Los sistemas integrales completos del habla, del lenguaje y la interfaz humana

Comprehensive basic systems for speech, language and human interfaces

Alex Waibel

Esbozo sobre el momento actual en el que se encuentra la investigación en el campo del habla desde el punto de vista del reconocimiento de los mecanismos e interfaces empleados en el habla y de la tecnología de estos sistemas. Su autor, Alex Waibel, apunta varios modelos en los que ha investigado, resaltando los principales obstáculos de la investigación y sus aplicaciones.

The article outlines the current status of speech research seen from viewpoint of speech recognition mechanisms and interfaces currently in use for speech applications and their technology. The author, Alex Waibel, mentions various models on which he has been researching, emphasizing the main obstacles hampering research and applications.

A menudo medimos el porcentaje de errores en los sistemas de reconocimiento del habla, pero olvidamos que lo que realmente importa es la comunicación entre humanos. Al fin y al cabo, estamos intentado desarrollar máquinas que comuniquen con los humanos. Pero a las personas de la calle que desconocen este tipo de trabajos les resulta difícil la comprensión de nuestro trabajo.

– ¿A qué se dedica usted?
– Al reconocimiento del habla.
– ¡Ah! Se refiere a ordenadores que hablan.
– No, no; a ordenadores que reconocen mi habla.
– Ya, claro. Se refiere usted a una máquina que le entiende y ejecuta lo que le pide.
–¡Ya me gustaría que fuera así! Me refiero a máquinas que reconocen la palabra hablada y que son capaces de imprimirla.

En el terreno de la investigación nos centramos básicamente en la técnica. Hablamos de comunicación y no de transcripción porque es la interacción mas eficaz entre humanos y máquinas. Así que se trabaja en superar las actuales barreras de la comunicación. Una de ellas es la movilidad; por ejemplo, con nuestros ordenadores podemos trabajar en la oficina o en casa, pero cuando paseamos o conducimos es imposible realizar este tipo de tareas porque no es práctico. Hablar es mucho más natural. Otras dificultades de la comunicación se perciben al trabajar con la tecnología informática entre varios países debido a las barreras lingüísticas y el factor tiempo.
Dado que existen grandes compañías que ofrecen ordenadores que reconocen el habla, el público cree que todos los problemas están resueltos, pero no es así. Para el reconocimiento del habla es preciso contar con una interfaz entre el ser humano y el ordenador (fig. 1). No se trata sólo de introducir datos en el ordenador para que resuelva un problema; también buscamos una interacción, un nuevo entorno del diálogo.
Por añadidura, existen facetas del habla que no están disponibles de manera interactiva, por ejemplo el registro de un historial clínico o determinados datos específicos que el ordenador tiene que poder gestionar. Y, finalmente, lo más difícil: la interacción entre humanos perfeccionada por el ordenador. Es decir, que cuando un ordenador observe una interacción entre dos personas y se le faciliten palabras clave, un mecanismo pueda reconocer y emprender determinadas acciones, como hacer un café, por ejemplo.

Factores que influyen en el reconocimiento del habla

En los últimos años se han llevado a cabo importantes progresos en este campo, pero existen determinados parámetros o factores que afectan al resultado del sistema.

Ruido. Los fabricantes de aplicaciones señalan que en el 99% de los casos el sistema funciona; pero seguro que podemos encontrar un caso en el que este porcentaje es inferior. Por ejemplo, en los sistemas de dictado aparece el factor ruido. Cuando una persona adquiere uno de estos sistemas, con su CD y su micrófono, el fabricante intenta hacerle creer que el micrófono es de regalo. Sin embargo no se trata de un acto de gratitud sino que sin ese micrófono específico que reduce el ruido exterior el sistema no funcionará. Existen reverberaciones, interferencias o micrófonos determinados (bidireccionales u omnidireccionales) que afectan al resultado del dictado, sin olvidar las reverberaciones que se pueden producir en el canal de transmisión.

Amplitud del tamaño de la terminología. Con un vocabulario de sólo 10 o 20 palabras todo funciona bien. Si nos enfrentamos a un vocabulario más amplio pero sencillo, donde el reconocimiento de la gramática sea fácil tampoco aparecerán problemas. Y si la variedad de opciones que tiene que reconocer el sistema es baja, la tasa de errores correspondiente también será baja.

