Enfoque UTE
ISSN: 1390-6542
enfoque@ute.edu.ec
Universidad Tecnológica Equinoccial
Ecuador

Plaza, Johanna; Sánchez-Zhunio, Cristina; Acosta-Urigüen, María-Inés; Orellana, Marcos; Cedillo, Priscila; Cedillo, Priscila; Zambrano-Martinez, Jorge Luis

Reconocimiento del habla con acento español basado en un modelo acústico
Enfoque UTE, vol. 13, núm. 3, 2022, Julio-Septiembre, pp. 45-57
Universidad Tecnológica Equinoccial
Ecuador

DOI: https://doi.org/10.29019/enfoqueute.839

Disponible en: https://www.redalyc.org/articulo.oa?id=572270556005

Enfoque UTE, V.13-N.3, Jul. 2022, pp. 45-57
http://ingenieria.ute.edu.ec/enfoqueute/
e-ISSN: 1390-6542

Received: 2022/02/22
Accepted: 2022/05/13
Online: 2022/05/13
Published: 2022/07/01
https://doi.org/10.29019/enfoqueute.839

Reconocimiento del habla con acento español basado en un modelo acústico

(Speech recognition based on Spanish accent acoustic model)

Johanna Plaza,[1] Cristina Sánchez-Zhunio,[2] María-Inés Acosta-Urigüen,[3] Marcos Orellana,[4] Priscila Cedillo,[5 6] Jorge Luis Zambrano-Martinez[7]

Resumen

El objetivo de este artículo fue generar un modelo reconocimiento automático de voz (RAV) basado en la traducción de la voz humana a texto, a este proceso se lo ha considerado como una de las ramas de la inteligencia artificial. El análisis de voz permite identificar información sobre la acústica, fonética, sintáctica, semántica de las palabras, entre otros elementos que pueden identificar ambigüedad en términos, errores de pronunciación, sintáctica similar pero semántica diferente, que representan características propias del lenguaje humano. El modelo se centró en el análisis acústico de las palabras, proponiendo la generación de una metodología para reconocimiento acústico, a partir de transcripciones del habla de audios que contienen voz humana. Se utilizó la tasa de error por palabra para identificar la precisión del modelo. Los audios son llamadas de emergencia registrados por el Servicio Integrado de Seguridad ECU 911. El modelo fue entrenado con la herramienta CMUSphinx para idioma español sin conexión a internet. Los resultados mostraron que la tasa de error por palabra varía en relación con la cantidad de audios; es decir a mayor cantidad de audios menor cantidad de palabras erróneas y mayor exactitud del modelo. La investigación concluyó haciendo énfasis en la duración de cada audio como variable que afecta la precisión del modelo.

Palabras clave

RAV; Modelo de Lenguaje; CMUSphinx.

Abstract

The objective of the article was to generate an Automatic Speech Recognition (ASR) model based on the translation from human voice to text, being considered as one of the branches of artificial intelligence. Voice analysis allows identifying information about the acoustics, phonetics, syntax, semantics of words, among other elements where ambiguity in terms, pronunciation errors, similar syntax but different semantics can be identified, which represent characteristics of the language. The model focused on the acoustic analysis of words proposing the generation of a methodology for acoustic recognition from speech transcripts from audios containing human voice and the error rate per word was considered to identify the accuracy of the model. The audios were taken from the Integrated Security Service ECU 911 that represent emergency calls registered by the entity. The model was trained with the CMUSphinx tool for the Spanish language without internet connection. The results showed that the word error rate varies in relation to the number of audios; that is, the greater the number of audios, the smaller number of erroneous words and the greater the accuracy of the model. The investigation concluded by emphasizing the duration of each audio as a variable that affects the accuracy of the model.

Keywords

Automatic Speech Recognition; Language Model; CMUSphinx.

1 Universidad del Azuay, Cuenca - Ecuador [jgabyplaza97@es.uazuay.edu.ec, https://orcid.org/0000-0003-1998-441X].
2 Universidad del Azuay, Cuenca - Ecuador [cristina.sanchezz@ucuenca.ec, https://orcid.org/0000-0002-9952-4853].
3 Universidad del Azuay, Cuenca - Ecuador [macosta@uazuay.edu.ec, https://orcid.org/0000-0003-4865-2983].
4 Universidad del Azuay, Cuenca - Ecuador [marore@uazuay.edu.ec, https://orcid.org/0000-0002-3671-9362].
5 Universidad del Azuay, Cuenca - Ecuador [icedillo@uazuay.edu.ec, https://orcid.org/0000-0002-6787-0655].
6 Universidad de Cuenca, Cuenca - Ecuador [priscila.cedillo@ucuenca.edu.ec, https://orcid.org/0000-0002-6787-0655].
7 Universidad del Azuay, Cuenca – Ecuador [jorge.zambrano@uazuay.edu.ec, https://orcid.org/0000-0002-5339-7860].

