

# Tecnología de reconocimiento del habla para teléfonos móviles

Stefan Dobler

La introducción de teléfonos móviles que pueden manejarse con la voz ha hecho que el reconocimiento del habla se esté convirtiendo en una función estándar en los aparatos celulares. La marcación vocal facilita el uso y puede aumentar la seguridad en el manejo del teléfono al conducir un automóvil.

El autor expone brevemente los tipos de sistema de reconocimiento del habla, ilustrando algunos problemas típicos del entorno celular, y argumenta sobre un uso eficiente de la memoria en los terminales móviles. Además, presenta detalles de la tecnología aplicada en el sistema de reconocimiento del habla de Ericsson. Finalmente, describe los avances potenciales previstos en los teléfonos móviles del futuro controlados por la voz.

## Introducción

El año pasado, Ericsson fue uno de los primeros fabricantes de teléfonos móviles que añadió avances tecnológicos importantes en sus aparatos. El modelo T18, introducido en primavera de 1999, fue el primer teléfono GSM del mercado que podía operarse mediante órdenes de voz usando un sistema de reconocimiento del habla automático, además del manejo tradicional con el teclado. Otros miembros de la familia de teléfonos que usan la primera generación de algoritmos de control de la voz son los modelos T28, R320 y A2618 (figura 1).

Estos aparatos usan el nuevo sistema de *marcación vocal*. Gracias a un uso eficiente de la memoria, actualmente es posible instruir y almacenar en la lista de teléfonos de los terminales antedichos etiquetas vocales para hasta 10 números. Cada etiqueta vocal se instruye en el aparato mediante una sola expresión de voz del usuario, y se asigna a un solo registro de teléfono en la lista. Cuando el usuario desea hacer una llamada, pulsa el botón y pronuncia el nombre de una persona. El teléfono contesta emitiendo la etiqueta vocal como "feedback" o retroali-

mentación acústica, y establece la llamada automáticamente.

Todos los teléfonos de Ericsson con funciones de reconocimiento del habla también incorporan la *respuesta a llamadas*, que permite al usuario aceptar o rechazar llamadas entrantes usando órdenes de voz. Evidentemente, esto tiene ventajas cuando el aparato se emplea con accesorios de manos libres.

En comparación con los programas para dictar textos que hay a la venta para PC de sobremesa, la aplicación aquí descrita puede parecer elemental. Sin embargo, los teléfonos móviles se emplean cada día, en un gran número de lugares, con ruidos de fondo de los tipos más variados imaginables. Por consiguiente, la cuestión decisiva para el reconocimiento del habla en los terminales móviles no es el tamaño del vocabulario, sino la robustez del sistema de reconocimiento.

Por un lado, el teléfono tiene que reconocer la voz correctamente, por ejemplo, en una oficina silenciosa, en un aeropuerto con gente hablando como telón de fondo acústico o en un automóvil que se desplace a 150 km/h. Y, por otro lado, no debe tener en cuenta ruidos fortuitos. Una puerta que se cierra o una carcajada no deberán confundirse con un nombre válido, provocando que el teléfono realice una llamada. Asimismo, el sistema de reconocimiento debe funcionar correctamente con cualquier tipo de micrófono, a diferentes distancias y ángulos entre la boca y el micrófono, y a pesar de cambios entre el uso normal y con accesorios de manos libres, y sin tener que instruir nuevamente al sistema.

## La necesidad de reconocimiento del habla

Las razones por las que el reconocimiento del habla se está convirtiendo en una función estándar en los terminales móviles son varias. El uso del teléfono celular al conducir se considera peligroso porque distrae. Para hacer una llamada, el conductor tiene que apartar la mano del volante y pulsar el número en el teclado. Y para controlar el número, debe apartar los ojos de la carretera y dirigir la mirada a la pantalla del teléfono. Durante la conversación le es preciso sujetar el teléfono con la mano. Ya se han impuesto restricciones legales o están en camino de imponerse (en Japón y Alemania, por ejemplo), que prohíben usar un teléfono celular mientras se conduce, a menos que se emplee un accesorio de manos libres.

El accesorio de manos libre es un importante avance, puesto que permite al conductor conservar ambas manos en el volante durante la conversación. Al establecer la llamada, la marcación vocal ofrece al usuario la posibilidad de pronunciar el nombre de la persona en lugar de teclear su número de teléfono. Además, la repetición acústica de "feedback" del nombre recono-

Figura 1  
Teléfonos que usan la primera generación de tecnología de control vocal de Ericsson.



cido contribuye a que el conductor no tenga que apartar la mirada de la carretera ni las manos del volante. Por tanto, estas funciones pueden incrementar la seguridad de uso de los aparatos celulares en un automóvil.