Capacidad de generar confusión. Si aparecen problemas de claridad en la expresión, o tenemos una mezcla de dígitos o atendemos a problemas de ortografía, la tasa de errores se aproximará al 90 %. La ortografía de muchas palabras se presta a confusión, por ejemplo a la hora de dictar una letra P y diferenciarla de una B. El motivo por el que la amplitud de vocabulario hace que la situación sea más compleja es porque implícitamente hay palabras susceptibles de confusión. Por ejemplo, en un vocabulario de 20 000 palabras en inglés, una de cada dos palabras se diferenciará de otra por un fonema, aunque su pronunciación sea similar. Además, a mayor amplitud de vocabulario, mayor dificultad para recordar la palabra, lo que no ocurre cuando se trata de dígitos. Por último, no hay que olvidar que irán surgiendo nuevas palabras en el idioma que no son reconocidas por el sistema del ordenador.

La variabilidad del orador. Prácticamente todos los sistemas de reconocimiento del habla son independientes del orador que lo utiliza, siendo el sistema el que se adapta al orador. Normalmente, estos sistemas son utilizados por personas adultas que conocen los sistemas comerciales de reconocimiento del habla. Sin embargo, a medida que introducimos estas tecnologías en nuevos grupos, como niños o personas de la tercera edad, aparecen factores que deben ajustarse a las nuevas necesidades. En la actualidad se concentran importantes esfuerzos en diseñar tecnología que ayude a mejorar la calidad de vida de las personas de edad avanzada. Todo ello sin dejar de lado el acento del orador: un español o un francés hablando en inglés tendrán acentos diferentes.

El estilo. Considerado antes como un tema aislado, en la actualidad se considera como una de las cuestiones más difíciles. En el habla espontánea, al pronunciar una frase o expresión sin premeditación se cometen muchos errores. Dado que el sistema tiene dificultades para trabajar con una frase fragmentada en una conversación entre dos personas, el reconocimiento del habla resulta bastante difícil. ¿Cómo se pueden resumir las dificultades que afectan al resultado del sistema? El vocabulario aumenta en función de los diferentes registros y por lo tanto aumenta la dificultad. Todo parece indicar que a medida que aumenta el tamaño del vocabulario se incrementa la tasa de errores, pero no es así. La dificultad del vocabulario es uno de los factores, pero hay que añadir otros, como la confusión generada o la ortografía.
Los sistemas de reconocimiento son muy sencillos. No saben nada del contexto cultural, no conocen el significado de lo que cuenta el orador y, lo único que hacen es juzgar las frases desde el punto de vista acústico. Todo ello ilustra la gran dificultad del reconocimiento de una conversación entre humanos por parte de una máquina. La tasa de errores se sitúa entre el 40 y el 45 %.
DARPA *(Speech Programs: Development of the State of the Art)*, una entidad dedicada al estudio de reconocimiento del habla, realiza cada año una valoración de los mejores resultados obtenidos en cada una de las diferentes tareas del reconocimiento del habla. Según los datos presentados, las tasas de errores aumentan, pero en un análisis detallado se detecta una ligera mejora una vez que la tarea tiene un rodaje.
Se seleccionan tareas que presentan cada vez mayor dificultad y complejidad, por ejemplo el dictado o el diálogo. En una conclusión rápida, podríamos señalar con un símil que el reconocimiento del habla en la actualidad es como una cebolla. Se supera una dificultad, es decir, quitamos una capa de

la cebolla, y enseguida aparece otra capa en forma de otro problema imprevisto y más difícil de superar que el anterior.
En definitiva, todavía existen grandes desafíos en este campo, así que nos centramos en una parte de la ciencia informática en la que se pueden obtener medidas objetivas, fáciles de generar y computar, de modo que se aprecian grandes progresos.
Pero en otros, el progreso no es tan perceptible, como en el caso de nuevas palabras. Cuando establecemos un sistema de valores de referencia, las nuevas palabras figuran como errores y lo que hacen los diseñadores es realizar un vocabulario lo bastante grande como para que dichas palabras puedan resultar legibles. En lo tocante a la semántica, el problema es que no la podemos medir, aunque se hayan realizado algunos esfuerzos iniciales.
Falta integrar otras modalidades de comunicación humana a la aplicación de la tecnología: la voz, los gestos, el lenguaje corporal, la escritura manual u otras. El habla es un elemento muy importante, pero la riqueza de la interacción humana trasciende la voz y hay que tener en cuenta todos esos elementos en la interacción. El lenguaje natural, escrito y corporal, es otro elemento a integrar.

