Los nuevos desarrollos en inteligencia artificial han conseguido poder identificar los patrones de la voz humana en el aire y convertirlos en información textual y cadenas de bits procesables. La base de ello son los nuevos algoritmos de aprendizaje automático y una mayor capacidad de procesamiento de datos. Sin embargo, ¿es suficiente? La comunicación oral, como veremos, requiere muchos datos además de la voz, pero la voz es una primera base para ese camino.

Antes de aceptar la propuesta del SMS surgieron serias dudas: ¿por qué las personas iban a querer escribir algo pudiendo llamar por teléfono y hablar? ¿Tenía sentido? (Hillebrand, 2010). Finalmente se apostó por esta tecnología iniciando una era de intercambio de mensajes escritos que heredarían progresivamente el correo electrónico, las redes sociales y las aplicaciones de mensajería, haciendo que la oralidad, la voz, perdiera cada vez más relevancia en las comunicaciones en la red[1].

Hoy, gracias a los algoritmos de aprendizaje automático (*machine learning*) y a una mayor capacidad de cómputo[2], la voz vuelve otra vez a recuperar terreno. Pero irónicamente no tanto para hablar más con otros humanos sino como nuevo interfaz para interactuar con máquinas. Y no únicamente entre humanos y máquinas, sino también como forma de comunicación de las máquinas entre sí.

La creación de lenguajes que sirvieran para comunicarse con la máquina se podría decir que es, en sí, la ciencia de la computación y es consustancial al desarrollo de los primeros computadores, colaborando desde el principio lógicos especialistas en el lenguaje como Gödel. E igualmente, desde los inicios, se ha tratado de que esa comunicación sea lo más próxima al lenguaje humano[3].

Ya en los años 80 se popularizan a nivel lúdico —que nunca debe ser subestimado dado su impacto comercial en el desarrollo de la informática— los juegos conversacionales (Montfort, 2003) y la comunicación de las máquinas empleando tonos de voz. Era habitual en las redes telefonía para hablar entre ellas y configurar la red. En concreto, el empleo de este tipo de tonos de voz suponía una vulneración de seguridad importante y había humanos que, haciéndose pasar por máquinas y hablando como ellas, conseguían engañar a otras máquinas. Vulnerabilidad que llegó a obligar a replantear toda la estandarización mundial de comunicaciones (Boone, 2011).

Hoy, sin embargo, hemos pasado desde aquella época en que conversar con un vehículo daba para argumento de series de televisión de ciencia ficción a ser un complemento de serie en todos los vehículos. Y también en nuestra casa y en los dispositivos personales.

Gracias a los algoritmos de aprendizaje automático y a una mayor capacidad de cómputo, la voz vuelve otra vez a recuperar terreno

Las máquinas hoy pueden, a diferencia de los módems antiguos, hablar entre ellas empleando palabras

humanas y solicitarse servicios mutuamente (Debin, 2015), o mantener cortos debates[4]. Incluso uno de los servicios más sorprendentes en el ámbito del *Internet of Things* de Telefónica es la posibilidad de transmitir la voz de un técnico para mantener y reparar elevadores. Más aún, cada vez es más habitual que interfaces de voz hagan de traductores no solo entre lenguajes humanos en tiempo real sino también entre humanos y máquinas, traduciendo las palabras del operador/programador a los comandos que requiere la máquina para ejecutar la acción.

¿Es correcto pues hablar de inteligencia artificial? ¿Qué es lo que falta a la conversación con uno de estos interfaces que hace que a menudo no sea amigable? ¿Es robótica esta voz inteligente o está más cercana a la humana?

La clave está en el contexto (Garten, Kennedy, Sagae, y Dehghani, 2019). Veamos unos ejemplos:

1) A: Cuesta doce veces más hacer un cliente nuevo que mantener a uno antiguo (Escandell, 2005).

2) A: ¿Conducirías un Mercedes?
B: No conduciría ningún coche caro.

3) A: ¿Qué tienes pensado hacer hoy?
B: Tengo un terrible dolor de cabeza.

4) -Leo vendió un cuadro a Pedro
-Pedro compró un cuadro a Leo[5].

Interpretar estas frases es una acción trivial para cualquier humano, ¿pero serían igualmente entendibles por una máquina? La primera expresión puede corresponder a un profesor transmitiendo un dato a sus alumnos o a un cliente lanzando una amenaza velada. ¿Percibiría Siri el nivel de la amenaza o considerará que se trata solo de un dato que queremos que aprenda?

El segundo caso, ¿interpretaría la inteligencia artificial que no queremos conducir el Mercedes?

Para el tercero, ¿qué significa exactamente esa respuesta? ¿Que se va a quedar en casa, que no tiene ganas de ocio, que piensa hacer algo como ir al médico, …?