Aparte de las ventajas en funcionalidad, el desarrollo habido en el sector durante los últimos años se ha caracterizado por una disminución del tamaño físico de los teléfonos móviles. Cada vez se fabrican aparatos más pequeños. Además, los usuarios piden pantallas más grandes para nuevos servicios. Como resultado de ello los teclados de los aparatos han disminuido a un tamaño que a veces hace el manejo muy incómodo. De vez en cuando se difunden noticias de teléfonos en forma de lápiz, sin teclado.

Los humanos consideramos el habla como el medio de comunicación más natural. Por tanto, el reconocimiento automático de la voz puede convertirse en un interfaz lógico para el usuario celular. Reduce la interacción manual con los teléfonos móviles. Al marcar por voz, por ejemplo, en lugar de buscar en la lista de teléfonos el nombre en cuestión, basta con que el usuario pronuncie esta palabra y el aparato establece la llamada automáticamente.

Continuamente se introducen nuevas formas de utilizar los teléfonos celulares. Por ejemplo, los minicascos telefónicos Bluetooth permitirán emplear aparatos móviles con equipo de manos libres inalámbrico. Estos minicascos y un interfaz de reconocimiento del habla constituyen una alternativa útil a los teclados de los teléfonos. Otras innovaciones incrementarán también la necesidad de un reconocimiento del habla fiable.

## Tipos de sistema de reconocimiento del habla

En un teléfono móvil, la elección del tipo de sistema de reconocimiento del habla es crucial. Cada sistema tiene ventajas y desventajas.

### Palabras aisladas y palabras interconectadas

Los sistemas para reconocer palabras aisladas pueden distinguir un vocablo en una ventana de reconocimiento. Para que tengan la posibilidad de analizar una secuencia de palabras, debe hacerse una pausa entre cada palabra y reiniciar el reconocimiento, lo cual comporta una pronunciación poco natural.

Como ya indica su nombre, los reconocedores de palabras interconectadas permiten que el usuario pronuncie varias palabras sin hacer pausas artificiales entre ellas. En comparación con los reconocedores de palabras aisladas tienen el inconveniente de su mayor complejidad. Un problema es la coarticulación, es decir, que los vocablos interconectados se pronuncian de modo diferente, tal como sucede con una secuencia de cifras. Entonces las palabras suelen

mezclarse y los límites entre ellas desaparecen. Sin embargo, debido a que permiten hablar de forma natural, estos reconocedores son los preferidos por su mayor comodidad.

### Reconocimiento dependiente e independiente de la persona

Los usuarios de teléfonos móviles que emplean la actual tecnología de reconocimiento tienen que instruir a sus aparatos antes de poder utilizar la nueva función. Esta instrucción el usuario la considera una tarea engorrosa. No obstante, una ventaja de los sistemas que dependen de la persona que habla es su independencia del idioma. Cualquier usuario puede instruir su teléfono en la lengua que desee, factor que simplifica la introducción de control de habla como una nueva función en un producto como el teléfono móvil, que se vende por diferentes países con distintos idiomas. Asimismo, debido a que el sistema aprende el comportamiento de habla de cada usuario, sus prestaciones son superiores a las de los procedimientos de reconocimiento que no dependen de la persona.

Los sistemas de reconocimiento independientes de quien habla se instruyen previamente empleando una gran diversidad de personas. El usuario puede empezar a usar el sistema de inmediato, lo cual facilita enormemente el manejo. Un obstáculo de ellos es la imposibilidad de modelar todas las variaciones posibles de habla de un idioma. Por consiguiente siempre habrá un pequeño porcentaje de usuarios cuyos teléfonos alcanzaran unas prestaciones subóptimas. Una solución a este problema sería dar a estas personas la opción de instruir el aparato en todas o algunas de las palabras del vocabulario que no funciona bien en ellos. Esto exigiría combinar en un sólo teléfono el reconocimiento dependiente y el independiente de quien habla.

### Reconocimiento de vocabularios pequeños y grandes

Para dar órdenes y para el control, basta con sistemas de reconocimiento que tengan vocabularios pequeños, de hasta 100 palabras. Normalmente estos sistemas se basan en vocablos, lo cual significa que cada elemento del vocabulario consiste en una palabra. Son de poca complejidad y el espacio de memoria necesario es limitado.

Como contraste, un sistema de reconocimiento de un gran vocabulario, para poder dictar cartas o mensajes electrónicos, puede contener hasta 100.000 palabras. Entonces, los elementos del vocabulario básico son subunidades de habla más pequeñas, tal como fonemas, que se emplean para construir las palabras del vocabulario. Son muy complejos y el espacio de memoria preciso es muy grande. Para implementar un auténtico sistema de dictado se necesitan varias docenas de megabytes, lo cual es un requisito inaceptable en un teléfono móvil.

## CUADRO A, TÉRMINOS Y ACRONIMOS

### Extracción de características

Un procedimiento de procesado previo en la operación de reconocimiento del habla que transforma la señal de voz en una representación espectral.

### Vector de características

El resultado producido por la extracción de características.

### HMM

Hidden Markov model.

### MIPS

Megainstrucciones por segundo.