Diferentes sistemas, diferentes problemas

En la actualidad se trabaja con varios modelos, desde el dictado hasta la interacción entre personas en la que el ordenador actúa de observador. Pero existen numerosos problemas tecnológicos sobre cada uno de estos modelos.

Dictado. En este modelo una persona trabaja con un micrófono situado cerca de la boca para evitar la entrada de ruido. Consiste en la lectura de un texto y se puede decir que es el mejor tipo de habla que se puede obtener. Pero también existen problemas, entre ellos el vocabulario, porque no se puede predecir el tipo de lenguaje que se utilizará.
Por ejemplo, si tenemos una carta comercial, no será la misma para todas las empresas y habrá que realizar un trabajo de adaptación a cada una de las situaciones. En cuanto a diccionarios y modelos lingüísticos, este sistema utiliza un vocabulario amplio, entre 60 000 y 100 000 palabras, y lo que tenemos que hacer es preocuparnos por cómo utilizan el sistema los humanos.
Pero incluso en este sencillo modelo, todavía subsisten problemas para mejorar el índice de corrección de errores. Por ejemplo, supongamos que el usuario quiere colaborar y se plantea construir un texto. Si producimos un documento y lo hacemos tecleando, tardamos un tiempo determinado y obtenemos un texto con errores de tipografía. En el caso del dictado por voz tardaremos menos tiempo, pues es evidente que hablar es más rápido que teclear. Pero, con el uso de la voz, tanto el reconocedor como el orador cometerán errores, que serán más difíciles de corregir y requerirán más tiempo.
Si para la corrección de un error volvemos a utilizar la voz, leyendo más despacio y acentuando la articulación, sólo se conseguirá distorsionar el reconocimiento de la voz. En cambio, si se deletrea o se teclea, mejora el rendimiento de la recuperación. Esto explica por qué muchas personas que compran un sistema de dictado acaban por no utilizarlo para producir documentos.

Interacción humana. En este caso se trata de una persona que lee correctamente ante un micrófono adecuado en un estudio en el que el nivel de ruido es muy reducido. En este caso también nos encontramos con variaciones en diferentes segmentos; por ejemplo, no es posible expandir el vocabulario.
El problema es que las noticias son dinámicas, es decir, el vocabulario cambia, y en consecuencia, se mantendrán los errores. Por ejemplo, hace una década nadie conocía la palabra ciberespacio, y este tipo de casos requiere una actualización. Si se trabaja con un margen de error del 30 %, el contenido queda bien representado para entender la información.
La tecnología actual permite realizar aplicaciones de extracción de información. Hoy en día, si se añaden palabras, se puede conseguir una adaptación de forma dinámica. Para ello se emplean documentos web que contengan temas relacionados con la palabra concreta que buscamos. Un recurso muy utilizado para este objetivo son las páginas web de los periódicos. Estas palabras capturadas de Internet permiten generar una nueva lista que se añade a un diccionario temporal; pero también hay que introducir la pronunciación. En el caso de los textos en español resulta bastante

fácil ajustar ortografía con fonemas, algo que no ocurre con el inglés. Finalmente, se genera un nuevo diccionario de pronunciación para las nuevas palabras y se modifican los modelos del lenguaje para que haya una cierta probabilidad asociada.
En nuestros trabajos tenemos un reconocedor en las noticias y en la trascripción aparecen algunos errores debido a las palabras nuevas, pero cuenta con un número suficiente de palabras para poder identificar el tema de la noticia.
En Alemania, desde hace tres años, trabajamos con un sistema de reconocimiento de voz para la televisión. Este sistema escucha las noticias cada noche, y se actualizan los diccionarios por Internet que se almacenarán en la base de datos; se trata de un modelo aplicable también a las reuniones. Así, al día siguiente, preguntamos al ordenador por alguna información y éste nos ofrecerá la grabación realizada. Es un modelo en el que los errores son aceptables y pueden ser controlados.

