Sistema de reconocimiento de voz: un enlace en la comunicación hombre-máquina

Voice recognition system: a link in man-machine communication

Kiara Barrios[1], José López[1], Samy Mendieta[1], Rilda Benavides[2], Yessica Sáez[3]*
[1] Licenciatura en Ingeniería Electrónica y Telecomunicaciones - Centro Regional de Azuero - Universidad Tecnológica de Panamá,
[2]Licenciatura en Ingeniería Eléctrica – Electrónica – Centro Regional de Azuero – Universidad Tecnológica de Panamá
[3]Facultad de Ingeniería Eléctrica – Centro Regional de Azuero – Universidad Tecnológica de Panamá

Resumen Un sistema de reconocimiento de voz es la capacidad que presenta un ordenador para recibir los datos de voz de un usuario, transformar la señal en código binario, el cual es asimilado por la computadora y luego establece la comunicación hombre-máquina necesaria para resolver diferentes problemas que requieran para su resolución la utilización de este método. Desde el punto de vista social se emplea como una herramienta útil y necesaria para personas con discapacidad física (carencia en sus extremidades superiores); además de agilizar la búsqueda de información propia o a través de la red para cualquier usuario que posea un ordenador con esta herramienta tecnológica. Esta tecnología podría estar convirtiéndose en un estándar en los nuevos dispositivos, pero su precisión es finalmente lo que determina si realmente se convierte en un recurso sin posibilidad de uso. En este artículo se evalúan y comparan los sistemas de reconocimiento de voz utilizados en dispositivos Android, IOS y el *software* para computadoras Cortana utilizando el sistema operativo Windows. El objetivo de esta comparación es determinar si esta tecnología se ha convertido en una opción práctica y determinar, de las aplicaciones presentadas, cuál sería la mejor opción. Los resultados muestran que el *software* para computadoras Cortana utilizando el sistema operativo Windows presenta la menor tasa de error.

Palabras claves Errores, palabras, frases, reconocimiento de voz, comunicación.

Abstract The voice recognition system is the ability of a computer to receive an user's voice data, transform the signal into binary code, which is assimilated by the computer, and then establish the necessary man-machine comunication to solve different problems that require, for their solution, the use of this method. From the social point of view, it is used as an useful and necessary tool for people with physical disability (lack in their upper extremities), in addition to expediting the search of information of its own or through the network for any user who owns a computer with this technological tool. This technology could become a standard in new devices, but its accuracy is ultimately what determines if it really becomes a resource without possibility of use. This article evaluates and compares voice recognition systems used on Android devices, IOS and Cortana Computer Software using the Windows operating system. The objective of this comparison is to determine if this technology has become a practical option and to determine from the applications presented, what would be the best option. The results show that the Cortana Software has the lowest word error rate.

Keywords Errors, words, phrases, voice recognition, communication.

* Corresponding author: yessica.saez@utp.ac.pa

1. Introducción

Desde sus inicios, el hombre ha sentido la necesidad de relacionarse y comunicarse con el mundo que lo rodea, razón por la cual a través de los años hemos estado en constante búsqueda de aquellos métodos que nos ayuden a suplir dicha necesidad. Hace más de 50 años, en 1940 para ser exactos, se inicia el desarrollo de una tecnología capaz de crear un enlace entre personas y máquinas conocida como "reconocimiento de voz". La misma surge a raíz de querer construir un sistema que hiciera el habla visible a las personas que presentaban problemas auditivos [1]. En un principio este sistema se basó en dispositivos mecánicos que poco a poco fueron evolucionando a dispositivos electrónicos y que hoy día son sistemas complejos desarrollados gracias al avance creciente en el campo de la informática.

Un sistema de reconocimiento de voz no es más que la capacidad que posee un ordenador, de convertir, las palabras de la voz humana a un código binario comprensible por la computadora [1]. Los primeros sistemas creados estaban basados en conversaciones de tipo discretas, en las cuales se utilizaba un lenguaje verbal puntuado por pausas y con un límite de palabras que no excedía de las 50 [2]. Años después se comienza a trabajar en un sistema de conversaciones continuas que no requería que el usuario realizara pausa entre palabras, esto debido a que se emplearon técnicas para minimizar la diferencia en la velocidad del habla [2]. Pasa el tiempo y con él se crean nuevas y mejores técnicas que hacen que los sistemas de reconocimiento de voz tomen un papel mucho más importante dentro de las grandes industrias conocidas. En un principio su utilidad estaba destinada al sector salud, cuya función era la de reemplazar las tradicionales transcripciones médicas y a su vez mantener un orden en las citas médicas de cada paciente [1]. Además de este sector, otros sectores también se ven beneficiados con dicha tecnología, entre los cuales se pueden mencionar el sector comercio, artículos como computadoras, celulares y automóviles inteligentes, domótica, telecomunicaciones, servicios militares y muchos más [1].