### MMI

Man-machine interface.

### Comparación de configuraciones

También denominada búsqueda; el procedimiento en el reconocimiento del habla que compara la expresión pronunciada por el usuario con etiquetas vocales. Las etiquetas vocales a veces también se llaman configuraciones.

### RAM

Random access memory.

### SMS

Short message service.

### Vocabulario

Término colectivo de todas las etiquetas vocales instruidas.

### Etiqueta vocal

Una representación de voz adaptada para un reconocedor de habla.

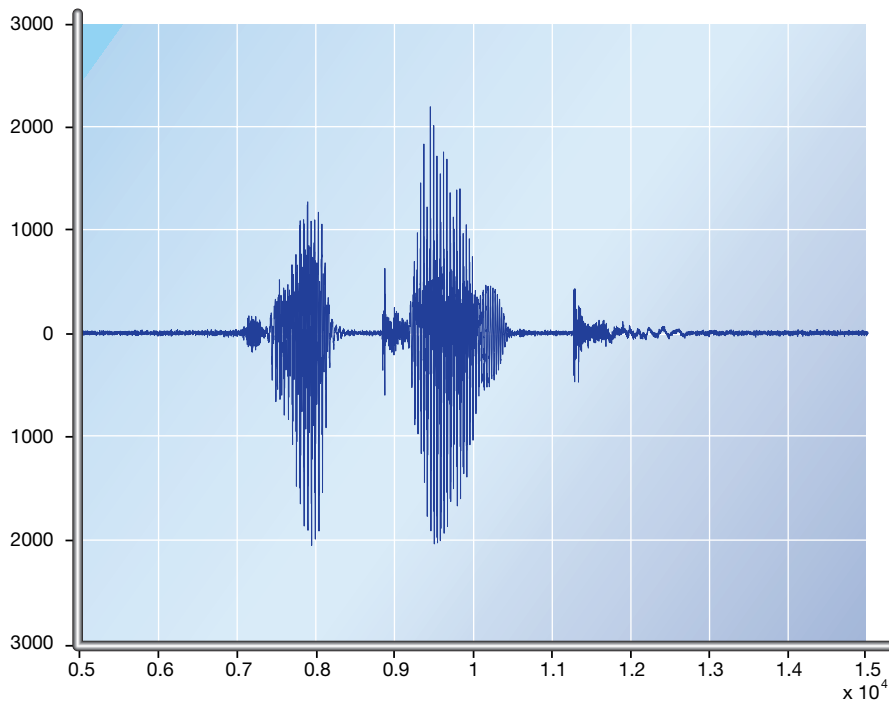
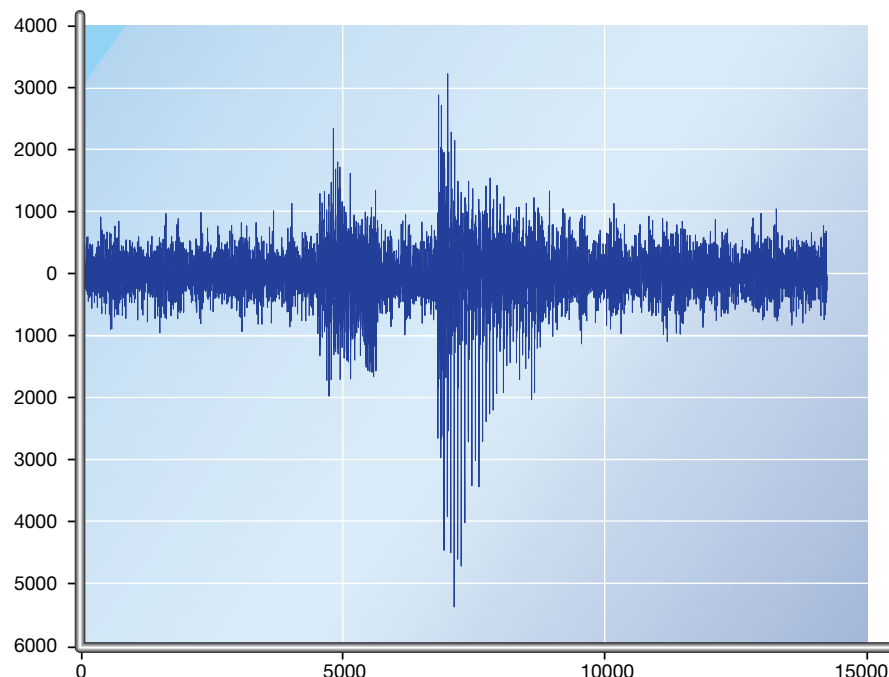


Figura 2  
Sin ruidos; archivo de voz de la palabra *repeat* (repita).

Figura 3  
Con ruido estacionario; archivo de voz de la palabra *repeat*.