Diálogo hombre-máquina. Este modelo se aleja de las tareas que van en una única dirección y busca la respuesta del sistema. Es por ello que las condiciones de grabación durante la interacción tienen un papel muy importante; por ejemplo que el micrófono esté situado cerca de quien habla o que se reduzca al mínimo el ruido.
En general, existen sistemas de diálogo con muy buenas condiciones de grabación con el micrófono cerca de quien habla, algo que no ocurre cuando intentamos aplicarlo por ejemplo a un automóvil, ya que se interpone el ruido del motor u otros ruidos de fondo que impedirán una buena grabación.
Existen una serie de ejemplos en los que este sistema funciona bastante bien, y son aquellos en los que existen temas de objetivo muy limitado, como conseguir un saldo bancario con las aplicaciones de diálogo por teléfono. En este campo algunos vendedores de sistemas de diálogo han descubierto que estos sistemas pueden conseguir muy buenos resultados, no por la precisión del reconocimiento sino por la cognición. Es decir, en un sistema de diálogo el objetivo es realizar alguna actuación o recuperar información que podría ser bastante amplia e implicaría dificultades en el caso de realizarlo una persona.
Pero algunas de las cuestiones que no debemos pasar por alto son, por ejemplo, los micrófonos remotos que implican una degradación del sonido, la espontaneidad del hablante o la gestión y control del diálogo.
En el modelo de diálogo se debe tener en cuenta que estamos añadiendo un estrato superior al reconocimiento del habla. ¿Cómo se gestiona la interacción persona-máquina, y cómo se concluye la tarea? Lo más normal sería que el usuario pudiese recorrer un menú de opciones por tonos diferentes. Por ejemplo, si el usuario busca alojamiento, el número 1 correspondería a reservar y el 2 a información. Pero se trata de un procedimiento engorroso. Existe un sistema que va formulando las preguntas en busca de una repuesta concreta, bien afirmativa o negativa, que tarda menos tiempo. O también se puede establecer un diálogo libre en el que el usuario sólo dé la respuesta afirmativa ante lo que necesita, pero implica una dificultad del reconocimiento y además el usuario algunas veces se puede confundir en la respuesta. En cierta manera, es una iniciativa mixta, donde el sistema da libertad al usuario y la máquina le guía por el diálogo hasta obtener el objetivo deseado. El problema es que al hablar se hacen comentarios y también se abusa del sistema, con lo que el diálogo se vuelve más complejo. Hacer un diálogo de empatía, en el caso de los humanos, es una cuestión social en la que se aporta información personal para generar una cierta amistad; afortunadamente esto no se da entre el ser humano y la máquina: sabe que tiene que ser directo y no hace falta ser amable con ella.
Partiendo de estos supuestos se pueden desarrollar unos sistemas de diálogos que proporcionen este tipo de orientaciones. Existen planteamientos diferentes; el más popular es el que se basa en la gramática. Las gramáticas generan la comprensión de una frase determinada y con ello intentan predecir la respuesta correcta.
Otros métodos se basan en teorías del diálogo que permiten una introducción de datos más libres en la pregunta. En ellos el sistema, en un marco semántico, podría formular una pregunta al usuario para que incluya la información que le falta. Se trata de métodos que requieren una preparación de una gramática que permita abordar el diálogo; esta gramática puede ser compleja, con técnicas de aprendizaje reforzado para optimizar la estrategia del sistema. Pero el diálogo no se establece sólo por el habla; hay que contar también con el dibujo o la señalización. Por ejemplo, marcar en un mapa la distancia entre dos puntos. Se pueden desarrollar estrategias basadas en un diálogo que implique el habla, la escritura y los gestos; posteriormente, se podrían corregir las deficiencias.

Por ejemplo, se puede encontrar ya en el mercado este tipo de sistemas para su uso durante la conducción, pero todavía requieren de una programación y del uso de algunos mandos que no se pueden utilizar mientras se conduce. Realizamos un trabajo para probar la consulta de datos desde el coche mediante la voz. La tecnología que permite estas consultas es similar a la que describimos en este apartado. Se trata de un reconocedor de gran amplitud de vocabulario, una base de datos; no es sólo una herramienta de reconocimiento, sino también un marco semántico con un diálogo que permite soluciones. El problema era el ruido. Habíamos conseguido una amplia base de datos con tipos de ruido (conducción, motor, carretera, circulación y fenómenos meteorológicos, entre otros). Además, descubrimos que si el micrófono iba adherido al cinturón de seguridad, los resultados eran casi tan buenos como en un reconocedor de voz; pero, por diferentes motivos, tuvimos que trabajar con micrófonos remotos en el retrovisor, que recogían ruidos ambientales. Por último, aparece otro problema: los cambios en las pautas de la voz humana cuando se está en una situación de estrés; en consecuencia, no encaja con las pautas embebidas en el sistema. Es decir, hablar al micrófono a distancia o los ruidos adicionales en el canal, son dos tipos de ruido que se tienen que gestionar.

