https://doi.org/10.37815/rte.v33n2.840
Artículos originales

Evaluación del reconocimiento de voz entre los servicios de Google y Amazon aplicado al Sistema Integrado de Seguridad ECU 911

Evaluation of voice recognition between Google and Amazon services applied to the ECU 911 Integrated Security System

Juan José Peralta Vásconez[1], Carlos Andrés Narváez Ortiz[1], Marcos Patricio Orellana Cordero[1] https://orcid.org/0000-0002-3671-9362, Paúl Andrés Patiño León[1] https://orcid.org/0000-0001-9504-6498, Priscila Cedillo Orellana[1, 2] https://orcid.org/0000-0002-6787-0655

[1]*Universidad del Azuay*, Cuenca, Ecuador
jjperalta@es.uazuay.edu.ec,
carlos.05.narvaez@es.uazuay.edu.ec, marore@uazuay.edu.ec,
andpatino@uazuay.edu.ec, icedillo@uazuay.edu.ec

[2]*Universidad de Cuenca*, Cuenca, Ecuador
priscila.cedillo@ucuenca.edu.ec

Enviado: 2021/07/09
Aceptado: 2021/09/28
Publicado: 2021/11/30

Resumen

El reconocimiento automático de voz (ASR) es una de las ramas de la inteligencia artificial que hace posible la comunicación entre el humano y la máquina, logrando que el usuario pueda interactuar con las máquinas de manera natural. En los últimos años, los sistemas ASR se han incrementado hasta el punto de lograr transcripciones casi perfectas, en la actualidad son muchas las empresas que desarrollan sistemas ASR tales como Google, Amazon, IBM, Microsoft. El objetivo de este trabajo es evaluar los sistemas de reconocimiento de voz de Google Speech to Text y Amazon Transcribe con el fin de determinar cuál de ellas ofrece una mayor precisión al momento de convertir el audio en texto. La precisión de las transcripciones se evalúa a través de la tasa de error por palabra (WER) la cual analiza las palabras eliminadas, sustituidas e insertadas con respecto a un texto de referencia de transcripción humana. Después de someter estos sistemas a diferentes ambientes de ruido se observa que el sistema con mayor rendimiento en el proceso de transcripción es el de Amazon Transcribe; por tal razón, se concluye que el servicio de Amazon muestra un mayor desempeño con respecto al servicio de Google tanto con audios con un nivel de ruido de fondo más alto y con audios con un nivel de ruido de fondo más bajo.

Como citar: Peralta, J., Narváez, C., Orellana, M., Patiño, P. & Cedillo, O. (2021). Evaluación del reconocimiento de voz entre los servicios de Google y Amazon aplicado al Sistema Integrado de Seguridad ECU 911. *Revista Tecnológica - Espol, 33(2), 147-158.*
http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/840

Palabras clave: Amazon Transcribe, ASR, Google Speech to Text, transcripción, WER.

Abstract

Automatic Speech Recognition (ASR) is one of the branches of artificial intelligence that makes communication between humans and machines possible, making it the closest thing to the interaction between humans. In recent years, ASR systems have increased to the point of achieving near-perfect transcriptions; today, many companies develop ASR systems, such as Google, Amazon, IBM, and Microsoft. This study aims to evaluate the voice recognition systems of Google Speech to Text and Amazon Transcribe to determine which of them offers greater precision when converting audio into text. The accuracy of transcripts was evaluated through the Word Error Rate (WER), which analyzes the deleted, substituted, and inserted words concerning a human transcription reference text. After subjecting the systems to different noise environments, it was observed that the system with the highest performance in transcripts was Amazon Transcribe; therefore, it was concluded that Amazon services showed a higher performance compared to Google services both with audios with a higher background noise level and with audios with a lower background noise level.

Keywords: Amazon Transcribe, ASR, Google Speech to Text, transcripton, WER.

Introducción

Dentro del dominio de la inteligencia artificial, el reconocimiento automático de voz (ASR) permite la comunicación entre el humano y la máquina, tratando de asemejar al máximo, la forma en que los humanos interactúan entre sí (Këpuska, 2017). Anteriormente, el lenguaje humano no se consideraba una variante importante entre la comunicación humano - máquina, debido a que no existía una tecnología que cumpliera con los parámetros necesarios para el desarrollo y la implementación en casos de la vida real (Yu Dong, 2015). Los parámetros importantes que se deben tener en cuenta en el proceso de reconocimiento de voz son el retraso y la precisión, ya que afectan a la calidad de la experiencia del usuario de dichos servicios (Assefi et al., 2015). Los sistemas de reconocimiento de voz van avanzando y mejorando con el tiempo, de tal forma que estos sistemas realizan tareas de transcripción con una precisión que bordea el 90% (Mashao et al., 2010). En la actualidad, son muchas las empresas que optan por implementar el reconocimiento de voz en sus servicios, ya sea por la comodidad que ofrecen al momento de realizar ciertas tareas o porque cada año la tasa de error va descendiendo, lo que provoca que estos servicios sean mucho más eficaces y confiables.