## Retos

### Acústica

En comparación con los teléfonos móviles, los aparatos de línea fija suelen usarse en un entorno acústico estable. Se emplean en el mismo recinto, una oficina o sala de estar, por ejemplo, con casi el mismo fondo acústico diariamente.

En las comunicaciones móviles, el ruido de fondo siempre está presente y es muy variable. Los teléfonos celulares se utilizan en cualquier entorno concebible. Puede ser una oficina, o un aeropuerto, una estación de ferrocarril, o al aire libre en un entorno que impone retos acústicos. El interior de los vehículos —con la interferencia provocada por el ruido de fondo, que depende del tipo de automóvil, la velocidad, etc.— es un reto que sólo tienen que afrontarlo los teléfonos móviles. En consecuencia, en las comunicaciones móviles no se pueden hacer suposiciones sobre las propiedades acústicas del entorno de uso o el ruido de fondo.

Asimismo, una cierta proporción de usuarios móviles cambian con frecuencia de manejo directo a manejo con manos libres, ya sea con equipo especial en el automóvil o accesorios portátiles de manos libres. Esto ocasiona grandes variaciones en la señal de voz, además de la variación convencional de atenuación de un usuario a otro.

Las figuras 2-4 muestran las diferencias en señales de habla causadas por el ruido de fondo en entornos típicos. La primera ilustración muestra la señal de voz de una palabra pronunciada sin fondo sonoro. Reconocerla debería ser una tarea sencilla para cualquier sistema de reconocimiento.

La segunda ilustración muestra la señal de voz producida por la misma palabra pronunciada teniendo un fondo sonoro alto. Esta situación corresponde a grandes rasgos al uso de manos libres dentro de un automóvil. La distancia de la boca al micrófono es de unos 30 centímetros. El ruido de fondo lo causa el motor, el viento y los automóviles que van pasando. Una comparación de las señales de habla en ambos casos presenta problemas de reconocimiento. Un alto ruido de fondo disminuye las prestaciones de los sistemas reconocedores. Hay técnicas, tales como la sustracción espectral, para atajar los ruidos durante las fases de procesamiento previo o extracción de las características de la voz, siempre que los ruidos sean estacionarios.

En la tercera señal de voz se presenta una situación más extrema (figura 4). Aquí el ruido no estacionario procede de gente que habla en segundo plano. Es la situación más exigente para un sistema de reconocimiento vocal automático, debido a que en este caso no sirven los métodos que abordan el ruido estacionario en la extracción de las características de la voz. El ruido de fondo también es habla humana, por lo que los métodos estadísticos no contribuyen a dis-

tinguir entre la aportación deseada del usuario y la no deseada de personas hablando en el entorno. No obstante, es una situación bastante típica en el uso del teléfono móvil, y los usuarios esperan que sus aparatos puedan funcionar en todos los entornos acústicos posibles.

### Restricciones del hardware

Algunas características de los teléfonos móviles influyen sobre los sistemas de reconocimiento del habla que pueden utilizarse. En general, estos terminales son:

- Aparatos muy pequeños;
- Fabricados en grandes volúmenes; y
- Con un costo altamente optimizado.

Esto restringe considerablemente el hardware para los algoritmos de reconocimiento del habla, puesto que limita la capacidad en términos de carga computacional, tamaño de memoria y velocidad de acceso a la memoria.

### Carga computacional

Para reducir al mínimo el tamaño y costo de los aparatos móviles, no se les incorpora un procesador adicional sólo para el control del habla. Los recursos informáticos existentes se han de compartir con otras aplicaciones en el teléfono. A pesar de todo, los procesadores de señales digitales usados actualmente en los teléfonos móviles ofrecen como mínimo 40 megainstrucciones por segundo (MIPS), y las prestaciones pueden alcanzar los 180 MIPS. Por eso, este cuello de botella no es un problema grave.

### Tamaño de memoria

La primera generación de teléfonos Ericsson con marcación vocal tenía un vocabulario de 10 nombres. Ericsson Research ha desarrollado un sistema que reduce al mínimo la memoria adicional precisa para almacenar las etiquetas vocales y el "feedback" auditivo de los nombres instruidos. El problema de la memoria disminuye con las modernas memorias programables flash, que permiten alojar un vocabulario de marcación vocal relativamente grande e implementar en los teléfonos móviles nuevas funciones de control por voz. Sin embargo, la memoria disponible es todavía demasiado pequeña para implementar un sistema de gran vocabulario que soporte el dictado de mensajes electrónicos y aplicaciones similares.