Sin embargo, aún con todos los avances tecnológicos, lograr alcanzar que los sistemas de reconocimiento de voz trabajen a la par de los humanos sigue siendo una tarea difícil. Generalmente, los humanos pierden una o dos palabras de cada 20 que escuchan. En una conversación de cinco minutos, eso podría ser como 80 palabras en promedio. Aunque para la mayoría los seres humanos esto no representa un gran problema, debemos considerar lo difícil que es esta tarea para una computadora.

Una vez mencionado lo anterior, el objetivo de este proyecto no es más que evaluar y comparar diferentes sistemas de reconocimiento de voz mayormente utilizados en la actualidad. Con esta comparación se espera determinar si esta tecnología se ha convertido en una opción práctica y de las aplicaciones presentadas, cuál sería la mejor opción.

2. Marco teórico

Un sistema de reconocimiento de voz es el encargado de establecer la comunicación entre los humanos y las computadoras con el desarrollo de aplicaciones capaces de reconocer diferentes voces en la medida de que el habla sea de manera natural. Aunque tiene muchas dificultades, ya que no se cuenta con un sistema que resuelva los problemas existentes relacionados con la variabilidad de las características de la señal acústica, estos sistemas se pueden clasificar de acuerdo a las siguientes restricciones [3], [4].

Pueden ser dependientes o independientes del locutor, en el caso de que sea dependiente, la probabilidad de que se dé un buen reconocimiento de la voz es alta, ya que las muestras que va a tener el reconocedor solo pertenecerán a una persona en particular facilitando que se reconozca el vocabulario. En el caso de que sea independiente del locutor, o sea que es un sistema que puede reconocer la voz de cualquier persona, existe cierta dificultad debido a que las representaciones paramétricas de la voz dependen altamente del locutor [3], [4].

Cuando el locutor habla lentamente, las palabras son más aisladas una de otras y esto da como resultado que la probabilidad de reconocimiento del sistema sea mayor, en cambio cuando el locutor habla de forma natural, realizando las pausas correspondientes que cada frase requiere, el grado de dificultad aumenta, ya que se hace difícil identificar donde inicia y donde termina una palabra, y si están muy juntas cabe la posibilidad de que se confundan con una sola [3], [4].

La extensión del vocabulario es importante, ya que en la medida de que vaya creciendo así también va aumentado la dificultad de reconocimiento, o bien surgen nuevos problemas como lo es cuando se confunde una palabra por otra o cuando el sistema tarda más en reconocer las palabras, etc. También se presenta la similitud entre palabras puesto que si tenemos un vocabulario extenso puede haber palabras que se parezcan aumentando la dificultad de reconocimiento y la búsqueda realizada para encontrar la palabra correcta tiene mayor probabilidad de que sea errónea [3], [4].

En cuanto al ruido que es un factor que afecta en gran medida el nivel de desempeño del reconocedor, puede ser producido por el ambiente, por parte del locutor, música, etc. Y también varía la eficiencia de acuerdo al estado de ánimo del locutor, la calidad del micrófono, entre otros factores [3], [4].

En uno de sus mensajes, la capitalista de riesgo de Silicon Valley, Mary Meeker, en su informe anual de Tendencias de Internet señala que: "la entrada de voz tiene el potencial de ser la forma más eficiente de computación: los seres humanos pueden pronunciar 150 palabras por minuto en promedio, pero sólo puede escribir 40. También agregó que ahora es el momento para que el reconocimiento de voz asuma el control, ya que la tecnología es un ajuste lógico con dispositivos conectados a Internet de cosas, como Amazon Echo o Apple Watch" [5]. Sin embargo, lo que impide que el reconocimiento de voz se convierta en una forma dominante de la informática es su falta de fiabilidad.

A pesar de los impresionantes avances en los últimos años, alcanzar el nivel de rendimiento humano en las tareas de inteligencia artificial como el reconocimiento de voz sigue siendo un reto científico. De hecho, los estándares de referencia no siempre revelan las variaciones y complejidades de los datos reales.

El WER (*Word Error Rate*) es una forma de medición utilizada principalmente en sistemas de reconocimiento automático del habla [6]. El WER es una valiosa herramienta para comparar diferentes sistemas de reconocimiento de voz, así como para evaluar las mejoras dentro de un sistema. El mismo consiste en comparar una referencia con una hipótesis y se define mediante:

$$WER = (S + B + I)/N \quad (1)$$

En donde,
S es el número de sustituciones
B es el número de borrados
I es el número de inserciones
N es el número de palabras de referencia

3. Materiales y métodos

Para realizar este proyecto de investigación se utilizó como base de estudio sistemas de reconocimiento de voz existentes en dispositivos Android, IOS y el sistema operativo Windows, utilizando las siguientes aplicaciones, respectivamente:

- S Voice: el cual es un *software* de reconocimiento de voz desarrollada por Samsung, disponible en sus dispositivos móviles. Es su asistente personal móvil virtual capaz de ejecutar una gran cantidad de tareas a través de comandos de voz por sí solo para ahorrar tiempo y esfuerzo. Actualmente está disponible en varios idiomas [7].
- Siri: una aplicación con funciones de asistente personal para iOS, macOS, tvOS y watchOS [8]. Esta herramienta ha sido desarrollada para contar con su propia personalidad y utiliza procesamiento del lenguaje natural para responder preguntas, hacer recomendaciones y realizar otras acciones mediante la delegación de solicitudes. Actualmente, es considerado uno de los mejores del mundo.
- Cortana: es un asistente personal capaz de realizar diversas tareas a través de comandos de voz. Este *software*, disponible en varios idiomas, es una de las principales competencias de Siri. Este *software* fue creado por Microsoft para Windows, iOS, Android, entre otros [9].

La metodología a utilizar en las pruebas fue la siguiente:

- Realizar repeticiones seguidas de diferentes frases para observar el comportamiento del sistema de reconocimiento de voz ante las mismas.
- Hacer una comparación de los tres sistemas de reconocimiento de voz utilizados, así como evaluar la cantidad de errores o fallas que presenten una vez realizado lo estipulado anteriormente.

Las frases utilizadas en la experimentación fueron las siguientes:

- Frase 1: "Donde reina el amor sobran las leyes" (7 palabras).
- Frase 2: "El hombre nunca sabe de lo que es capaz hasta que lo intenta" (13 palabras).
- Frase 3: "La sabiduría no ejerce ninguna autoridad y aquellos que ejercen la autoridad no son sabios" (15 palabras).

Estas frases fueron seleccionadas aleatoriamente. Para todas las frases y para todos los sistemas de reconocimiento en estudio, se utilizó una voz femenina procurando que las frases o muestras del reconocedor pertenecieran a una persona para facilitar que se reconociera mejor el vocabulario. Cada frase fue repetida diez veces, realizando una pronunciación considerablemente adecuada a una velocidad normal con las pausas necesarias. Se obtuvo un promedio de las variables S, B e I de la Ecuación (1). Luego estos valores promedio fueron utilizados para obtener el WER promedio en cada frase. Todas las pruebas fueron realizadas en un aula de clases apartado, procurando tener la menor cantidad de ruido posible.

4. Análisis y resultados

Luego de realizar varias pruebas en los diferentes dispositivos se obtuvieron los resultados mostrados en las tablas 1, 2 y 3, mostradas a continuación.

Tabla 1. Resultados de reconocimiento de voz en Android

	FRASE 1 (N=7)	**FRASE 2 (N=13)**	**FRASE 3 (N=15)**
X	10	10	10
S	0.2	1	1.4
B	0	0	0.3
I	0	0	0
WER	0.0286	0.0786	0.1086

Los datos de la tabla 1 fueron obtenidos utilizando un Samsung Galaxy S6 a través de la aplicación S Voice.

Tabla 2. Resultados de reconocimiento de voz en Windows

	FRASE 1 (N=7)	**FRASE 2 (N=13)**	**FRASE 3 (N=15)**
X	10	10	10
S	0	1	2.2
B	0	0.1	0.4
I	0	0.6	0.1
WER	0	0.0538	0.18

Los datos presentados en la tabla 2 se obtuvieron a través de una computadora portátil ASUS, utilizando el asistente de reconocimiento de voz Cortana.