Empresas como Microsoft, IBM, Google y Amazon desarrollan plataformas que ofrecen servicios en la nube y que utilizan algoritmos ASR (IANCU, 2019). Microsoft utiliza Azure Cloud Service, el cual brinda la posibilidad de transcribir a texto los datos de audio en tiempo real, o por otro lado, permite subir archivos almacenados en el dispositivo del cliente para transcribirlos (Microsoft, 2011). La empresa IBM cuenta con el servicio llamado *IBM Cloud Speech to Text*, el cual ofrece a los clientes la posibilidad de transcribir archivos de audio de una forma continua y con baja latencia; la empresa ofrece servicios estándar y premium (Service et al., 2019), finalmente está Google con el servicio denominado Google Cloud Speech-to-Text.

El presente trabajo tiene como objetivo la comparación entre las dos grandes empresas que se posicionan en el mercado del reconocimiento de voz, es el caso de Google y Amazon. Google con el servicio de *Google Cloud Speech-to-Text* trabaja en la Web y permite cargar archivos de audio o procesar sonidos en tiempo real. El servicio de Google proporciona a los usuarios la posibilidad de escoger el idioma y permite la detección de contenido inadecuado o

vulgar mediante filtros (Morbini et al., 2013). Google Speech-to-Text también ofrece otras características como la detección automática de idiomas o el reconocimiento automático de puntuación (IANCU, 2019). El servicio Amazon Transcribe, de la empresa Amazon, es una herramienta que ofrece a los consumidores la capacidad de agregar funciones de transformación de voz a texto en sus aplicaciones. *Amazon Transcribe* también utiliza el reconocimiento automático de voz (ASR) para generar transcripciones rápidas y de gran calidad (Amazon, n.d.).

Por otra parte, la necesidad de realizar tareas de transcripción automática varía de acuerdo al dominio que es usado. Un caso puntual son los centros de comando y control (C2s), conocidos comúnmente como centros de operaciones, las cuales son entidades de orden público que tienen la función de recopilar los datos provenientes de sensores o cámaras de monitoreo, que se encuentran distribuidos por todo el territorio nacional, toda la información que se recolecta se almacena para ser procesada y analizada a través de un conjunto de plataformas que ayudan en la toma de decisiones (Muse et al., 2020). En complemento a estas tareas, los C2 reciben constantemente llamadas de la comunidad en busca de despachos para emergencias o de soporte policial.

Este estudio realiza una comparativa, que permite evaluar las características de los productos existentes al sistema integrado de seguridad ECU 911, que es la entidad encargada de organizar y dar respuesta ante las situaciones de emergencia que se puedan presentar en el territorio nacional, coordinando la atención con los diferentes organismos en caso de siniestros o desastres. Esta entidad proporciona las grabaciones relacionadas con las alertas emitidas, las cuales se analizan para su registro y codificación considerando las particularidades de cada región. Además, una vez concluidas las configuraciones y entrenamientos necesarios, se determina que producto ofrece una mayor precisión al momento de convertir el audio a texto. También, se evalúa la calidad de las transcripciones de las herramientas utilizadas con relación a las transcripciones realizadas por humanos.

Este estudio se organiza de la siguiente manera: En la Sección 2 se presentan los trabajos relacionados, en la Sección 3 se explica la metodología que se usó para realizar las pruebas, en la Sección 4 se muestran los resultados obtenidos y en la Sección 5 se exponen las conclusiones.

Trabajos relacionados

En el campo de la conversión de voz a texto, existen varios aportes, mismos que abordan la evaluación de ciertos factores clave, que influyen en la calidad del resultado ofrecido. Sin embargo, estos sistemas trabajan con idiomas como el rumano, inglés o japonés. Dichos estudios consideran factores como el nivel de ruido o la tasa de error de las palabras convertidas. Los trabajos que se presentan a continuación se relacionan directamente con los métodos de evaluación y las tareas de transcripción.