### Velocidad de acceso a la memoria

Normalmente, los teléfonos móviles consisten en un sistema de multiprocesador con distintas memorias unidas a los procesadores disponibles. La transferencia de datos entre estas memorias tiende a ser un problema, debido a que las arquitecturas de hardware ordinarias de los teléfonos móviles no se han diseñado para las altas velocidades de transmisión de datos necesarias para el reconocimiento automático del habla. Las prestaciones computacionales de los proce-

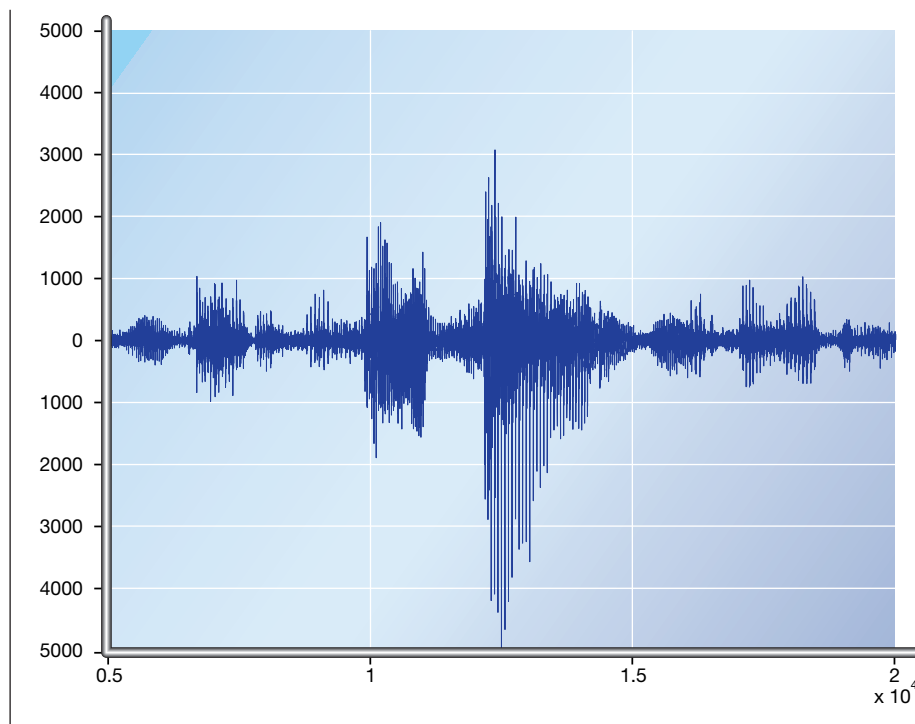


Figura 4  
Ruido de fondo no estacionario; archivo de voz de la palabra *repeat*.

sadores de señales digitales son buenas, pero las de la memoria de acceso aleatorio (RAM) son bastante limitadas. En consecuencia, se precisa disponer de acceso a la memoria flash del procesador de control..

## Tecnología

### El proceso de reconocimiento del habla

Por definición, los sistemas de reconocimiento del habla dependientes del usuario exigen que éste realice un programa de instrucción para crear el vocabulario para el reconocedor. El programa lee la secuencia de los vectores de características que corresponden a la pronunciación de la palabra a instruir y luego crea y almacena un modelo de la misma. Más tarde, cuando la palabra se usa como orden, el sistema extrae las características, intenta comparar la secuencia de características en la pronunciación, y luego ejecuta una determinación de confianza (figura 5).

La señal de voz  $v(t)$  captada por el micrófono se alimenta a la función de extracción de características, que obtiene información esencial de la señal. El resultado de esta función es el vector de características,  $y_k$ , que describe las propiedades acústicas básicas de dicha señal en un intervalo de tiempo dado. La letra  $k$  indica el ín-

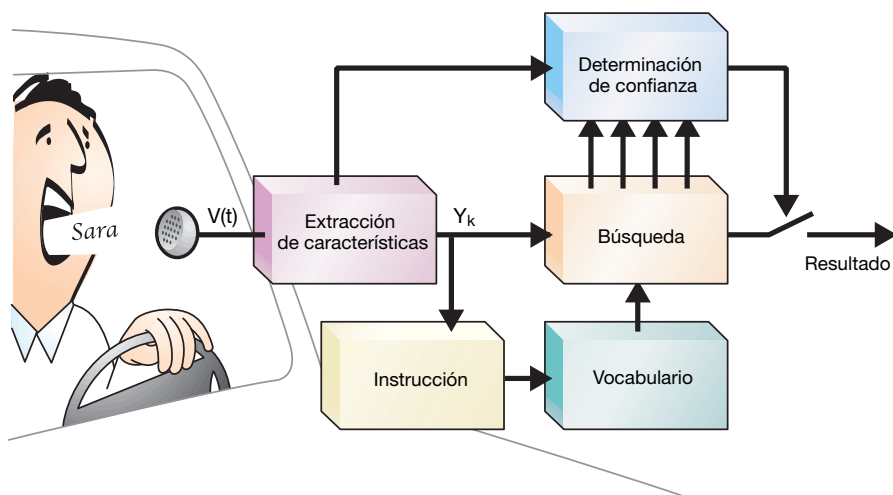


Figura 5  
Diagrama de bloques general del reconocedor de habla.

secuencia de los vectores de características entrantes,  $y_k$ , con el vocabulario disponible. La palabra del vocabulario más parecida se convierte en el resultado del reconocimiento. Si un usuario acciona involuntariamente el sistema de reconocimiento, cualquier actividad de señal de habla puede producir un resultado, estableciendo por consiguiente una llamada. Lógicamente, esto es inaceptable, por lo que se ha añadido un bloque adicional: la función de determinación de confianza.