El cuarto, dependiendo de que usemos una u otra expresión estaremos subrayando uno u otro matiz. ¿Sería una inteligencia artificial capaz de captarlo? Si se lo hemos enseñado previamente sí; si no, es imposible. Una inteligencia artificial suele desenvolverse en un cierto contexto, pero no domina el casi infinito registro de contextos que domina un ser humano.

Por ejemplo, podemos diseñar una inteligencia artificial de atención al cliente que pueda interpretar el primer caso como una amenaza pero, si situamos de pronto esa misma inteligencia artificial en un aula, sería incapaz de procesar correctamente lo que sucede de la misma forma que lo haría un ser humano.

El ámbito de la comunicación lingüística en la que se basa la voz siempre ha oscilado entre dos tendencias: una primera que se focalizaba en la gramática y en el lenguaje en sí, con su lógica asociada —destacaríamos a Russell, el primer Wittgenstein y a Chomsky, por citar algunos— y en los estudiosos del entorno en que se produce esa conversación, la ubicación de ese contexto: la llamada pragmática.

Los primeros modelos de comunicación se basaron en los mismos que se empleaban precisamente para la

transferencia de información entre máquinas[6] y que aún se estudian en el bachiller. En ellos, el hablante tiene una imagen mental de lo que quiere decir, lo codifica mediante su voz y se lo hace llegar al receptor mediante un canal, que a su vez vuelve a decodificar estas palabras. Pero, en realidad, las personas no funcionamos así: paralelo a este proceso de codificación y decodificación hay un proceso de inferencia en el receptor en el cual asocia lo que decodifica a su situación de contexto. Esto es lo que se conoce como Teoría de la Relevancia y es donde nuestras máquinas encuentran problemas (Sperber y Wilson, 2012).

El principio de relevancia es la base de la comunicación verbal básica con las máquinas. Por ejemplo, a la hora de buscar en Google el buscador nos responde, en principio, por el orden de lo que entiende más relevante. Para ello se basa en el carácter epistemológico de la fuente de información y los datos personales del que hace la pregunta que ha conseguido colectar[7]. Sin embargo, no es lo mismo tratar de buscar una información determinada que mantener un diálogo. En un diálogo se trabaja en común entre emisor y receptor para compartir y acordar una información: actuamos mediante actos de habla[8]. La máquina debería tener esa misma flexibilidad para cualquier situación en un todo coherente.

Estamos solo en el inicio del empleo de la voz entre hombres y máquinas: nuestro siguiente gran reto es dotar a las máquinas de capacidad de contextualizar y entender la relevancia

Hay formas de paliar esto con mayor o menor resultado. Por ejemplo, es posible dotar a las máquinas de entendimiento sobre contextos haciéndoles aprender sobre los mismos mediante el uso de guiones predefinidos[9].

En algunas máquinas, para reforzar esta ausencia de contexto, se añade al reconocimiento de la voz la capacidad de reconocer imágenes o un reconocimiento de patrones más profundo. De ese modo, las máquinas pueden percibir la emoción del rostro del que habla, la emoción que tiene por el tono de voz o el uso del lenguaje e incluso constantes fisiológicas como el pulso o la respiración que enriquecerían la información sobre su estado emocional. El Internet de las Cosas puede dotar además de información sobre el entorno físico en el que se encuentra la inteligencia artificial.

Existen además capacidades para argumentar y debatir, principalmente mediante búsquedas en fuentes de cultura general como Wikipedia. Lo que es otra forma de dotar a nuestras máquinas de conocimiento de contexto. Esta técnica, así como el hecho de que existen patrones comunes, se emplea para detectar otra peculiaridad de la voz humana difícilmente gestionable por máquinas: la ironía[10] o las fórmulas poéticas como las metáforas o analogías (Nehaniy, 1999).

La importancia del contexto en las comunicaciones de voz llega al punto de ser considerada por la inteligencia militar. ¿Cómo detectar las fuentes de propaganda o radicalismo online de una forma óptima y certera? La can- tidad de información en Internet es colosal y fallar en la identificación de patrones es problemático[11].

Como vemos, estamos solo en el inicio del empleo de la voz entre hombres y máquinas: nuestro siguiente gran reto es dotar a las máquinas de capacidad de contextualizar y entender la relevancia. Esto supone un desafío tanto tecnológico —no existe capacidad de proceso actual como para dotar a las máquinas de esta habilidad—, como matemático—mejores algoritmos de aprendizaje de estos contextos— o lingüístico—¿es la actual Teoría de la Relevancia el marco que mejor sirve para trabajar las cuestiones de la voz humano-máquina?—. Y pese a ello terminamos con el argumento de Searle[12], el filósofo que entendió el lenguaje como una forma de actuar: identificar patrones y generar respuestas de una forma muy rápida, ¿de verdad puede llamarse inteligencia? En cualquier caso, cuando Searle planteó su famoso argumento, conocido como “la habitación china”, frente al test de Turing, no pareció tener en cuenta la cuestión del contexto.