1. Introducción

El Procesamiento del Lenguaje Natural (PLN) es un enfoque computacional para el análisis de texto que permite estudiar los textos que ocurren naturalmente en diferentes niveles de análisis lingüístico, con el propósito de lograr un procesamiento del lenguaje similar al humano para una variedad de aplicaciones (Belinkov y Glass, 2019). El análisis lingüístico abarca el análisis fonético, morfológico, léxico, sintáctico, semántico, discursivo y pragmático del lenguaje (Zhao et al., 2020). Entre algunas aplicaciones que se pueden encontrar del PLN están la generación de resúmenes automáticos de texto, la traducción automática, la búsqueda de información y las respuestas a preguntas (Belinkov et al., 2019).

El reconocimiento automático de voz es una de las aplicaciones en el área del PLN (Ankit et al., 2016), que tiene como objetivo fundamental la transcripción del habla, que se basa en secuencias de palabras representadas a través de ondas de los audios. (Alharbi et al., 2021) Una conversación comúnmente puede darse entre actores humanos y agentes artificiales, donde la naturaleza del discurso, el tamaño del vocabulario y el ancho de banda son aspectos relevantes y primordiales al momento de entrenar un sistema de Reconocimiento Automático de Voz (RAV) (Alharbi et al., 2021). Además, el RAV considera aspectos del lenguaje natural como semántica, sintaxis, gramática y la fonética, dada la variedad de sonidos del habla que pueden producir los seres humanos, que incluyen el ritmo, el acento, la pronunciación dialéctica, las entonaciones peculiares de una palabra para dar un significado u otro, e incluso las distintas malas pronunciaciones en ciertos fonemas como por ejemplo el rotacismo (Aguiar de Lima y Da Costa-Abreu, 2020).

Se han desarrollado algoritmos de aprendizaje por transferencia para el entrenamiento de modelos de reconocimiento de voz (IBM-Custom-Speech y Microsoft-Custom-Speech) donde se convierte un conjunto de diálogos de prueba en documentos de texto y se calcula la tasa de error de caracteres, siendo considerada una métrica general para evaluar el modelo (Kim et al., 2021). Estos sistemas de reconocimiento de voz implementan algoritmos que incluyen la caracterización del efecto de la expresión emocional en el habla, a más de técnicas que están vinculadas para mejorar el rendimiento de los sistemas de procesamiento del habla. Los estados emocionales tales como la ira, la felicidad, el miedo, la tristeza, la sorpresa y el disgusto, influyen en la modulación de la voz en aspectos como el volumen, la entonación, el ritmo y la pronunciación del que habla, que alteran la onda, la intensidad, la calidad del habla, la prosodia y la sincronización (Kim et al., 2021).

Por otro lado, el avance de la tecnología ha permitido reconocer el nivel de eficacia que estos sistemas de reconocimiento de voz obtienen en casos específicos. Así, los sistemas RAV basados en redes neuronales convolucionales (convolutional neural network-ConvNet/CNN) que se generan especialmente durante las llamadas de emergencia y cuyo fin es el de detectar el estado emocional y verificar la autenticidad intencional del hablante (Tavi et al., 2019).

Un claro ejemplo de implementación de RAV, es a través del proyecto Julius, sistema que decodifica el sonido emitido por el usuario y busca, en el modelo acústico alguna coincidencia, a través un modelo lingüístico que estima de manera estadística la probabilidad de que cierto fonema sea pronunciado en un segmento de audio. Al evaluar una palabra, el decodificador comienza a determinar que fonema es el correspondiente al sonido mediante su respectiva evaluación hasta que llega una pausa de parte del usuario, donde la palabra termina. El modelo lingüístico busca esta secuencia de fonemas para llegar a identificar una serie equivalente o posible palabra (Medina et al., 2014). En estos sistemas es imprescindible cuantificar la eficien-

cia, entendida como el porcentaje de éxito del reconocimiento de palabras u oraciones. Por tal motivo es necesario contar con un diccionario fonético, un modelo acústico para la distribución de probabilidad de los fonemas en la señal de audio y un modelo del lenguaje para la distribución de probabilidad de una secuencia de palabras (Celis et al., 2017).

Actualmente existen varios modelos acústicos y de lenguajes que son desarrollados para diferentes idiomas, incluyendo el español y sus variaciones por región. Entre los factores que afectan la eficiencia del sistema RAV están aspectos relacionados al acento que tiene un idioma, la gramática que se emplea, la velocidad del habla, el modelo de lenguaje que se utiliza, y las características propias que posee el funcionamiento del reconocedor (Ankit et al., 2016). Existen diversas herramientas y lenguajes para desarrollar estos modelos, como: Praat, Julius, openSMIL, Sphinx, Kaldi C, C++ Python y Java, que son ampliamente utilizados, y mucho de ellos siendo de código abierto (Aguiar de Lima & Da Costa-Abreu, 2020).

La tasa de error por palabras (WER) es la métrica más conocida que se utiliza para cuantificar la precisión del sistema RAV. Esta métrica considera la secuencia de palabras hipotetizada por el sistema RAV y la alinea con una transcripción de referencia y el número de errores se calcula como la suma de sustituciones (*S*), inserciones (*I*) y eliminaciones (*D*) (Ali y Renals, 2018). También existe otra métrica que proporciona la información cuantificada del porcentaje de palabras que han sido reconocidas por un sistema de reconocimiento de voz denominada "precisión de palabras" (WAcc). Sin duda, esta métrica no es más que la proporción del total de palabras reconocidas en el sistema contra el número total de las palabras que contiene el texto de referencia (N), además guarda una estrecha correlación con la métrica WER (Errattahi et. al, 2018).

En este contexto, el objetivo en que se fundamenta este trabajo es generar un sistema que permita reconocer y convertir archivos de audio automáticamente con la voz humana integrada en texto; la transcripción se realizará en el idioma español, y los audios a utilizar serán tomados de las llamadas de emergencia provenientes del Servicio Integrado de Seguridad ECU 911 de la ciudad de Cuenca, Ecuador. Para el procesamiento del audio se considerarán las palabras en español que comúnmente son usadas por los habitantes de la ciudad, así como sus abreviaciones, jerga o argot, lenguaje coloquial, variedades lingüísticas y formalidades que poseen las personas de una región en su dialecto junto con el acento y la velocidad de la pronunciación.

Este artículo se estructura de la siguiente manera: la Sección 2 contiene la discusión de trabajos relacionados que influyeron en esta investigación realizada. En la Sección 3 se describe la metodología, la cual posee el marco teórico que se ha utilizado en esta investigación. Luego, en la Sección 4 se presenta el análisis de los resultados obtenidos en la investigación y, por último, la Sección 5 contiene las conclusiones obtenidas.

2. Trabajos Relacionados

Alharbi et. al. llevaron a cabo una revisión sistemática de literatura de los sistemas RAV, en el que identificaron las principales técnicas implementadas en este campo. Su principal contribución está en la clasificación de la literatura según los problemas que el dominio presentó, el uso de técnicas de PLN basados en la eficiencia del dispositivo. El artículo presenta que los sistemas RAV tienen problemas fundamentalmente asociados al ruido y reverberación del audio, la superposición de voz o conversaciones simultáneas, el procesamiento de la señal acústica y la adaptación para resolver problemas de ajuste. El vocabulario, la pronunciación y el dialecto son las principales técnicas en las que el PLN se enfoca, por lo que hace énfasis en la cantidad

de palabras que deberían incluirse en el vocabulario, los problemas que pueden ocasionar una mala pronunciación de las palabras y el problema del reconocimiento del dialecto de distintas regiones donde manejan el idioma implementado. Por último, el uso del micrófono también es analizado, debido a que es un dispositivo que captura la voz y que incide radicalmente cuando se hacen entrenamientos y pruebas de datos con y sin su uso (Alharbi et al., 2021).

Tavi et al. en su artículo relacionado a los sistemas RAV para llamadas de emergencia incluye el análisis de voz aguda y desagradable también llamada voz chirriante, alevines vocales o estación laríngea. Por lo tanto, estos son un factor que altera la fonación del hablante. El modelo está implementado con redes neuronales convolucionales (CNN), busca la detección automática de las anomalías de la voz como crujidos para el análisis a gran escala de las llamadas de emergencia, y conjuntamente con un kit de herramientas de análisis de voz, lograron un f-measure de 0.41 con grabaciones de llamadas de emergencia. La red neuronal convolucional alcanzó 0.64 en el mismo conjunto de datos de prueba. Los resultados demuestran la vital importancia de la cantidad de datos de entrenamiento como la calidad de los datos de prueba, así como los elementos que afectan la eficiencia del modelo (Tavi et al., 2019).

Es conocido que existen sistemas RAV desarrollados por empresas pioneras como Microsoft, IBM, Google, Amazon cuya finalidad es lograr transcripciones casi perfectas. Los autores Vásconez et al. presentan un estudio donde evalúan a los sistemas RAV de Google Speech to Text y Amazon Transcribe, con el objetivo de determinar cuál de esos sistemas RAV ofrece una mayor precisión al transcribir en el idioma español los audios del Servicio Integrado de Seguridad ECU 911, en texto. La evaluación de la precisión de la transcripción en este estudio es realizada a través de la métrica WER, donde demostraron con los resultados obtenidos que Amazon Transcribe es el sistema RAV con mayor rendimiento y desempeño en el proceso de transcripción con audios en español que poseen un alto y bajo nivel de ruido de fondo (Peralta Vásconez et al., 2021).

Por otro lado, existen diversas herramientas para implementar los sistemas RAV, Peinl et al. presentan un análisis comparativo de kits de herramientas RAV de código abierto. Así tenemos las herramientas HDecode y Julius que son entrenadas utilizando el kit de herramientas del modelo oculto de Markov (HTK). Estos modelos incluyen el entrenamiento de los monófonos para el manejo del silencio y la realineación para mejorar la correspondencia entre las pronunciaciones y los datos acústicos. Otras herramientas que existen son Pocketsphinx y Sphinx-4, que son entrenados con el kit de herramientas CMU sphinxtrain y que proporciona todas las herramientas necesarias para sistemas RAV. Los modelos están entrenados con vectores Mel Frequency Cepstral, que consta de 13 cepstral, 13 delta y 13 coeficientes de aceleración. Los kits de herramientas de Kaldi proporcionan varias canalizaciones para diferentes corpus. Las capacidades de estas canalizaciones incluyen el entrenamiento adaptativo del hablante y la regresión lineal de máxima verosimilitud. El entrenamiento es de alto costo computacional, la implementación y las canalizaciones están optimizados para computación paralela (Peinl et al., 2020).

Los autores Ankit et. al. presentan el uso de una arquitectura para la conversión de voz denominada "CMU-Sphinx". Esta arquitectura está compuesta por bloques de Front End, lingüística y decodificación. El bloque Front End se encarga del procesamiento de la señal digital que consiste en la extracción de características de la voz; el bloque de lingüística posee los módulos acústicos para la representación estadística de cada sonido, una dicción que almacena las diferentes formas de pronunciación de las palabras; y el modelo del lenguaje que muestra la probabilidad de ocurrencia de una palabra (Ankit et al., 2016; Peinl et al., 2020). Para generar un modelo funcional basado en CMU-Sphinx, se recomienda tener una cantidad sustancial de datos sin procesar que

abarca un rango de 50 a 60 horas de archivos de audio. La información contenida en cada archivo de audio se divide en partes pequeñas que no contengan más de una oración ya en formato de texto, conocido como "Archivo de transcripción" (Lakdawala et al., 2018).

