¿Qué es el reconocimiento de voz?

El reconocimiento de voz, también conocido como reconocimiento automático de voz (ASR), reconocimiento de voz por ordenador o conversión de voz a texto, es una capacidad que permite a un programa procesar el habla humana en un formato escrito.

Aunque el reconocimiento del habla suele confundirse con el reconocimiento de la voz, este último se centra en la traducción del habla de un formato verbal a otro textual, mientras que el primero sólo trata de identificar la voz de un usuario concreto.

IBM ha desempeñado un papel destacado en el reconocimiento de voz desde su creación, con el lanzamiento de "Shoebox" en 1962. Esta máquina era capaz de reconocer 16 palabras diferentes, lo que suponía un avance del trabajo inicial de los Laboratorios Bell de los años cincuenta. Sin embargo, IBM no se detuvo ahí, sino que continuó innovando a lo largo de los años, lanzando la aplicación VoiceType Simply Speaking en 1996. Este software de reconocimiento de voz tenía un vocabulario de 42 000 palabras, admitía inglés y español, e incluía un diccionario de ortografía de 100 000 palabras.

Aunque la tecnología del habla tenía un vocabulario limitado en sus inicios, hoy en día se utiliza en un gran número de sectores, como la automoción, la tecnología y la sanidad. Su adopción no ha hecho más que acelerarse en los últimos años debido a los avances en el deep learning y el big data. [Las investigaciones](https://www.marketsandmarkets.com/PressReleases/speech-voice-recognition.asp) (enlace externo a ibm.com) muestran que se espera que este mercado tenga un valor de 24 900 millones de dólares en 2025.

[Libro electrónico](https://www.ibm.com/account/reg/signup?formid=urx-52384" \t "_blank)

[Conocer y utilizar Presto](https://www.ibm.com/account/reg/signup?formid=urx-52384" \t "_blank)

[Explore el libro electrónico gratuito de O’Reilly para aprender a dar los primeros pasos con Presto, el motor SQL de código abierto para el análisis de datos.](https://www.ibm.com/account/reg/signup?formid=urx-52384" \t "_blank)

Contenido relacionado

[Regístrese para recibir la guía sobre modelos fundacionales](https://www.ibm.com/account/reg/signup?formid=urx-52620)

Características clave del reconocimiento de voz eficaz

Hay muchas aplicaciones y dispositivos de reconocimiento de voz disponibles, pero las soluciones más avanzadas utilizan [la IA](https://www.ibm.com/es-es/topics/artificial-intelligence) y el [machine learning](https://www.ibm.com/es-es/topics/machine-learning). Integran la gramática, la sintaxis, la estructura y la composición de las señales de audio y voz para comprender y procesar el habla humana. Lo ideal es que aprendan sobre la marcha, haciendo evolucionar las respuestas con cada interacción.

El mejor tipo de sistemas también permite a las organizaciones personalizar y adaptar la tecnología a sus requisitos específicos, desde el lenguaje y los matices del habla hasta el reconocimiento de la marca. Por ejemplo:

* **Ponderación lingüística:** mejore la precisión ponderando las palabras específicas que se dicen con frecuencia (como los nombres de los productos o la jerga industrial), más allá de los términos que ya están en el vocabulario base.
* **Etiquetado de hablantes:** produzca una transcripción que cite o etiquete las contribuciones de cada hablante en una conversación con varios participantes.
* **Formación en acústica:** ocúpese de la parte acústica del negocio. Entrene el sistema para que se adapte a un entorno acústico (como el ruido ambiental de un centro de llamadas) y a los estilos del orador (como el tono de voz, el volumen y el ritmo).
* **Filtrado de palabras malsonantes:** utilice filtros para identificar ciertas palabras o frases y desinfectar la salida de voz.

Mientras tanto, el reconocimiento de voz continúa avanzando. Las empresas, como IBM, están haciendo incursiones en varias áreas, para mejorar mejor la interacción entre humanos y máquinas.

Algoritmos de reconocimiento de voz

Los caprichos del habla humana han dificultado el desarrollo. Se considera una de las áreas más complejas de la informática, ya que involucra la lingüística, las matemáticas y la estadística. Los reconocedores de voz constan de varios componentes, como la entrada de voz, la extracción de características, los vectores de características, un descodificador y una salida de palabras. El decodificador aprovecha modelos acústicos, un diccionario de pronunciación y modelos de lenguaje para determinar la salida adecuada.

La tecnología de reconocimiento de voz se evalúa en función de su tasa de precisión, es decir, la tasa de error de palabras (WER), y la velocidad. Hay una serie de factores que pueden afectar a la tasa de error de las palabras, como la pronunciación, el acento, el tono, el volumen y el ruido de fondo. Lograr la paridad humana, es decir, una tasa de error igual a la de dos personas hablando, es desde hace tiempo el objetivo de los sistemas de reconocimiento de voz. [Una investigación de Lippmann](https://www.ee.columbia.edu/~dpwe/classes/e6820-2005-01/papers/Lipp97-hummach.pdf) (enlace externo a ibm.com) estima que la tasa de error de palabras ronda el 4 por ciento, pero ha sido difícil replicar los resultados de este artículo.

