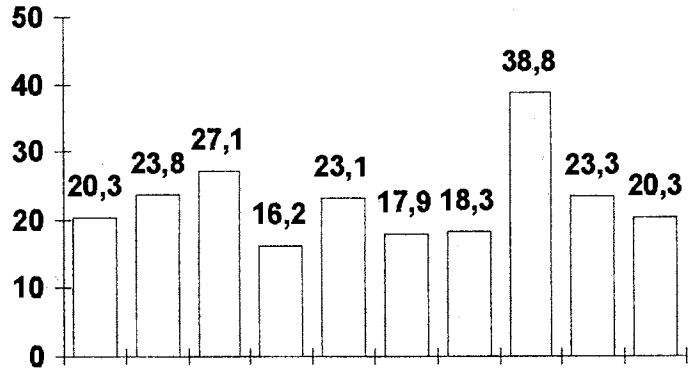

Figura 8. *Resultados de las evaluaciones de DARPA de 1997*

Como se puede apreciar, las tasas de error varían bastante de unos centros a otros. En 1996, los mejores resultados los consiguió el sistema de LIMSI, con un 27.1% de error, seguido muy de cerca por el sistema HTK de la Universidad de Cambridge (UC). En 1997, los mejores resultados los obtuvo el sistema HTK de UC, con un 16.2% de error, seguido por el sistema de IBM. El hecho de que la tasa de error haya bajado casi once puntos de un año a otro se debe por un lado a la mejora incremental de los sistemas que participan en las evaluaciones y por otro a que la base de datos con la que se evaluaron los reconocedores en 1997 tenía una proporción mucho mayor de pronunciaciones realizadas en las mismas condiciones de la base de datos de entrenamiento.

La primera pregunta que surge al analizar los resultados y compararlos con los del ATOS es por qué éste con 2.000 palabras tiene una tasa de error similar a la de un sistema que reconoce programas de radio con vocabularios de tamaño ilimitado. Ante esta pregunta, la primera y obligada respuesta es que no se pueden comparar dos sistemas de reconocimiento si no se fijan con mucho cuidado los parámetros de las pruebas. En primer lugar, las tasas de error, para que sean comparables, deben medirse en la misma prueba, cosa que no se cumple en este caso. En segundo lugar, estamos comparando dos sistemas completamente distintos: el reconocedor del ATOS está orientado a tiempo real. Esto es, debe reconocer en menos de uno o dos segundos para poder dar un servicio sin que el usuario se aburra. Sin embargo, los reconocedores evaluados en DARPA no son de tiempo real, porque su objetivo es reducir las tasas de error al mínimo. Así pues, pueden dedicar todo el tiempo que sea necesario y emplear tanta información como sea posible en el proceso de reconocimiento. Igualmente, no tienen restricciones en el tamaño de la memoria que necesitan para reconocer, mientras que el reconocedor del ATOS si tiene estas restricciones para no encarecer desmesuradamente el producto resultante. Adicionalmente, el programa de radio que se reconoce en estas evaluaciones tiene un número de personajes que hablan en todos los programas, por lo que los sistemas de reconocimiento pueden estar adaptados a sus voces y así aumentar las tasas de acierto.

7. CONCLUSIONES

La tecnología actual de reconocimiento de voz ha alcanzado un nivel que permite su explotación en servicios reales. Hemos visto, sin embargo, que todavía quedan muchos problemas por resolver, que surgen cuando las condiciones en que se ha entrenado el sistema difieren de las condiciones en las que se va a emplear. Por ejemplo, cuando la voz del usuario no está bien representada en la base de datos de entrenamiento, o cuando el nivel o el tipo de ruido de fondo es distinto del empleado al entrenar.

Por este motivo, Telefónica Investigación y Desarrollo, al igual que el resto de los centros más avanzados del mundo, está haciendo un importante esfuerzo para resolver o al menos mitigar el efecto nocivo del cambio de las condiciones en las tasas de reconocimiento. Este esfuerzo se refleja, año a año, en mejoras incrementales que se plasman tanto en los productos como en la calidad y complejidad de los servicios que se realizan con los mismos. Valga como ejemplo el caso de los sistemas conversacionales, que hace unos años eran patrimonio de las películas de ciencia ficción y que ya, en este momento, empiezan a surgir tímidamente en aplicaciones moderadamente complejas. Seguro que dentro de unos años formarán parte de nuestra vida cotidiana.

BIBLIOGRAFÍA

Alvar, M. 1996a. *Manual de dialectología hispánica: El Español de España.* Barcelona: Ariel.

Alvar, M. 1996b. *Manual de dialectología hispánica: El Español de América.* Barcelona: Ariel.

Álvarez, J., Tapias, D., Crespo, C., Cortazar, I., y Martínez, F. 1997. "Development and Evaluation of the ATOS Spontaneous Speech Conversational System". *Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Munich.

Crespo, C., Tapias, D., Escalada, G., Álvarez, J. 1997. "Language Model Adaptation for Conversational Speech Recognition Using Automatically Tagged Pseudo-morphological Classes". *Proceedings of the International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, Munich.

Eskénazi, M. 1993. "Trends in Speaking Styles Research". *Proceedings of Eurospeech'93*, Berlín. 501-509.

Junqua, J.C. y Haton, J.P. 1996. *Robustness in Automatic Speech Recognition. Fundamentals and Applications.* Kluwer Academic Publishers.

Lippmann, R.P. 1997. "Speech Recognition by Machines and Humans". *Speech Communication* 22: 1-15.

Llisterri, J. 1992. "Speaking Styles in Speech Research", *ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublín.