Entrevista

Xuedong Huang, director mundial de tecnologías del habla de Microsoft

«La próxima revolución será la de la voz»

«The next revolution will be the speech revolution»

Susana Arasa

En el campo de la informática aplicada al lenguaje queda mucho terreno por investigar. De momento, según Xuedong Huang, director mundial de tecnologías del habla de Microsoft, hay que recurrir a la aplicación de estándares y al uso de aplicaciones multimodo.

HAL 9000, el protagonista de *2001: Una odisea en el espacio*, el mítico ordenador inteligente que habla y piensa podría encontrarse más cerca de lo que creemos. A juzgar por las explicaciones y demostraciones técnicas de Xuedong Huang, director mundial de tecnologías del habla de Microsoft, falta muy poco tiempo para que la *criatura* de Stanley Kubrick cobre vida real en nuestros escritorios. El advenimiento de *HAL* se producirá con la llegada de Windows XP, la más reciente versión del sistema operativo de dicha empresa.
Con el lanzamiento de este sistema operativo, que incluye funciones de control de mandatos, dictado y conversión de voz a texto, la compañía de Bill Gates ha demostrado que entre sus prioridades también figura la de que los ordenadores hablen y escuchen. De momento sólo disfrutarán de las nuevas posibilidades que ofrece Windows XP quienes utilicen el inglés, el chino y el japonés, y en todo caso con resultados muy discretos.

El peso del error

«La tecnología se encuentra lejos de la perfección», reconoce Huang. Y prosigue: «La voz no es como un teclado, pues se basa en la modalidad. Siempre se producirán errores en el reconocimiento de la voz, en la comprensión o simplemente al convertir la voz a texto».
Para este ingeniero electrónico, que desde el año pasado dirige la unidad *Speech.net* del gigante del software, «el habla es, probablemente, el modelo de interacción más consistente de todos». Sin embargo, su enorme complejidad plantea grandes retos a los desarrolladores. Entre ellos, destaca la gestión de errores.
Los sistemas de reconocimiento de voz suelen enfrentarse a tres tipos de errores: la omisión de palabras, la inserción de incorrecciones no emitidas y la sustitución de una palabra por otra. A estos errores hay que sumar la dificultad que dichos sistemas experimentan ocasionalmente para discriminar el ruido ambiental y, por supuesto, la ignorancia ante términos o construcciones no contenidos en sus gramáticas.
«Existen muchas formas de corregir estos errores, pero el uso de aplicaciones multimodo que combinen la voz con otros tipos de interfaces es probablemente la mejor manera de solventarlos», señala Huang.

Hacia el estándar

Otro de los aspectos que dificultan el despegue de las tecnologías de reconocimiento del habla es la ausencia de una interfaz de programación (API) estándar, reconocida realmente por toda la industria. «El uso de estándares permite que los motores de voz de distintos fabricantes puedan trabajar con aplicaciones de otras empresas, de modo que los desarrolladores pueden centrarse en lo que debe hacer su aplicación», defiende Huang.
En la actualidad, los modelos de interfaz más extendidos son JSAPI *(Java Speech API)*, de Sun Microsytems; SRAPI *(Speech Recognition API)*, de Novell y SAPI *(Speech API)*, de la propia Microsoft.

Existe también un estándar para aplicaciones de telefonía denominado Voice XML *(VXML)*, impulsado por el consorcio W3C. Sin embargo, en opinión de Huang, VXML tiene un futuro incierto. «No es exactamente una aplicación para la Web y las aplicaciones de voz del futuro deben ser multimodo.»
En este sentido, Microsoft ya está considerando un nuevo estándar que pueda soportar múltiples dispositivos y multimodalidad. Este desarrollo incluye el PC, los teléfonos inteligentes y los convencionales. Huang opina que se trata de un modelo muy consistente en cuanto a programación para la Web, que «a diferencia de Voice XML, no exige aprender un nuevo lenguaje».

Un mercado con futuro

A pesar de la gran cantidad de obstáculos que deberán vencer, las aplicaciones de voz cuentan con un excelente futuro. La consultora Cahners In-Stat Group, por ejemplo, espera que el mercado del software de reconocimiento del habla se multiplique por diez en los próximos cuatro años, pasando de un volumen de negocio de 200 millones de dólares a 2700 millones en 2005.
Por razones de reducción de costes, los sectores más propensos al uso de aplicaciones del habla han sido tradicionalmente los mercados verticales científicos y profesionales. El experto de Microsoft asegura que ello se debe a que los programas de dictado resultan muy útiles en los escenarios en los que las manos y los ojos están ocupados: «En estos casos, lo único que se quiere es hablar». Sin embargo, lo que dota de alta eficacia a estas aplicaciones en tales contextos es «la escasa ambigüedad del lenguaje profesional, que contiene mucha estructura y facilita su proceso por parte de los ordenadores».
Xuedong Huang asegura que con el lanzamiento de Windows XP se popularizará el uso de las aplicaciones de voz y se abaratará su precio. Su mercado natural pasará a ser el de los millones de usuarios que debido a minusvalías en la vista tienen dificultades de accesibilidad a las interfaces gráficas de usuario (IGU) y que representan entre un 10 % y un 15 % de la población. A ellos se sumarán los usuarios con escasa formación técnica. Aunque en este segundo caso no se cuenta con cifras, sí se sabe que uno de los principales frenos a la popularización de la informática es la complejidad para interactuar con ella por vías tradicionales.
Huang añade a la lista de entornos idóneos para la interacción con la voz otros segmentos del mercado, como el hogar, el automóvil o los teléfonos móviles, situaciones en las que las manos están ocupadas o no existe espacio físico para una interfaz.

¿Bajo control?

En la actualidad, recuerda Huang, las aplicaciones de reconocimiento del habla se dividen en tres grandes tipos: las que se destinan al control de mandatos del ordenador, las que permiten el dictado y los sistemas de conversión de texto a voz, también denominados *Text To Speech o TTS.*
Las aplicaciones más veteranas están representadas por los motores de control de mandatos por voz, cuyos primeros prototipos permitían en la década de los sesenta ejecutar órdenes simples, como mover las fichas de un tablero de ajedrez, por ejemplo. La reciente *suite* ofimática Office XP de Microsoft incluye funciones avanzadas de control por voz que permiten ordenar acciones en las diversas aplicaciones que la forman, como abrir archivos, seleccionar texto, cambiar formato o activar macroinstrucciones.
Estos sistemas trabajan con gramáticas que definen sucesos aunque según Huang suelen ser excesivamente rígidas: «En ocasiones rechazan frases por una pequeña diferencia sobre los parámetros establecidos, lo que acarrea graves problemas de utilización». Sin embargo, afirma que estas aplicaciones tienen futuro en áreas como los juegos o la edición de textos, combinadas en este segundo caso con las de dictado.
Las gramáticas más extendidas en el diseño de interfaces de reconocimiento de voz son las denominadas CFG (*Context-Free Grammar)* y N-Gram. Huang recomienda el uso de alguna de ellas, el empleo de nombres de mandatos intuitivos para los usuarios y el trabajo con gran cantidad de información real sobre ellos. Y advierte: «Lo difícil no es identificar la gramática, sino disponer de un asistente que ofrezca una buena cobertura de la casuística de los usuarios».

La dictadura del dictado

Estas gramáticas pueden solucionar las necesidades de sistemas de control de mandatos, cuyos

requisitos son limitados. Sin embargo, las aplicaciones de dictado exigen una funcionalidad superior. El dictado, explica Huang, obliga a sofisticados desarrollos, capaces de «convertir el habla en el texto que nos gustaría ver de forma escrita». Esta aparente obviedad comporta un gran trabajo sobre la pronunciación y la presentación de los datos, entre otros aspectos.
También implica funciones de corrección de errores y automejora basadas en modelos de lengua, modelos acústicos y el *feedback* del usuario. «En la actualidad lo que estamos haciendo [en Windows XP] es ofrecer un módulo de reconocimiento basado en la expresión habitual, que permite personalizar las diferentes formas de expresarse que tiene la gente», revela Huang.
El producto para el dictado de Microsoft ha sido adaptado al inglés, al chino y al japonés, tres mercados lingüísticos que, sumados, acumulan el 60 % de la demanda mundial de equipos PC. Según el representante de la compañía, el sistema de dictado en inglés todavía resulta más lento que el empleo del teclado, pero en las lenguas orientales la voz es el doble de rápida.
De momento, la empresa de Redmond no tiene previsto ampliar esta función a otros idiomas, aunque se ha comprometido a facilitar herramientas de desarrollo para que otros fabricantes puedan desarrollar sus motores para Windows XP en otras lenguas.

Máquinas parlantes

Para que la interacción por voz con el ordenador sea absoluta debe ser bidireccional. No basta con que las máquinas nos entiendan cuando les hablamos; también tenemos que poder oír sus respuestas. Se trata de que hablen.
Este diálogo no es algo nuevo. Las operadoras han utilizado tradicionalmente sistemas de *output* de texto a voz para ofrecer servicios telefónicos como información sobre la hora o los números de teléfono de los abonados. Sin embargo, según Huang, «la mayoría de estos sistemas se basan en audio grabado porque la calidad de los TTS todavía no es óptima».
Comienzan a proliferar ahora los portales de voz como *Tell Me*, que permiten a los usuarios llamar por teléfono para obtener información personalizada o leer el correo electrónico basado en web desde un terminal telefónico cualquiera. «Se trata esencialmente de una extensión de los sistemas IVR *(Interactive Voice Response)*, utilizando un vocabulario básico, más flexible, y una adaptación a la web», describe Huang.
Los usuarios solicitan la información específica que desean escuchar (sólo deporte o una determinada cotización, por ejemplo). «Es algo así como una radio personalizada», afirma Huang. Sin embargo, de momento, siguen ofreciendo estos servicios mediante audio grabado y no con sistemas de conversión de texto a voz (TTS).
Microsoft también trabaja en algo parecido, si bien de momento sus esfuerzos se centran en los servicios basados en la pantalla. «Para tener una experiencia de usuario más interesante se sigue necesitando una pantalla», asegura Huang.

Somos humanos

Tales avances no pueden hacer olvidar a los diseñadores de interfaces de voz que los usuarios, como seres humanos que son, también tienen sus limitaciones. «No poseemos superpoderes», advierte Huang. «Tenemos un sistema audiovisual concreto y unos límites modales.»
Según este experto de origen chino que en la actualidad trabaja en los laboratorios de Microsoft en Redmond, no siempre que el ordenador dice algo al usuario éste es capaz de entenderlo y memorizarlo. Además, recuerda, «está comprobado que en la comunicación oral el ser humano sólo puede retener una secuencia de hasta siete órdenes». Este dato, unido a la característica unidimensional de la voz, debería aconsejar la creación de interfaces de voz que no superen esa cantidad de mandatos.
Otro de los errores habituales en el diseño de interfaces de voz consiste en pretender que el usuario se adapte a los esquemas de funcionamiento del ordenador. «Con todas esas gramáticas estamos forzando a que la gente hable el lenguaje de la máquina y no al revés», denuncia Huang, que sugiere resolver este tipo de problemas adoptando una «visión clientecéntrica» ajustada al modelo mental, la tarea y el momento de uso del consumidor.
En opinión de Huang, los diseñadores de interfaces deben tener siempre presentes las peculiaridades del habla: no requiere espacio real en la interfaz gráfica, puede ser utilizada a distancia, es muy expresiva y se desarrolla en una secuencia temporal. Estas características únicas influyen positiva y

negativamente en las aplicaciones de voz y convierten en muy importante el uso del multimodo. «La voz y las interfaces gráficas tienen fuerzas y debilidades complementarias, de modo que la combinación de ambas tecnologías es el procedimiento más efectivo», añade Huang.

El sistema MiPad

Este convencimiento ha movido a Microsoft a invertir en el desarrollo de MiPad, un dispositivo inalámbrico que combina funciones de voz y pulsación sobre una pantalla sensible al tacto, algo que el fabricante denomina *tap and talk*. A través de este sistema se puede suministrar una información muy importante para el reconocimiento de voz y la comprensión, ahorrando tiempo de ejecución. «Si queremos enviar un correo electrónico, marcamos con el lápiz el campo PARA: y damos el nombre hablando», señala Huang. Con esta operación, se activa el sistema de voz y, a la vez, se le indica al sistema que queremos introducir un nombre de persona o una dirección de correo. «Es una información de contexto crucial para el reconocimiento de voz», explica Huang.
En la actualidad, las funciones de lápiz y de habla de MiPad son independientes y funcionan por separado. El micrófono sólo se conecta en el momento de dar órdenes habladas; si no fuera así, el sistema consumiría la batería al tener que lidiar de forma permanente con el ruido del entorno. Huang cree que esta tecnología aprovecha lo mejor de cada sistema de forma complementaria: «El lápiz resulta muy útil para la manipulación directa de acciones simples aunque precisa emplear la mano y los ojos. La voz permite manipular acciones mucho más complejas. No es tan precisa pero a cambio no requiere tanta atención».

Conclusión

A pesar de los indudables progresos conseguidos por las tecnologías del habla, los ordenadores todavía no pueden comunicarse con los seres humanos de forma oral normalizada. Lo más cercano a esta meta es la actual combinación de tecnologías de voz e interfaz gráfica, como las que describe Huang, aunque reconoce que sus funciones son todavía muy limitadas. Con todo, estamos más cerca que nunca de dar el salto. «En su día el PC lo cambió todo y ahora le va a llegar el turno a una nueva revolución, la de la voz», declara este científico de Microsoft. En este 2001, *HAL* llama a nuestra puerta.

Xuedong Huang

Director general del equipo de Tecnología del Habla de Microsoft desde 1993. Xuedong y su equipo han desarrollado trabajos tecnológicos de productos de Microsoft como Office XP o Windows XP. Los trabajos no se basan únicamente en los componentes del lenguaje hablado, sino que incluyen trabajos de cómo proveer una solución multimodal en ambientes móviles. Es profesor asociado de ingeniería eléctrica de la Universidad de Washington y profesor honorario de ciencia computacional de la Universidad Hunan. Coautor de dos libros: *Hidden Markov Models for Speech Recognition* (1990) y *Spoken Language Processing* (2001). En este último libro trata aspectos relacionados con los avances y las técnicas sobre ciencia computacional, ingeniería eléctrica y producción y percepción del habla, entre otros.
xueh@microsoft.com