Una vez la función de búsqueda termina pasada de reconocimiento, la función de determinación de confianza lee parámetros de la extracción de características (tales como la relación entre señal y ruido durante el reconocimiento), y busca funciones (por ejemplo la distancia entre la primera y segunda comparación mejores). Basándose en estos parámetros, se toma una decisión sobre si aceptar y transferir los resultados del interfaz hombre-máquina superpuesto (MMI) o rechazar los resultados de la búsqueda.

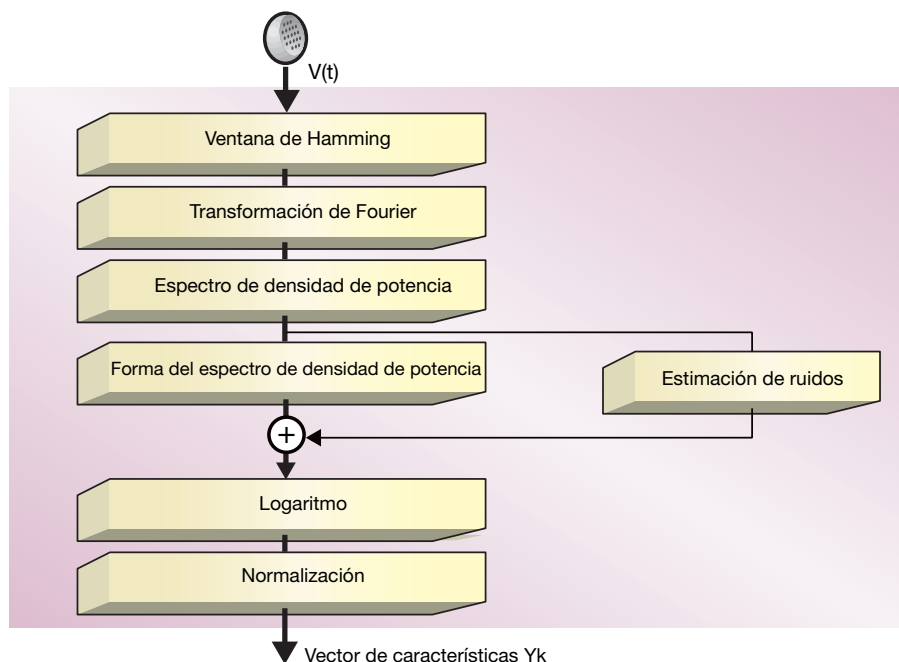
dice de un intervalo de tiempo. Típicamente, se genera un nuevo vector de características cada 10 ms.

Los vectores de características se alimentan a la función de búsqueda, que posee datos adicionales del vocabulario instruido; en el caso de teléfonos móviles, este vocabulario consiste normalmente en modelos de una sola palabra. La tarea de la función de búsqueda es comparar la

### Extracción de características

La figura 6 muestra los principales bloques de la función de extracción de características. La señal de habla,  $v(t)$ , que consiste en una corriente continua de ejemplos de voz, se divide para el procesamiento de tramas. Partiendo de una frecuencia de muestreo de 8 kHz, se combinan típicamente 256 muestras (es decir, una longitud de ventana de 32 ms) para construir una trama con un solape de 176 muestras sobre tramas conse-

Figura 6  
Bloques principales de la función de extracción de características.



cutivas. El despliegue de una ventana de Hamming en cada trama proporciona un alisado entre las tramas. La transformación de Fourier convierte la señal del dominio de tiempo al de frecuencia. Esto proporciona un vector de características que representa las propiedades espectrales básicas de la señal de voz.

Las fases de procesamiento subsiguientes usan el espectro de densidad de potencia en lugar del complejo espectro de Fourier. Las investigaciones realizadas han comprobado que el oído humano es insensible a esas variaciones, por lo que en esta etapa se retira la información de fase de la señal. La estructura fina del espectro de densidad de potencia transporta información amplia sobre la persona que habla y sus características vocales. La información sobre el vocablo pronunciado propiamente dicho puede encontrarse en el envolvente o la forma del espectro. El envolvente se deriva muestreando el espectro a distintas frecuencias centrales, que se distribuyen de una forma no lineal por el eje de frecuencias, nuevamente similar a la audición humana. El sistema de reconocimiento de Ericsson usa  $\mu = 15$  frecuencias centrales.

