La comunicación humano-máquina es hoy más natural y efectiva

Las tecnologías de la voz no son algo nuevo. En 1952 se construyó el primer reconocedor automático de voz, que identificaba dígitos. El auge en los últimos años de las interfaces de voz en el mercado se debe a mejoras sustanciales como la reducción de errores de comprensión en estas tecnologías. Pero ¿qué retos quedan por resolver para que podamos interactuar con la tecnología de manera más natural?

[ILUSTRACIÓN: LAURA PÉREZ]

El auge actual de los dispositivos controlados por voz, como los altavoces inteligentes, hace que hoy más que nunca se hable de las tecnologías del habla. Estas tecnologías, desde hace ya unas décadas, nos permiten interactuar con las máquinas a través de la voz. Como toda tecnología, en sus inicios tenía limitaciones que impedían que la comunicación hablada fuera más natural y eficiente. Ahora, con los últimos avances tecnológicos en aprendizaje automático e inteligencia artificial, muchas de las limitaciones han sido resueltas. Esto ha supuesto una mejora en aplicaciones de automatización de procesos, respuestas a preguntas frecuentes e incluso, la interacción con nuestro entorno físico; son aplicaciones que rompen la barrera digital entre muchas personas que lo necesitan y la tecnología gracias a la interacción por voz.

Pero empecemos por entender dónde se aplican las tecnologías del habla. Crear una buena experiencia conversacional requiere entender de teorías de la comunicación. Entender cómo hablamos y cuál es el objetivo, analizar las técnicas que usamos cuando nos comunicamos, para después llevar ese aprendizaje a modelos y algoritmos.

Así, el proceso de una interacción por voz comprende cinco fases. En la primera, la de reconocimiento automático de voz (ASR, *Automatic Speech Recognizer* en sus siglas en inglés), el mensaje hablado es recibido y transcrito a palabras gracias a modelos lingüísticos que realizan esa traducción de fonemas a caracteres. En la segunda fase, la de comprensión del lenguaje natural (NLU, *Natural Language Understanding*), el objetivo es entender en esas palabras escritas la intención del emisor. Además, se extraen elementos im- portantes o entidades que aportan valor al mensaje. Aquí no solo se analizan las palabras morfosintácticamente sino que la semántica y la pragmática entran en juego. En la tercera, la de gestión del diálogo (DM, *Dialog Management*), se decide qué hacer o responder y se busca información para formar una respuesta si es necesario. Además, para una comunicación efectiva, es importante tener en cuenta el contexto de la conversación y saber con quién hablamos. En la cuarta etapa, la de generación de respuesta (RG, *Response Generation*), se forma una frase que tenga sentido con la información a responder. Por último, en la etapa de síntesis de voz (TTS, *Text to Speech*), de la frase de respuesta se genera una respuesta de audio con una voz sintética, basada en modelos fonéticos.

Las tecnologías del habla que aplican a cada una de las etapas anteriores han visto su evolución a lo largo de las décadas, salvando parte de las limitaciones para esa comunicación efectiva entre humano y máquina.

La clave para el éxito de todo el proceso de comunicación hablada ha sido la mejora en el reconocimiento de voz, la primera de las etapas. Esto se debe a que un error de reconocimiento conlleva una sucesión de

malentendidos en cascada en las siguientes fases. Entender los problemas y limitaciones que han sufrido las tecnologías del habla es muy útil para comprender el alcance y la velocidad con la que las interfaces de voz se han posicionado en los últimos años y su posible evolución e impacto.

El primer reconocedor automático de voz, llamado *Audrey*, fue construido por AT&T Bell Labs en el año 1952. Su objetivo era identificar por teléfono dígitos del 0 al 9. El sistema reconocía con una exactitud del 90 por ciento los fonemas de personas concretas —como la de su creador, HK Davis—, pero no servía para reconocer la voz de cualquiera.

No fue hasta 1971, casi veinte años más tarde, cuando estos sistemas fueron capaces de reconocer frases. El grupo de investigación DARPA[1], del Departamento de Defensa de los Esta- dos Unidos, apostó por crear un fondo para el desarrollo del reconocimiento del habla, con el objetivo de entender como mínimo un vocabulario de 1.000 palabras en inglés. Fue entonces cuando nació *Harpy*, en 1976. Fue el primer sistema que utilizó modelos de lenguaje para determinar qué secuencias de pa- labras tenían más sentido juntas y reducir así errores en el reconocimiento de voz. De hecho, fue una de las primeras aplicaciones de técnicas estadísticas basadas en modelos ocultos de Markov[2].

En esta década, esta tecnología se empezó a aplicar en sistemas automatizados de respuestas interactivas (IVR, *Interactive Voice Response*), en centros de llamadas de teléfono, que permitían a los interlocutores navegar por menús de servicios a través de la voz.

En la década de los noventa, gracias a *Dragon*, la tecnología de reconocimiento del habla llegó al mercado con productos de dictado. La aplicación reconocía habla continua a un ritmo de cien palabras por minuto, aunque previo a su utilización, era necesario entrenar el modelo con cada hablante durante 45 minutos. En la primera década de los años 2000 se estancó la tecnología de reconocimiento, con una exactitud de palabras reconocidas de un 80 por ciento.

En 2010 se empieza a reconocer la voz de un público generalista, sin necesidad de entrenamiento, gracias a modelos basados en ingentes cantidades de datos. Fue Google, entre 2008 y 2012, quien lanzó Voice Search, el precursor de Google Assistant, en dispositivos móviles y navegadores web. Gracias a esto pudo obtener datos de millones de consultas de búsqueda para mejorar el reconocedor de voz y predecir qué decían los usuarios.

En 2012 aparecen los algoritmos de aprendizaje profundo (*deep learning*) basados en redes neuronales. Esta técnica ha permitido resolver retos clave para la efectividad de las etapas de una interfaz de voz. De hecho, uno de estos grandes retos del reconocimiento de voz, resueltos en parte con este problema[3], ha sido separar a diferentes interlocutores en entornos de ruido o música. Este problema es conocido como c*ocktail party problem*[4].

Otro reto, el de la comprensión de la intención del usuario, ha mejorado gracias a técnicas de aprendizaje automático aplicadas a textos. Una técnica común es crear clasificadores de intenciones con frases reales de los usuarios. Un método flexible, pues no depende de la estructura de la frase como las técnicas anteriores más estrictas basadas en gramáticas. Además, entre otros métodos, es útil la representación vectorial de palabras (*word embeddings*). La técnica *word2vec*[5] consiste en crear un espacio vectorial semántico para entender qué palabras están re- lacionadas según el contexto e incluso aplicar operaciones vectoriales (rey + femenino = reina).

Otra de las mejoras es la aplicada a las últimas fases de generación de respuesta y la síntesis de voz. En la primera, diferentes técnicas de generación del lenguaje natural, que extraen patrones sobre cómo nos comunicamos, pueden generar cadenas de texto con respuestas eficientes y cada vez más naturales. En la segunda, la gene- ración de voces sintéticas cada vez más humanas, como el proyecto WaveNet[6] nacido en 2016. Lo más destacable es que en ciertos casos no es necesario tener muchos datos de partida para poder obtener resultados naturales y útiles.

Gracias a tener más datos y mejor capacidad de cómputo en la nube, empresas como Google crearon asistentes de voz de ámbito general. Apple lanza *Siri* en 2011. Le siguen Microsoft con *Cortana* (2014) y Amazon con *Alexa* (2014). En los últimos cinco años, estos asistentes de voz, y otros como *Bixby* de Samsung, se han integrado en el móvil y en infinidad de dispositivos, como los altavoces inteligentes. No solo son más ubicuos sino que son capaces de entendernos mejor cada día. Es el caso de Google, cuyo reconocedor de voz alcanzó un ratio de 95 por ciento de palabras reconocidas con éxito, equiparable a capacidades humanas, en 2017[7].

Sí, hay momentos en los que la interacción exclusiva por voz carece de sentido. Es el momento de apoyarse en las pantallas

Esto ha sido clave para crear un nuevo ecosistema en el que se han desarrollado herramientas para conseguir las primeras aplicaciones para los asistentes de voz. Gracias a esto aparecen nuevos casos de uso donde la voz y la interacción natural con la tecnología es clave para las personas.

Pero, ¿cuándo es útil la voz? Desde el inicio, en las IVRs, la voz se ha utilizado para optimizar procesos y servicios de atención al cliente. El objetivo era el ahorro de dinero. Ahora podemos preguntarle a nuestro asistente de voz favorito por el tiempo, pedirle que ponga música o escuchar algún *pódcast* mientras desayunamos. Aún así, más allá del dispositivo o de las funcionalidades del asistente, la interacción por voz no siempre es útil, no vale para todo. Pensemos en situaciones en las que necesitamos hablar para conseguir algo. Si el entorno nos permite comunicarnos, es un buen caso de uso. Por ejemplo, informarnos del tráfico mientras estamos en el coche conduciendo de camino al trabajo. O mientras estamos en la cocina preparando la comida. Y sí, hay situaciones en las que la interacción exclusiva por voz carece de sentido. Es el momento de apoyarse en pantallas, para que el usuario pueda tener diferentes maneras de interacción según el contexto. Hablamos de una interacción multimodal.

Independiente del caso de uso, no olvidemos que las interfaces de voz rompen la barrera digital que a veces supone la tecnología. Las interfaces de voz, cuando además están bien diseñadas, son inclusivas[8]. Son idóneas para niños, personas mayores, personas con alguna demencia como el Alzheimer, con movilidad reducida, invidentes, analfabetos digitales o incluso para quien no tiene acceso a la tecnología. Y aquí las interfaces aportan mejoras sustanciales en pequeñas situaciones del día a día, ya que aportan independencia al individuo.

Aún así, seguimos investigando y descubriendo. Quedan muchos retos y limitaciones a resolver. La tecnología para la toma de decisiones no sesgadas, inclusiva y ética. Ética en la recogida de datos para entrenamiento de algoritmos[9], y la aplicación de estos algoritmos de tecnologías del habla. Sesgos derivados del entrenamiento con datos poco inclusivos y de los diseños centrados en el usuario estándar, poco representativos de una sociedad multicultural.

Lo positivo es que nos esperan años de mejoras en los algoritmos de aprendizaje automático, que harán que las interfaces de voz entiendan mejor el uso del lenguaje, la semántica y, no menos importante, el contexto de la conversación. Mejoras en la gestión del diálogo, en la toma de decisiones a la hora de responder e incluso en la automatización de tareas. Las interfaces por voz serán realmente asistentes conversacionales ubicuos en diferentes situaciones y casos de uso, que nos ayudarán a tomar mejores decisiones.

Davis, K.; Biddulph, R. y Balashek, S. (1952): “Automatic recognition of spoken digits” en *The Journal of the Acoustic Society of America*, 24(6), 637-642.

Lowerre, B.T. (1976): *The HARPY Speech Recognition System*. Stanford University. Disponible en: https://stacks.stanford. edu/file/druid:rq916rn6924/rq916rn6924.pdf