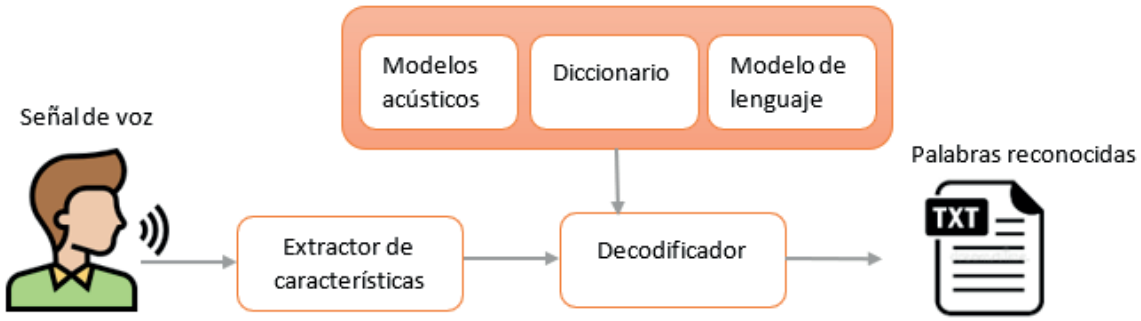
El rendimiento de estos sistemas RAV se mide a través de la métrica WER, que compara la transcripción manual frente a la realizada por el RAV. Para obtener una estimación confiable del WER, se requieren al menos dos horas de archivos de audio de prueba, lo que llega a ser un proceso costoso y lento de transcripción manual. Sin embargo, el uso de sistemas RAV ha ganado terreno en el campo de investigación en los últimos años, ya que estos sistemas permiten mejorar la calidad de transcripción, así como la cantidad de audios procesados (Ali & Renals, 2018).

3. Metodología

Luego de haber realizado un análisis de las herramientas disponibles para sistemas RAV, se escogió a CMUSphinx debido a su notable rendimiento y precisión, además de un conjunto de audios provenientes del Servicio Integrado de Seguridad ECU 911 en idioma español. El proceso de desarrollo está basado en la metodología propuesta en Celis et al. (2017) que consiste en el entrenamiento del modelo acústico adaptándolo al dialecto cucuteño. El aporte de esta metodología fue la adaptación del modelo cucuteño hacia un modelo para el español que incluye el cálculo de la métrica WER, mediante el uso de la herramienta CMU-Sphinx, y el *software* CMU-Cambridge Statistical Language Modeling toolkit (CMUCLMTK). Esta herramienta permite el reconocimiento de voz, además, ofrece una serie de paquetes y herramientas, tanto para plataformas Unix como para Windows, y permite medir los porcentajes de eficiencia del sistema (Lakdawala et al., 2018).

El proceso de desarrollo del sistema RAV para el idioma español se expresa en la figura 1. Hay que considerar que para su desarrollo se empleó un sistema de reconocimiento de locutor dependiente de texto, en otras palabras, que conocemos el contenido léxico de los audios y que está basado en la propuesta de Celis et. al. (2017).

Figura 1. Sistema RAV



Inicialmente, este sistema pasó por un proceso de cuantización vectorial, es decir que la señal acústica se transformó a un vector numérico, dando paso a la extracción de características. Entonces, para la siguiente etapa que es la decodificación fue necesario el vector de características; además en esta etapa intervino el modelo acústico, el diccionario y el modelo de lenguaje. Finalmente, se realizó la transformación a texto (Celis et al., 2017).

El sistema adaptado a un corpus español dependerá de estos tres elementos:

1. *Diccionario fonético*: es un conjunto de palabras en donde se caracteriza por tener su división fonética de cada una de ellas, los fonemas son sonidos reales que se usan para poder pronunciar las palabras (Medina et al., 2014), que, en el caso de nuestro dialecto, se buscarán los fonemas más utilizados localmente dentro del alfabeto fonético internacional (IPA). En la siguiente tabla se presenta un ejemplo:

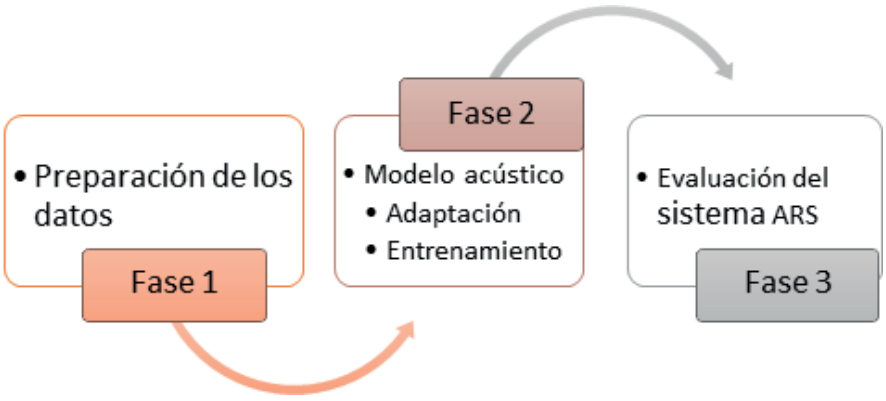
Tabla 1. Ejemplo del alfabeto fonético internacional