Tabla 3. Resultados de reconocimiento de voz en iOS

	FRASE 1 (N=7)	FRASE 2 (N=13)	FRASE 3 (N=15)
X	10	10	10
S	1.5	0.6	1.9
B	1.6	9.4	3.9
I	0	0	0.2
WER	0.443	0.714	0.4

Los datos mostrados en la tabla 3 se obtuvieron a través de un iPhone 6 mediante el asistente de reconocimiento de voz Siri.

Los resultados obtenidos son bastante aceptables para la cantidad de intentos que se hicieron en dicho experimento, ya que entre más intentos los resultados se estarían acercando a los datos proporcionados por las estadísticas de los fabricantes. Podemos observar que el sistema de reconocimiento de voz utilizado en el iPhone 6, es el que presenta el mayor WER promedio para todas las frases. Este resultado es bastante sorpresivo, ya que este asistente personal, el cual es el más utilizado de Estados Unidos, se encuentra catalogado como uno de los mejores del mundo. Estos resultados pueden tener diferentes razones: el acento, el tono de voz, el ruido de fondo, entre otros. Además, consideramos que para juzgar la fiabilidad de estos sistemas, el simple reconocimiento de los palabras no es suficiente, ya que para tener cualquier nivel de efectividad, los sistemas necesitan ser capaces de distinguir entre homófonos (palabras con la misma pronunciación pero significados diferentes) y aprender nuevas palabras y nombres propios.

5. Conclusiones

Los sistemas de reconocimiento de voz han ido mejorando a través de los años. Hemos observado y comprobado que uno de los mejores reconocedores de voz es el de Microsoft en el sistema operativo Windows, ya que la tasa de error se aproxima bastante al que nos da el fabricante, por lo que se puede decir que es muy aceptable a la hora de reconocer la voz humana.

El reconocimiento de voz todavía tiene muchos defectos y limitaciones que se basan principalmente en deficiencias de la inteligencia artificial. Podemos decir que la tecnología que poseen los sistemas de reconocimiento de voz estudiados en este artículo, funciona más que nada como un traductor de comandos determinados, ya que las computadoras no pueden filtrar el contexto o la motivación de las órdenes. Quizás este es el motivo principal por el que Siri no obtuvo los resultados esperados, pues generalmente siempre se utilizan frases como "Siri, llama a Jorge", por ejemplo. Al mismo tiempo, hablar de procesamiento de lenguaje es más fácil decirlo que hacerlo. Lo que indica la literatura y lo que comprueban los resultados en este trabajo, es que a las computadoras les es difícil procesar múltiples frases y reconocer los comandos fácilmente. Por otro lado, consideramos que la mayoría del *software* de reconocimiento de voz tiene que ser configurados para funcionar correctamente. Los mismos deben adecuarse al tono de voz del locutor para que pueda reconocer las órdenes y comandos que el mismo le ordena. Sin embargo, se espera que en un futuro el *software* de reconocimiento de voz, sea una parte integral de las computadoras, no solo en las industrias, sino también dentro de nuestros hogares (domótica, Internet de las cosas).

Como trabajo futuro esperamos realizar estudios comparativos más profundos con una mayor cantidad de repeticiones en los experimentos. Además, esperamos que en asignaturas como Procesamiento Digital de Señales podamos darle continuidad a este estudio y programar nuestro propio sistema de reconocimiento de voz.

REFERENCIAS

[1] [Online] Disponible en: http://www.articulosinformativos.com/Reconocimiento_de_Voz-a963743.html

[2] Lumenvox. (s.f.). Obtenido de La Historia de la Tecnología del Reconociemiento de voz: http://www.lumenvox.com/espanol/resources/tips/historyOfSpeechRecognition.aspx

[3] Puebla, U. d. (s.f.). Sistemas de reconocimiento y síntesis. *Tesis Digitales*. Obtenido de http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/ahuactzin_l_a/capitulo1.pdf

[4] C.L. Pablo, J. M. (s.f.). *La tercera revolución: comunicación, tecnología y su nomenclatura en inglés*.

[5] [Online] Disponible en: https://www.inc.com/tess-townsend/mary-meeker-says-voice-search-is-going-to-be-huge.html

[6] *T. Martin* (s.f.). Obtenido de World Error Rate Calculation: https://martin-thoma.com/word-error-rate-calculation/

[7] [Online] Disponible en: http://www.samsung.com/global/galaxy/what-is/s-voice/

[8] [Online] Disponible en: https://www.apple.com/ios/siri/

[9] [Online] Disponible en: https://www.thestreet.com/story/12534433/1/why-cortana-assistant-can-help-microsoft-in-the-smartphone-market.html