En Këpuska (2017), se diseña una herramienta que permite comparar varios sistemas de reconocimiento automático de voz (ASR) como Microsoft Speech Api, Google Speech Api y Sphinx-4, que son ASR de código abierto. Para ello, se utilizan grabaciones de diferentes fuentes, en idioma inglés, para calcular la tasa de error de palabras (WER) y la exactitud de las mismas. Allí, se determina que la API de Google fue superior.

En el trabajo que realiza Kimura et al. (2018) se utiliza la métrica WER para comparar el rendimiento entre los ASR de Kaldi y la API de Google. Para el estudio se emplean audios de habla japonesa en tiempo real, lo que permite concluir que el ASR de Kaldi muestra una

alta precisión de reconocimiento cuando los datos se encuentran en un dominio cerrado y el nivel de ruido en los audios no es alto. En su lugar, el API de Google presenta una gran precisión y estabilidad en varios entornos y dominios, así como un tiempo de respuesta mucho menor. En Filippidou y Moussiades (2020), se compara la precisión de Google respecto a IBM Watson y Wit, utilizando WER junto con los cálculos de la tasa de error de palabras independiente de la posición de hipótesis (Hper) y la tasa de error de palabras independiente de la posición de referencia (Rper). Concluyen que el ASR de Google es más eficaz que los otros sistemas evaluados.

En (IANCU, 2019), se evalúa la API de Google Cloud Speech-to-Text, utilizando vídeos disponibles de YouTube en rumano. Los datos que se obtienen se indexan para transformarse en material para búsquedas. El autor utiliza WER para medir la precisión del contenido multimedia y concluye que los resultados para la indexación de estos recursos son satisfactorios. En el trabajo de Kodish-Wachs et al, (Kodish-Wachs et al., 2018) los autores desarrollan una comparación sistemática de ASR aplicada al lenguaje clínico convencional. En el estudio se usan audios grabados de escenarios clínicos y se evalúan ocho motores de ASR usando la tasa de error de palabra (WER). Como resultado se encuentra que los motores de ASR generan una amplia gama de errores de palabras.

Por otro lado, en Assefi, Liu, et al. (2015), comparan dos sistemas basados en la nube, Apple Siri y Google Speech Recognition (GSR), evaluando los sistemas de acuerdo a los parámetros de retraso y precisión de las transcripciones de cada herramienta. Se llega a la conclusión de que los sistemas que están basados en la nube son afectados por la pérdida de paquetes o por la fluctuación que se da comúnmente en sistemas conectados a la red inalámbrica y celular. En Assefi, Wittie, et al., (2015), se estudian los mismos sistemas de reconocimiento de voz basados en la nube, y se evalúa el rendimiento ante varias situaciones de red, en parámetros de precisión de reconocimiento de comandos y retardo de ida y vuelta. Los resultados obtenidos son similares a los detallados anteriormente (Assefi, Liu, et al., 2015); sin embargo, se propone una solución de transporte de codificación de red para aumentar la calidad de las transmisiones de voz.

En Wang (2019), se realiza un estudio sobre el reconocimiento de emociones basado en un modelo de red neuronal convolucional y recurrente (ARCNN), que es aplicado a diálogos telefónicos de atención al cliente. El modelo se entrena con el texto convertido de la voz del teléfono del cliente mediante la API de transcripción de Amazon. Así también, en Munot y Nenkova, (2019), se analiza cómo las emociones en los diálogos influyen en el desempeño de los sistemas. Se concluye que el rendimiento de los sistemas de reconocimiento depende de factores como el contenido léxico, la identidad del hablante y el dialecto. Se evalúan varias aplicaciones comerciales de compañías como Amazon, IBM y Google.

En Mashao et al., (2010), se realiza una investigación para la implementación del reconocimiento de voz aplicado a dispositivos móviles, bien sea en red o con reconocimiento de voz distribuido. Se llega a la conclusión de que el reconocimiento distribuido obtiene una mayor precisión.

Como se analiza, existen múltiples estudios de evaluación y comparación de sistemas de reconocimiento automático de voz (ASR); sin embargo, no se precisan estudios de comparación de dos grandes empresas como Google y Amazon. Tampoco se encuentran trabajos relacionados al reconocimiento de voz aplicado a los centros de comando y control, ni soluciones dirigidas al idioma español. Por ello, el presente estudio se centra en evaluar los sistemas de reconocimiento automático de voz de las empresas Google y Amazon, mediante

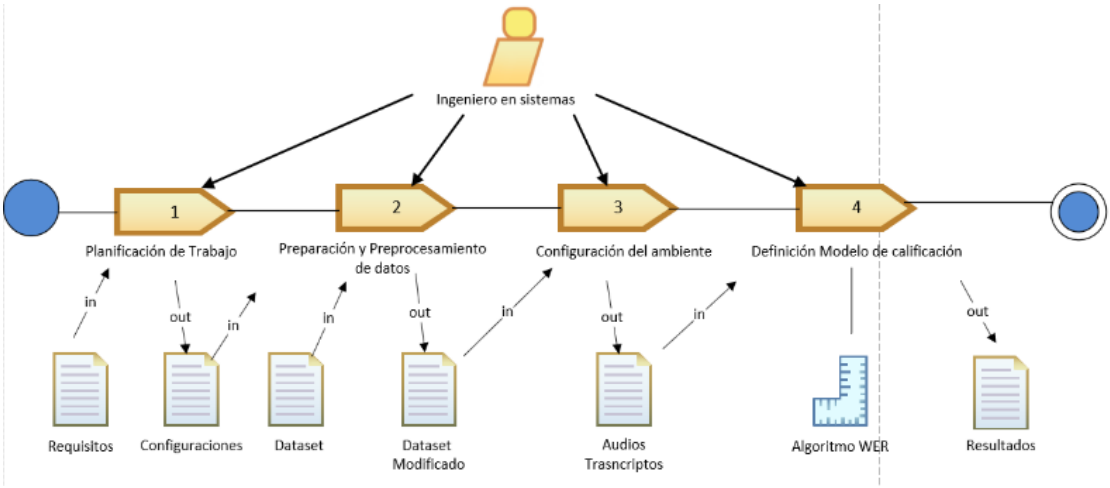
grabaciones de audio en situaciones de emergencia, determinando su calidad con respecto a transcripciones realizadas por humanos.

Metodología

La metodología empleada durante el proceso de comparación estuvo conformada por cuatro actividades detalladas en la Figura 1. Estas actividades utilizaron como entrada los datos provenientes de la recopilación de llamadas de emergencia realizadas al Servicio Integrado de Seguridad ECU 911, los registros de llamadas incluían tanto el audio como su transcripción manual.

Figura 1

Actividades relacionadas a la metodología de evaluación de los sistemas ASR



Planificación de Trabajo

En esta actividad se detallan los pasos para configurar el entorno de trabajo, como se puede observar en la Figura 2.

Figura 2

Esquema de Planificación de Trabajo

Se creó una cuenta en Google Cloud para tener acceso al servicio de Google Speech to Text. Debido a que la mayoría de audios tenían una duración superior a un minuto, se generó un storage en donde se almacenaron estos datos. Por otra parte, se creó una cuenta en los servicios Web de Amazon para poder tener acceso a los servicios de Amazon Transcribe, se precisó un espacio de almacenamiento para los audios (bucket) y se generó un espacio de trabajo para realizar las transcripciones (transcription job).

Finalmente, se creó un entorno de trabajo en Google Collaboratory, se trata de un documento compartido al cual se puede acceder desde cualquier dispositivo y que permite

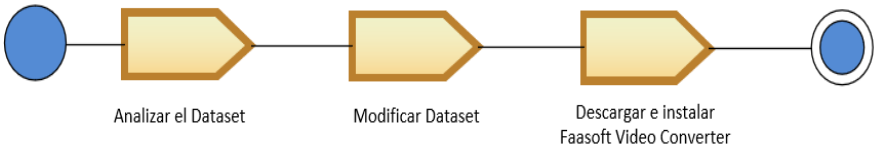
trabajar con el lenguaje de programación Python sin la necesidad de descargar ningún programa. En este documento se implementó tanto la API de Google Speech to Text como la API de Amazon Transcribe.

Preparación y Preprocesamiento de Datos

En la Figura 3, se detallan los pasos para la preparación y el preprocesamiento de los datos.

Figura 3

Esquema de Preparación y Preprocesamiento de Datos



Se analizó el Dataset proporcionado por el ECU 911, el mismo que contaba con diversos campos tales como: nombre del archivo de los audios, la dirección y las transcripciones, tanto del operador como del alertante.

Las distintas herramientas analizadas requerían que el archivo de audio fuera de 16 bits y que el canal del audio fuera estéreo (2 canales). Como parte de ajuste del requerimiento se pudo utilizar el software Fassoft Video Converter para modificar la codificación de las llamadas según las características deseadas.

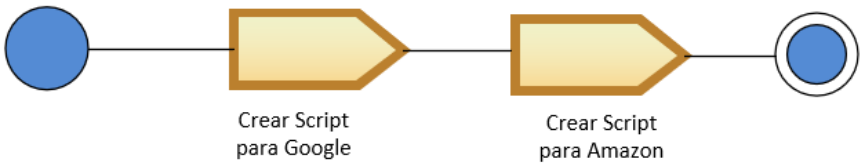
Para continuar, se dividieron los archivos de audio en dos categorías. En la primera categoría se agruparon los audios que no contenían un alto nivel de ruido de fondo y en la segunda categoría se agruparon los audios que tenían un mayor nivel de ruido de fondo.

Configuración del Ambiente

En la actividad que se presenta en la Figura 4, se detalla la creación de los Scripts para la transcripción tanto para Google como para Amazon.

Figura 4

Esquema de Configuración del Ambiente



Se creó el Script para la codificación e implementación de la API de Google, el cual permitió realizar las transcripciones de los audios. Para que la comparación esté dentro de los mismos parámetros fue necesario configurar la transcripción resultante, para lo cual se eliminaron todos los signos de puntuación, así como se utilizó un método *lowercase* el cual retornó todas las palabras en minúsculas.

Por otra parte, Amazon permitió realizar las transcripciones desde sus servicios Web (AWS), esta proporcionó un archivo de formato JSON con la transcripción del audio. Para el

análisis del mismo se creó un Script que permitió configurar la transcripción resultante para igualar a los parámetros anteriormente señalados, donde se utilizó un método *replace* para reemplazar los signos de puntuación por espacios en blanco y el método *lowercase* que permitió dejar el texto en minúsculas.

Definición del modelo de calificación

Para medir la efectividad de las transcripciones a texto de los dos servicios antes mencionados se utilizó el cálculo de la tasa de error por palabras (WER). El cual, en primera instancia, tomó un texto de referencia que contenía la transcripción del audio sin ningún error. Posteriormente, se consideró un texto de análisis el cual fue generado a partir de las transcripciones realizadas por los servicios de Google y Amazon. Se compararon los dos textos tomando en cuenta las palabras que fueron eliminadas, cambiadas y añadidas. Este cálculo dio como resultado el porcentaje de error de las transcripciones respecto al texto de referencia. El cálculo del WER se dio bajo la siguiente fórmula:

$$WER = \frac{I + D + S}{N}$$

Donde (I) son las palabras insertadas, (D) son las palabras eliminadas, (S) son las palabras que se sustituyeron y (N) es el total de palabras del texto de referencia.

Resultados

La Tabla 1muestra los resultados que se obtienen para la tasa de error por palabra (WER) de 10 audios proporcionados por el sistema ECU911, los cuales no contienen ruido de fondo. El índice de WER que logra la API de Amazon Transcribe es menor que el índice que alcanza la API de Google Speech to Text, es decir la API de Amazon es más efectiva al momento de realizar las transcripciones en comparación a la API de Google.

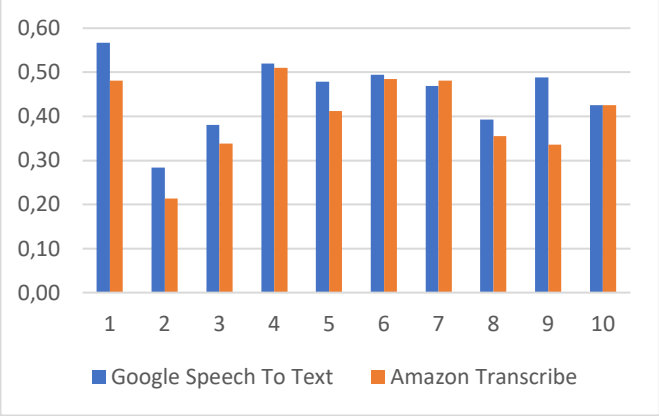
Tabla 1

Resultados de las Transcripciones (Sin Ruido de Fondo)

Nombre del archivo	Google Speech to Text	Amazon Transcribe
	WER (%)	
f58a547c-b84b-4d46-b2cfd2d5f21f8d2b	0,57	0,48
f2a3ae7f-aa13-4742-8030-82a274be4405	0,28	0,21
dddd5d0c-4453-4389-b960-fed73caa2bf7	0,38	0,34
d0fa229f-a710-465d-ac3f-d701a05ab071	0,52	0,51
cd82b677-0732-424e-a432-cc039c1eb520	0,48	0,41
9544243b-fc6a-4c1c-b622-c6812828d519	0,49	0,48
91ddd3ad-1ebf-43ca-b628-9f425c83d65b	0,47	0,48
56d63c4d-d64c-4022-bd4e-688c1f3a45b4	0,39	0,36
49bbd185-9b67-4355-b504-600a064f1049	0,49	0,34
45a51110-078a-4e5b-9366-196f4f12c3ce	0,43	0,43
Media	**0,45**	**0,40**

Figura 5

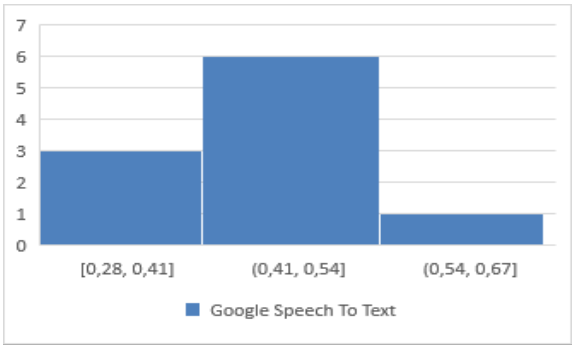
Comparación Entre los Dos Sistemas (Sin Ruido de Fondo)



La Figura 6, presenta un histograma de los resultados que se obtienen de la API de Google Speech to Text de los audios que no presentan ruido de fondo. Se puede observar que existe una frecuencia mayor para los resultados que oscilan entre el 41 y 54 por ciento de WER.

Figura 6

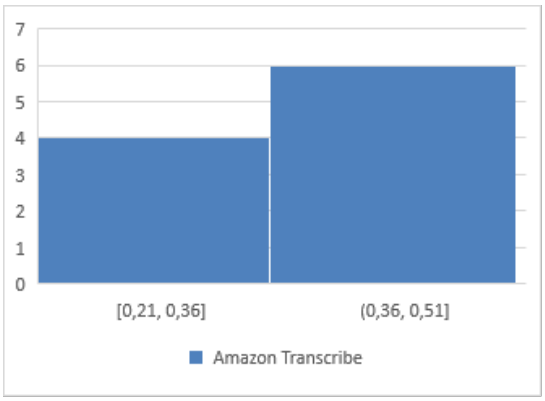
Histograma de las Transcripciones de Google (Sin Ruido de Fondo)



En la Figura 7, se observa el histograma con los resultados que se obtienen de la API de Amazon Transcribe de los audios que no presentan ruido de fondo, donde se puede observar que la mayor frecuencia se encuentra entre el 36 y 51 por ciento de WER.

Figura 7

Histograma de las Transcripciones de Amazon (Sin Ruido de Fondo)



La Tabla 2 muestra los resultados que se obtiene para la tasa de error por palabra (WER) de cada uno de los audios del estudio. Se consideran elementos con un ruido de fondo más intenso para observar el comportamiento de las API en diferentes escenarios. El índice WER que se logra por la API de Amazon Transcribe es menor que el índice WER que se obtiene por la API de Google Speech to Text, es decir, la API de Amazon una vez más tiene mayor efectividad al momento de realizar las transcripciones en comparación a la API de Google.

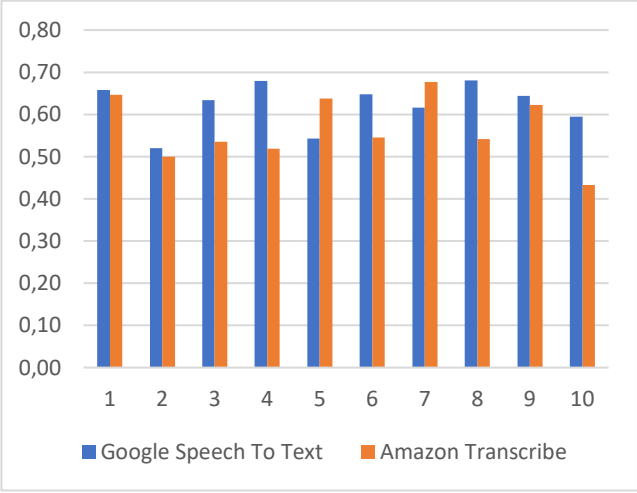
Tabla 2

Resultados de las Transcripciones (Con Ruido de Fondo)

Nombre del archivo	Google Speech to Text	Amazon Transcribe
	WER (%)	
c4095c6e-e4d1-44c2-a03d-6b007269da15	0,66	0,65
cdf112ec-4be5-4aa0-9a8e-b5da95320660	0,52	0,50
ce0685b3-1c76-4e19-ae50-1e59c64d4f98	0,63	0,54
d3d7d6b5-8b04-4778-96d6-98f638fa05c1	0,68	0,52
e185a546-a73f-4dbf-90b7-64c32edfcb74	0,54	0,64
edb78b82-e5af-4892-b429-be2fe23af8be	0,65	0,55
f42dbd3a-890b-4df0-836d-21a6d0216255	0,62	0,68
fa66d704-aacb-4333-a847-03badd6f08f7	0,68	0,54
fa9148cf-5020-4992-b1ab-c59806508da8	0,64	0,62
fcb4f4eb-4c21-45d9-ba71-321b0a1b3aab	0,59	0,43
Media	**0,62**	**0,58**

Figura 8

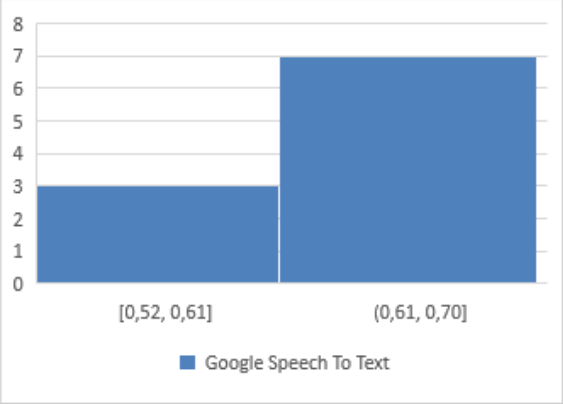
Comparación Entre los Dos Sistemas (Con Ruido de Fondo)



En la Figura 9. Se muestra el histograma de los resultados de la API de Google Speech to Text con audios que contienen ruidos de fondo, donde se puede observar que existe una mayor frecuencia en los resultados del WER que varían entre el 61 y 70 por ciento.

Figura 9

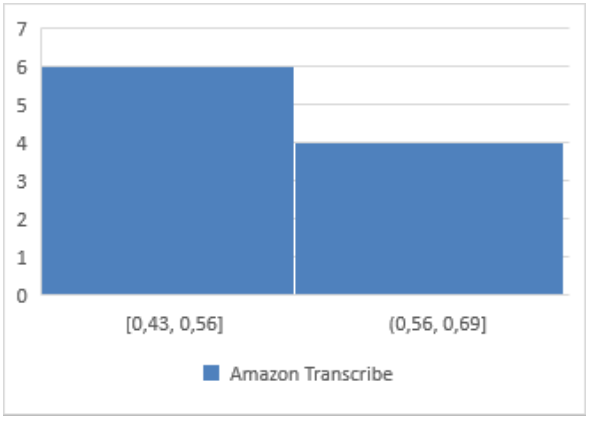
Histograma de las Transcripciones de Google (Con Ruido de Fondo)



En la Figura 10, se muestra el histograma de los resultados de la API de Amazon Transcribe con audios que contienen ruidos de fondo, donde se puede observar que existe una mayor frecuencia en los resultados del WER que varían entre el 43 y 56 por ciento.

Figura 10

Histograma de las Transcripciones de Amazon (Con Ruido de Fondo)



Conclusiones

Como parte de las conclusiones se puede identificar consideraciones a tener en cuenta para el procesamiento de los audios, tales como la separación de los canales, problemas de interferencia entre los dos interlocutores y en la codificación de los bits de los audios. Estos aspectos deben ser mejorados para elevar la calidad de los sistemas ASR en tareas de transcripción.

Con base en los resultados del estudio, los servicios de Amazon alcanzan una media de error del 40%, mientras que los servicios de Google alcanzan un 45% de media de error en aquellos audios que no contienen ruido de fondo. En audios con mayor ruido de fondo, los servicios de Amazon vuelven a tener un valor menor de tasa de error con un 58%, en comparación al 62% de los servicios de Google.

Se puede concluir que los servicios de Amazon muestran una mayor eficiencia respecto a los servicios de Google tanto con audios con un nivel de ruido de fondo más alto como en audios con un nivel de ruido de fondo más bajo. El método de evaluación como los resultados

que logra este estudio pueden ser considerados por otros trabajos que se enfoquen a realizar tareas de transcripción automática. Como trabajo futuro se propone buscar técnicas que mejoren la calidad de los audios, que permitan eliminar el ruido y elevar la claridad de la voz para una óptima transcripción.

