Tecnologías del habla y sus aplicaciones en la sociedad digital.

Las Tecnologías del Habla son el conjunto de técnicas que permiten la interacción entre personas y máquinas mediante el uso de la voz. Se engloban dentro de las Tecnologías del Lenguaje, las cuales incluyen además el procesamiento automático de lenguaje escrito (tanto compresión como generación artificial). Con el advenimiento de la sociedad digital, dichas tecnologías están cobrando un especial protagonismo, ejemplo del cual es el Plan Nacional de Impulso de las Tecnologías del Lenguaje dependiente de la Secretaría de Estado para el Avance Digital [1].

ASCENSIÓN GALLARDO ANTOLÍN
Profesora Titular de Universidad. Dpto. de Teoría de la Señal y Comunicaciones. Universidad Carlos III de Madrid.

JUAN MANUEL MONTERO MARTÍNEZ
Profesor Titular de Universidad.Dpto. de Ingeniería Electrónica Universidad Politécnica de Madrid.

En este artículo, nos centraremos en particular en los sistemas basados en el lenguaje hablado y sus aplicaciones.

El habla es la forma de comunicación más eficaz, directa y sencilla para los seres humanos, por lo que resulta imprescindible en las aplicaciones de interacción entre personas y máquinas. Sin embargo, el habla no sólo transmite un mensaje lingüístico, sino también información sobre características del hablante, como su identidad, edad, género, o estado emocional, incluso su estado de salud. Por ello se han desarrollado distintas tecnologías para su procesamiento [2], [3]:

- *Reconocimiento automático del habla,* que tiene como objeto la transcripción automática de lenguaje hablado en texto.

- *Síntesis o conversión texto-habla,* que genera mensajes orales inteligibles y naturales a partir de texto, imitando una o más características vocales de una persona: desde su timbre de voz (aquello que hace que su voz sea diferente a la de otros hablantes) hasta su estilo de habla o su estado anímico.

- *Sistemas de diálogo o conversacionales,* capaces de comunicarse con usuarios humanos en lenguaje natural. Generalmente, constan de módulos de reconocimiento y comprensión de habla y de generación de respuesta, así como de un gestor de diálogo, que controla el flujo de las interacciones con el usuario.

• *Identificación de idioma,* que tiene como objeto determinar en qué idioma está pronunciada una determinada frase, antes de reconocer su contenido.

• *Reconocimiento de locutor,* también denominado biometría de la voz, que consiste en la determinación o la verificación de la identidad de una persona mediante el análisis de su voz.

• *Detección de estrés y carga cognitiva,* cuya finalidad es la de determinar el nivel de estrés o carga cognitiva de una persona (es decir, la cantidad de recursos mentales que necesita para realizar una tarea dada), analizando simplemente su voz.

• *Identificación de otras características paralingüísticas del hablante:* la edad, el estado emocional o etílico y enfermedades como la depresión, el Parkinson o la Esclerosis Lateral Amiotrófica (ELA) afectan a la producción del habla, y mediante el uso de tecnologías del habla se puede contribuir a detectarlas o a estimar su nivel de gravedad (en casos previamente dignosticados) y realizar un seguimiento de su evolución temporal. En el futuro se prevé que este tipo de sistemas de vanguardo tengan aplicación en el ámbito de la salud y de la seguridad vial o ciudadana [4].

Desde el punto de vista técnico, son tecnologías multidisciplinares que combinan técnicas de procesado de señal, de inteligencia artificial (en particular, aprendizaje automático) y de lingüística computacional. De hecho, el gran desarrollo experimentado por la inteligencia artificial en los últimos años, llega de la mano de las denominadas redes neuronales artificiales *(Artificial Neural Networks)* y el aprendizaje profundo *(Deep Learning)* [5], [6], y ha producido una extraordinaria mejora en las prestaciones de estos sistemas en escenarios reales. Estas técnicas de aprendizaje profundo se suelen emplear en los módulos más básicos (reconocimiento y síntesis de habla) comunes a múltiples aplicaciones, pero las capas superiores, las que implementan la capacidad de comprensión e inteligencia del sistema, es habitual que sigan empleando mucho conocimiento experto, difícil de reutilizar de un campo de aplicación a otro.

Las Tecnologías del Habla proporcionan canales complementarios a las tecnologías de Internet o a los operadores humanos, para proporcionar servicios de valor añadido. Aunque el nivel de fiabilidad de las tecnologías del habla sea siempre inferior al 100% (debido a la extraordinaria complejidad y variablidad del habla misma), siempre pueden servir de filtro o de apoyo a la toma de decisiones por parte de humanos. En general, su objetivo no es tanto sustituir a especialistas u operadores, como proporcionar herramientas que permitan optimizar el rendimiento de su trabajo, disminuyendo las tareas más rutinarias.

Las Tecnologías del Habla resultan especialmente prometedoras allí donde se necesitan técnicas no invasivas o invisibles al usuario, dado que solo necesitan un micrófono, un altavoz y un procesador (por ejemplo, un teléfono móvil), con acceso a un servicio típicamente disponible en la nube. La invisibilidad es especialmente interesante en aplicaciones médicas o de detección de carga de trabajo, porque en una simple conversación se puede detectar qué casos requerirían un estudio más detallado por parte de un especialista.