Se utilizan varios algoritmos y técnicas de cálculo para reconocer el habla en texto y mejorar la precisión de la transcripción. A continuación se presentan breves explicaciones de algunos de los métodos más utilizados:

* **Procesamiento del lenguaje natural (PLN):** aunque el [PLN](https://www.ibm.com/es-es/topics/natural-language-processing) no es necesariamente un algoritmo específico utilizado en el reconocimiento de voz, es el área de [inteligencia artificial](https://www.ibm.com/es-es/consulting/artificial-intelligence) que se centra en la interacción entre humanos y máquinas a través del lenguaje y texto. Muchos dispositivos móviles incorporan el reconocimiento de voz en sus sistemas para realizar búsquedas por voz (por ejemplo, Siri) o proporcionar más accesibilidad a los mensajes de texto.
* **Modelos ocultos de Márkov (HMM):**los modelos ocultos de Márkov se basan en el modelo en cadena de Márkov, que estipula que la probabilidad de un estado determinado depende del estado actual, no de sus estados anteriores. Mientras que un modelo de cadena de Markov es útil para eventos observables, como entradas de texto, los modelos ocultos de Markov nos permiten incorporar eventos ocultos, como etiquetas de parte de la oración, en un modelo probabilístico. Se utilizan como modelos de secuencia dentro del reconocimiento de voz, asignando etiquetas a cada unidad, es decir, palabras, sílabas, oraciones, etc., en la secuencia. Estas etiquetas crean un mapeo con la entrada proporcionada, lo que le permite determinar la secuencia de etiquetas más adecuada.
* **N-gramas:**este es el tipo más simple de modelo de lenguaje (LM), que asigna probabilidades a oraciones o frases. Un N-grama es una secuencia de N-palabras. Por ejemplo, "pida la pizza" es un trigrama o 3 gramas y "por favor, pida la pizza" es un 5 grama. La gramática y la probabilidad de ciertas secuencias de palabras se utilizan para mejorar el reconocimiento y la precisión.
* **Redes neuronales :** principalmente aprovechadas para algoritmos de [deep learning,](https://www.ibm.com/es-es/topics/deep-learning) las redes neuronales procesan los datos de entrenamiento de entrada al imitar la interconectividad del cerebro humano a través de capas de nodos. Cada nodo se compone de entradas, ponderaciones, un sesgo (umbral) y una salida. Si ese valor de salida supera un umbral determinado, se "dispara" o activa el nodo, pasando los datos a la siguiente capa de la red. [Las redes neuronales](https://www.ibm.com/es-es/topics/neural-networks) aprenden esta función de mapeo a través del aprendizaje supervisado, ajustándose en función de la función de pérdida a través del proceso de descenso del gradiente.  Aunque las redes neuronales tienden a ser más precisas y pueden aceptar más datos, esto tiene un costo de eficiencia de rendimiento, ya que tienden a ser más lentas de entrenar en comparación con los modelos de lenguaje tradicionales.
* **Diarización de hablantes (SD):**los algoritmos de diarización de los hablantes identifican y segmentan el discurso según la identidad del hablante. Esto ayuda a los programas a distinguir mejor a las personas en una conversación y se aplica con frecuencia en los centros de llamadas para distinguir a los clientes y a los agentes de ventas.

Casos de uso de reconocimiento de voz

Un gran número de industrias están utilizando diferentes aplicaciones de la tecnología del habla en la actualidad, lo que ayuda a las empresas y a los consumidores a ahorrar tiempo e incluso vidas. Algunos ejemplos son:

**Automoción:** los reconocedores de voz mejoran la seguridad del conductor al habilitar los sistemas de navegación activados por voz y las capacidades de búsqueda en las radios de los coches.

**Tecnología:** [los agentes virtuales](https://www.ibm.com/es-es/products/watsonx-assistant) se están integrando cada vez más en nuestra vida diaria, especialmente en nuestros dispositivos móviles. Usamos comandos de voz para acceder a ellos a través de nuestros teléfonos inteligentes, como a través de Google Assistant o Siri de Apple, para tareas, como la búsqueda por voz, o a través de nuestros altavoces, a través de Alexa de Amazon o Cortana de Microsoft, para reproducir música. Seguirán integrándose en los productos cotidianos que utilizamos, alimentando el movimiento de la “Internet de las cosas”.

**Sanidad:** los médicos y los enfermeros utilizan las aplicaciones de dictado para capturar y registrar los diagnósticos y las notas de tratamiento de los pacientes.

**Ventas:** la tecnología de reconocimiento de voz tiene un par de aplicaciones en las ventas. Puede ayudar a un centro de llamadas a transcribir miles de llamadas telefónicas entre clientes y agentes para identificar patrones de llamadas y problemas comunes. [Los chatbots de IA](https://www.ibm.com/es-es/products/watsonx-assistant/artificial-intelligence) también pueden hablar con gente a través de una página web, respondiendo a consultas comunes y resolviendo peticiones básicas sin necesidad de esperar a que un agente del centro de contacto esté disponible. En ambos casos, los sistemas de reconocimiento de voz ayudan a reducir el tiempo de resolución de los problemas de los consumidores.

**Seguridad:** a medida que la tecnología se integra en nuestra vida diaria, los protocolos de seguridad son una prioridad cada vez mayor. La autenticación basada en voz agrega un nivel viable de seguridad.