Se planifica realizar, como próximos pasos, la comparación con otras herramientas tales como Microsoft e IBM. Además, se espera realizar una medición de otros aspectos como los modelos internos de cada uno de los servicios que se estudian, el tipo de redes neuronales que utilizan, análisis de corpus, arquitectura; además, características de calidad tales como: precisión, confiabilidad, eficiencia, entre otros.

Agradecimientos

Los autores desean agradecer al Vicerrectorado de Investigaciones de la Universidad del Azuay por el apoyo financiero y académico, así como a todo el personal de la escuela de Ingeniería de Sistemas y Telemática, y el Laboratorio de Investigación y Desarrollo en Informática (LIDI).

Referencias

Amazon. (n.d.). *Amazon Transcribe Guía para desarrolladores*. 1–334. https://docs.aws.amazon.com/es_es/transcribe/latest/dg/transcribe-dg.pdf

Assefi, M., Liu, G., Wittie, M. P., & Izurieta, C. (2015). An experimental evaluation of Apple Siri and Google Speech Recognition. *24th International Conference on Software Engineering and Data Engineering, SEDE 2015*.

Assefi, M., Wittie, M., & Knight, A. (2015). Impact of network performance on cloud speech recognition. *Proceedings - International Conference on Computer Communications and Networks, ICCCN*, *2015-Octob*. https://doi.org/10.1109/ICCCN.2015.7288417

Filippidou, F., & Moussiades, L. (2020). *A Benchmarking of IBM, Google and Wit*. https://doi.org/10.1007/978-3-030-49161-1

IANCU, B. (2019). Evaluating Google Speech-to-Text API's Performance for Romanian e-Learning Resources. *Informatica Economica*, *23*(1/2019), 17–25. https://doi.org/10.12948/issn14531305/23.1.2019.02

Këpuska, V. (2017). Comparing Speech Recognition Systems (Microsoft API, Google API And CMU Sphinx). *International Journal of Engineering Research and Applications*. https://doi.org/10.9790/9622-0703022024

Kodish-Wachs, J., Agassi, E., Kenny III, P., & Overhage., J. M. (2018). *A systematic comparison of contemprary automatic speech recognition.pdf*.

Mashao, Daniel J, Isaacs, D. (2010). *A Comparison of the Network Speech Recognition and Distributed Speech Recognition Systems and their eect on Speech Enabling Mobile Devices*. *February*, 1–94. https://open.uct.ac.za/handle/11427/11232

Microsoft. (2011). *What is the Speech service*.

Morbini, F., Audhkhasi, K., Sagae, K., Artstein, R., Can, D., Georgiou, P., Narayanan, S., Leuski, A., & Traum, D. (2013). Which ASR should i choose for my dialogue system? *SIGDIAL 2013 - 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Proceedings of the Conference*, *August*, 394–403.

Munot, R., & Nenkova, A. (2019). Emotion impacts speech recognition performance. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Student Research Workshop*, 16–21.

https://doi.org/10.18653/v1/n19-3003

Muse, L. P., Martins, P. R., Hojda, A., Abreu, P. A. De, & Almeida, P. C. De. (2020). The role of Urban Control and Command Centers in the face of COVID-19: The case of COR in Rio de Janeiro, Brazil. *2020 IEEE International Smart Cities Conference, ISC2 2020*. https://doi.org/10.1109/ISC251055.2020.9239068

Service, C., Sheets, P. D., Levels, S., & Support, T. (2019). *IBM Cloud Additional Service Description IBM Watson Speech to Text Service Levels and Technical Support*. *09*, 3–5. https://www-03.ibm.com/software/sla/sladb.nsf/8bd55c6b9fa8039c86256c6800578854/78a62403a2752f7f862583b3006435bd/$FILE/i126-6945-09_03-2019_en_US.pdf

Takashi Kimura, Takashi Nose, Shinji Hirooka, Yuya Chiba, A. I. (2018). *Comparison of Speech Recognition Performance Between Kaldi and Google Cloud Speech API* (Vol. 2).

Wang, Y. (2019). *Sentiment Analysis of Customer Support Phone Dialogues using Fusion-based Emotion Recognition Techniques*. https://repository.library.northeastern.edu/files/neu:m044wq01p/fulltext.pdf

Yu Dong, L. D. (2015). Automatic Speech Recognition. In *Lecture Notes in Electrical Engineering* (Vol. 686). https://doi.org/10.1007/978-981-15-7031-5_63