Finalmente, entre las cuestiones éticas que afectan a las tecnologías del habla, es importante señalar que su rendimiento puede ser sensible a parámetros del hablante (como su edad, género, grupo étnico, carácter nativo, enfermedades...) o del idioma (los sistemas comercialmente disponibles para idiomas con tantos hablantes como el inglés o el castellano tienen prestaciones algo más elevadas que los de idiomas menos mayoritarios), especialmente, si los desarrolladores no han diseñado cuidadosamente sus sistemas. Por ello, se debe extremar el cuidado durante su evaluación y contratación.

"En general, su objetivo no es tanto sustituir a especialistas u operadores, como proporcionar herramientas que permitan optimizar el rendimiento de su trabajo, disminuyendo las tareas más rutinarias."

ÁMBITOS DE APLICACIÓN

Los sistemas conversacionales *(chatbots)* permiten proporcionar información o intermediar en alguna gestión de atención al público, de asistencia a la ciudadanía o de cita previa. Tradicionalmente, estos sistemas eran diseñados para un idioma concreto, pero actualmente es posible implantar sistemas multilingües con identificación automática del idioma del hablante, que permiten que sea el sistema el que se adapte al hablante (en vez de al revés), especialmente en países con varios idiomas cooficiales como el nuestro.

Los sistemas de *Speech Analytics* extraen información de conversaciones, típicamente telefónicas entre agentes y usuarios: desde el contenido lingüístico de las conversaciones en sí, hasta características propias del usuario, como su edad, dialecto, estado de ánimo, etc. Esto permite analizar el motivo de las llamadas y calcular estadísticas sobre la frecuencia de ciertas consultas o gestiones, analizar el rendimiento de los agentes y su interacción con los usuarios, asegurar la calidad de atención al cliente y, en suma, mejorar el servicio. Originalmente, estos sistemas se desarrollaron para el ámbito concreto de los centros de llamadas *(call centers)*, pero tienen aplicación en cualquier otro servicio de atención e información, como los existentes en la Administración Pública.

Los sistemas de traducción automática habla-habla son capaces de oír y reconocer habla en una lengua, traducir el mensaje a otra lengua distinta y sintetizar habla en la misma. Estos sistemas tienen un interés turístico inmediato en un país como España. Hoy en día es posible incluso clonar características de la voz del hablante original (su estilo de habla o su estado emocional) en una lengua en la que nunca ha hablado.

Los sistemas biométricos de reconocimiento de locutor tienen utilidad en aplicaciones de seguridad para control de acceso a un recinto, para el control horario de entrada y salida y, especialmente, para servicios no presenciales (permitiendo la identificación sin clave o la verificación transparente al usuario). Uno de los principales riesgos de la biometría vocal es la suplantación de identidad *(spoofing)* por medio de mensajes pregrabados o sintetizados del usuario en cuestión. Las soluciones más frecuentes a este problema son los sistemas de reconocimiento *anti-spoofing* o los sistemas multimodales en los que se combinan la voz con otras técnicas alternativas de identificación, como el uso de contraseñas dinámicas o el reconocimiento facial.

Los sistemas de reconocimiento de sobrecarga cognitiva o de estrés son aplicables en situaciones de elevada carga de trabajo o cuando es necesario atender varios estímulos simultáneos (alertas, llamadas telefónicas, etc.) que puedan afectar negativamente al rendimiento de una persona. Por este motivo, la detección automática de los niveles de carga cognitiva y/o estrés tiene aplicación en escenarios reales como son aquellos en lo que están involucrados servicios sanitarios, de emergencia, bomberos y cuerpos y fuerzas de seguridad del Estado.

CONCLUSIONES

En conclusión, pese a las limitaciones actuales expuestas, los recientes avances experimentados por las tecnologías del habla permiten el desarrollo de un elevado número de aplicaciones en distintos ámbitos como atención al ciudadano, seguridad y sanidad, incrementando la presencia de dichas tecnologías en la vida cotidiana de los ciudadanos, y suponiendo, por tanto, un paso más hacia la consolidación de la sociedad digital. ✱

Bibliografía

[1] «Plan de Impulso de las Tecnologías del Lenguaje», 2015. URL: https://www.plantl.gob.es/tecnologias-lenguaje/PTL/Paginas/plan-impulso-tecnologias-lenguaje.aspx

[2] D. Jurafsky y J. H. Martin, Speech and Language Processing, ISBN: 978-0131873216: Prentice Hall, 2008.

[3] B. W. Schuller, Y. Zhang y F. Weninger, «Three recent trends in Paralinguistics on the way to omniscient machine intelligence», Journal on Multimodal User Interfaces, vol. 12, nffl 4, pp. 273-283, 2018.

[4] B. Schuller, «Can Affective Computing Save Lives? Meet Mobile Health», Computer, vol. 50, nffl 5, pp. 13-13, 2017.

[5] A. Graves, A. Mohamed y G. Hinton, «Speech recognition with deep recurrent neural networks», de IEEE International Conference on Acoustics, Speech and Sig-nal Processing (ICASSP), 2013.

[6] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville y Y. Bengio, «Char2Wav: End-to-end speech synthesis», de International Conference on Learning Representations (ICLR), 2017.