Hombre-máquina-hombre. Hoy por hoy es imposible entender todos los idiomas del mundo y el inglés, con distintos acentos, se convierte en el idioma de la interacción, donde el ordenador actúa como mediador. Sin embargo, pese a lo que se podría pensar, en el año 2020 el idioma más común en Internet no será el inglés sino el chino. Es muy importante contar con un enfoque equilibrado respecto al multilingüismo (en la actualidad existen en el mundo entre 4000 y 6000 idiomas).
Una aplicación en la que trabajamos es la traducción con «una especie de intermediario»: se trata del asistente lingüístico portátil. Este aparato ayuda a la navegación para poder encontrar, por ejemplo, la manera de llegar a un lugar determinado. Además, nos puede facilitar información sobre un monumento o puede realizar la traducción de un diálogo al hablar con una persona que no entiende nuestro idioma.
Respecto al multilingüismo, si tenemos un reconocedor de un idioma y queremos pasar a otro, debemos –entre otras tareas– desarrollar diccionarios, y esto se puede llevar a cabo con reconocimiento de voz o mediante traducción automática. Es una tarea en la que se está invirtiendo mucho tiempo, incluso décadas. Si se trata de un idioma hablado por poca gente, la tarea para el aprendizaje y los problemas de traducción se complican, porque nadie fuera de la comunidad habla dicho idioma.
Y para el reconocedor que permite traducir de un idioma a otro existen una serie de pasos importantes a seguir. El primero es el procesamiento de la señal; es decir, se hacen transformaciones matemáticas, a excepción de idiomas tonales, como el chino. En cuanto al modelo lingüístico, se puede adecuar si existe texto que se puede obtener de Internet, y aquí surge la dificultad de que no existen páginas en Internet en el idioma que queremos utilizar. Respecto al diccionario de pronunciación, normalmente es una tarea que se realiza «a mano» pero en la actualidad se está intentando generar diccionarios de pronunciación automáticos, que escuchan cómo habla la gente y luego cartografían la ortografía y la pronunciación. En cuanto a modelos acústicos, necesitamos una base de datos que requiere decenas de miles de frases recogidas en ese idioma. El tema gráfico se puede reducir a los caracteres romanos, pero el problema es cómo segmentar el texto para que tenga sentido y dotarlo de consistencia natural.
Podemos desarrollar un reconocedor de voz, intentar descubrir un nuevo modelo lingüístico acústico más eficaz. Se trata de un proceso laborioso, durante el cual desarrollamos amplias bases de datos multilingües con modelos acústicos multilingües. En este punto hay que tener en cuenta que la noción de palabra no es igual en todas las lenguas; por ejemplo, en japonés y chino es una secuencia de caracteres. Se puede desarrollar un modelo multilingüe que es un modelo acústico y después someterlo a prueba con una de las lenguas; el resultado es mejor que un modelo acústico que cubra todas las lenguas, un modelo más compacto y difícil de manipular.

Hombre-hombre. En este apartado el ordenador tiene el papel de observador. Se trata de proveer de un acceso rápido a los participantes de reuniones. En este proyecto, *Meeting Browser*, el papel del ordenador es el de un archivo con todas las informaciones que se expresan, incluyendo la identificación de los participantes en la reunión. Por ello abarca diferentes tareas, como la transcripción del discurso, la grabación y localización de los participantes en la reunión o la

recuperación de todas las ideas expresadas en la misma.
Este sistema se centra en cuatro componentes principales: un sistema de transcripción del habla, el resumen o una herramienta estadística para localizar los momentos en los que hay un cambio en la mesa de los interlocutores y un identificador del discurso; por último, incorpora una estructura visual, además de incluir un sistema de vídeo que permite grabar imágenes de las personas participantes y sus puntos de atención.

Alex Waibel

Profesor en la School of Computer Science en la Universidad Carnegie Mellon en Pittsburg y en el Departamento de Computer Science de la Universidad Karlsruhe en Alemania. Director del Interactive Systems Laboratories en los citados centros. Fue uno de los fundadores del consorcio C-STAR para la investigación de traducción del habla. Sus áreas de investigación se centran entre otros campos en las interfaces multimodales, el reconocimiento del habla o traducción automática. Ha publicado artículos en numerosas revistas especializadas en el campo de las tecnologías.
waibel@cs.cmu.edu