Un problema importante durante el reconocimiento es el ruido de fondo estacionario. Este ruido puede modelarse como un componente de ruido aditivo  $N$  a la señal de habla  $S$  en el dominio de densidad de potencia. Para cada una de las frecuencias centrales, esto puede describirse como:

$$X_{\Delta}(\mu) = S(\mu) + N(\mu)$$

El nivel sonoro  $N$  se valora en el bloque de estimación del nivel de ruidos. Sustrayendo la valoración del ruido de fondo se reduce al mínimo la influencia de la señal de ruido en  $X_{\Delta}$ , haciendo el sistema más robusto a los ruidos de fondo estacionarios.

Otro problema es un entorno acústico con variaciones en los niveles del micrófono. Un medio de atajarlo es utilizando un logaritmo del espectro de densidad de potencia.  $X_{\Delta}(\mu)$ , que es la frecuencia central número  $\mu$  antes de la logaritimización, se multiplica por un factor constante,  $V$ .

$$VX_{\Delta} = 10 \log\{V X_{\Delta}(\mu)\} \text{ con } \mu L \{1..15\}$$

Hay que tener en cuenta que en el sistema de Ericsson,  $\mu$  es entre 1 y 15. El factor  $V$  proviene de condiciones tales como que la voz de un usuario sea más alta que la de otro. Por tanto, la fórmula antes citada puede volverse a escribir del siguiente modo:

$$VX_{\Delta} = 10 \log[V] + 10 \log\{X_{\Delta}(\mu)\}$$

El factor de atenuación constante  $V$  se disocia en una suma de la señal deseada  $X_{\Delta}(\mu)$ . El uso de

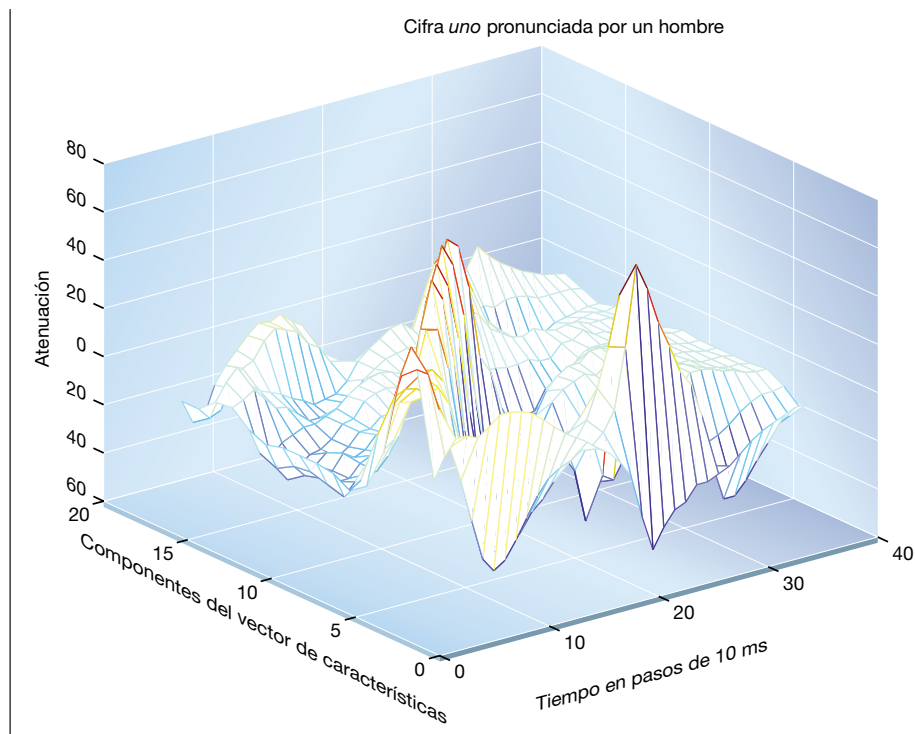


Figura 7  
Secuencia de vectores de características de la pronunciación de la cifra uno en inglés.

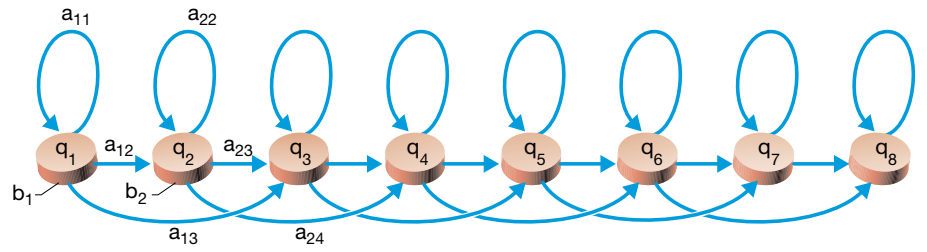
parámetros diferenciados es un medio eficaz de suprimir el término aditivo del factor de atenuación  $V$ . La energía media de todas las frecuencias centrales se sustrae y añade como componente número dieciséis. Otras etapas de normalización conducen al vector de características final  $Y_k$ .

La figura 7 muestra un ejemplo del resultado final de extracción de características: la secuencia de vectores de características para la palabra uno en inglés. Los componentes del vector de características 16 se llevan a la secuencia de etapas de 10 ms.

### Búsqueda

Como ya se ha mencionado, la función de búsqueda (o comparación de configuraciones) desempeña el cometido de cotejar, con el vocabulario instruido del reconocedor, la secuencia de vectores de características procedentes de una expresión de usuario. El sistema de reconocimiento del habla de Ericsson emplea modelos de Markov ocultos (HMM) para modelizar la voz humana. Con esta técnica, cada palabra del vocabulario consiste en un HMM. La estructura de

Figura 8  
Estructura de una palabra de vocabulario en forma de modelo de Markov oculto.



un HMM (figura 8) está formada por una cadena de estados (denominada  $q$ ), describiendo cada estado un segmento de la palabra del vocabulario. Por tanto,  $q1$  modela el inicio, y  $q8$  el final de la palabra.

Los estados están conectados con transiciones, que facilitan cambios de estado, dependiendo de las probabilidades de transición,  $a_{ij}$ . A los estados van unidas las probabilidades de emisión,  $b_j$ , que expresan la similitud espectral de un vector de características con un intervalo de tiempo de la referencia. Empezando con el estado inicial,  $q1$ , por el modelo pueden usarse diferentes trayectos, dependiendo de la secuencia de vectores de características entrante. La repetición o supresión de estados permite la adaptación a variaciones en el ritmo de habla del usuario. Una palabra se reconoce si un trayecto por la referencia ha al-

canzado su estado final,  $q8$ , con una probabilidad fiable.

### Futuros MMI de habla para teléfonos móviles

La experiencia obtenida con la primera generación de marcación vocal muestra que es una tecnología lo suficientemente robusta para emplearse en teléfonos móviles. Un vocabulario de 10 nombres fue suficiente como punto de partida, pero existe demanda de vocabularios de marcación mayores. El tamaño preferido difiere mucho entre cada individuo, pero la mayoría de los usuarios desean un vocabulario más grande que el que puede obtenerse en la actualidad.

Hasta ahora la principal aplicación de los sistemas de reconocimiento del habla ha sido la marcación vocal. No obstante, en el futuro habrá

Figura 9  
Minicasco telefónico inalámbrico Bluetooth.



más funciones que se controlarán con la voz. Por ejemplo, para activar directamente una función que en otro caso exigiría una secuencia de pulsaciones en el teclado dentro de una estructura de interfaz de menú en capas.

La simplificación de los menús mediante la voz es una idea muy atractiva, pero aún es más importante un interfaz hombre-máquina (MMI) que no obligue al usuario a depender de feedback visual o de una complicada interacción en el teclado cuando haya dado una orden de voz. Una ilustración clara de este aspecto es el uso de teléfonos móviles en un entorno de manos libres. Además de la utilización actual de equipo de manos libres en automóviles, esta forma de uso pronto se realizará con minicascos telefónicos basados en tecnología Bluetooth (figura 9).

La utilización de un minicasco telefónico permitirá que el usuario pueda conservar el teléfono en su bolsa o maletín. Bastará que como mínimo las funciones telefónicas básicas (aceptar o efectuar una llamada y el manejo de la lista de teléfonos) puedan controlarse desde el minicasco. El usuario de manos libres dependerá del feedback acústico como complemento de las palabras pronunciadas. En consecuencia, las funciones tales como establecimiento de llamada, perfil de conmutación y grabación de mensajes son apropiadas para control por la voz, mientras que otras tales como un examen de la agenda o el servicio de mensajes cortos (SMS) no lo son. Mediante una cuidadosa selección de funciones, el MMI de habla se convertirá en algo inestimable para el usuario móvil.

La tecnología todavía es relativamente elemental, debido a los recursos limitados de los teléfonos móviles. En el futuro aumentarán el poder computacional y el tamaño de la memoria, permitiendo aplicar tecnología más sofisticada y cómoda para el usuario, por ejemplo, el reconocimiento de palabras interconectadas. También se ha sugerido que la tecnología de reconocer la voz se use para aplicaciones que exijan grandes vocabularios (más de 10.000 palabras), tal como el dictado de cartas-e. Aun cuando los recursos en los teléfonos móviles mejorarán, puede suceder que la memoria necesaria para albergar reconocedores de voz complejos nunca quepa en un teléfono. Sin embargo, una solución podría ser combinar los sistemas para funciones telefónicas de uso frecuente basados en terminal, con reconocedores más complejos basados en la red.

## Conclusión

A pesar de los retos excepcionales que supone el entorno del usuario móvil, en muchos teléfonos celulares ya se emplean sistemas de reconocimiento del habla robustos. Los MMI de habla serán cada vez más importantes a medida que aumente el tamaño de las pantallas de los aparatos y sus teclados sean cada vez más pequeños. Aun cuando las restricciones del hardware limitarán durante algún tiempo el tamaño de los vocabularios incorporados, los usuarios pueden confiar en que el futuro traerá nuevos avances en funcionalidad.