Palabra	Fonema
Cerveza	/z/ /e/ /r/ /b/ /a/
Herrero	/e/ /rr/ /r/ /o/
Gara	/gˈ/ /a/ /ɾ/ /a/
Melloco	/m/ /e/ /ʎˈ/ /o/ /k/ /o/

2. *Modelo del lenguaje:* representa el orden estadístico de las palabras, es decir, indica la probabilidad de aparición que pueden tener las palabras en nuestro sistema, la función de nuestro modelo se basará en buscar las palabras adaptadas y reconocidas, en algunos casos cuando no se reconoce la palabra. El modelo se encarga de predecir cuál podría ser la palabra siguiente pero dependiente de la palabra anterior (Celis et al., 2017).

3. *Modelo acústico:* también conocido como corpus de voz, representa la distribución de probabilidades de los fonemas o sonidos que componen las palabras en la señal de audio. Su creación se basa en grandes bases de datos, así como en algoritmos de entrenamiento configurados orientados a diferente idioma (Celis et al., 2017).

Figura 2. Proceso para el desarrollo de sistemas RAV



El proceso que se siguió para desarrollar el sistema RAV en el idioma español se detalla en la figura 2, además que se encuentra dividido en tres fases:

Fase 1: Preparación de los datos

Para el mejoramiento del modelo en el idioma español que proporciona CMUSphinx se necesitó los archivos que se describen a continuación:

- Audios: son los archivos de audios de las llamadas de emergencia del ECU 911, estos fueron transformados a una frecuencia de muestreo de 16 kHz, a 16 bits por muestra, con un códec denominado "modulación por impulsos codificados de 16 bits o PCM16" y sonido monoaural, es decir un solo canal.
- Dataset: es un archivo que contiene las transcripciones realizadas por un humano de los audios del Servicio Integrado de Seguridad ECU 911.
- Corpus en el idioma español: este modelo es proporcionado por CMUSphinx, lo que permite que el proceso de la obtención del modelado no sea desde cero.

Fase 2: Modelo acústico

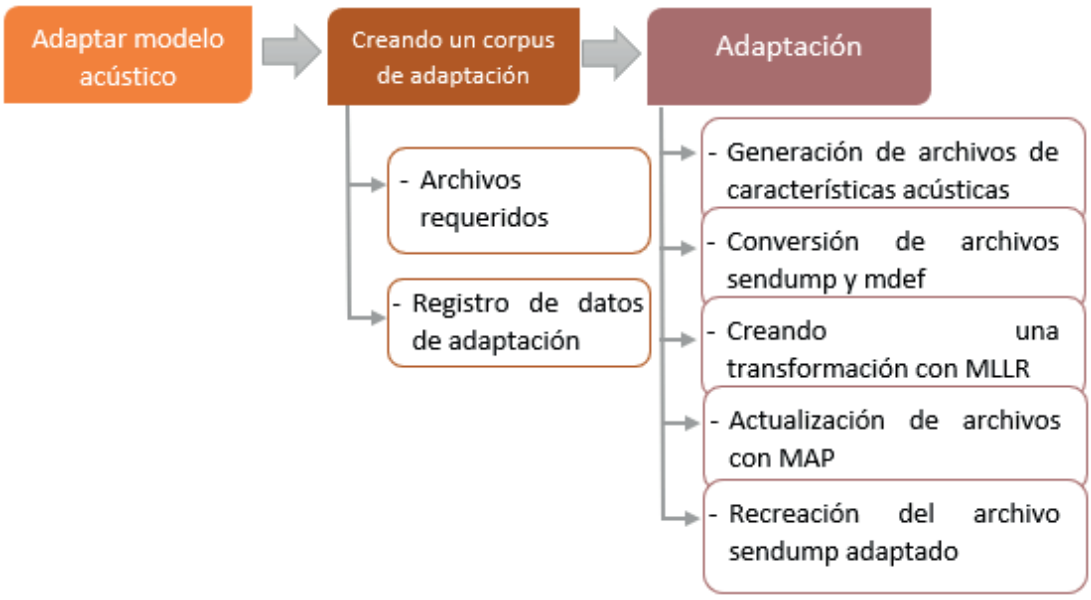
Esta fase se compone de las tareas que son la adaptación y el entrenamiento del modelo acústico. Dichas tareas se describen a continuación:

A. Adaptación del modelo acústico

Este proceso de adaptación del modelo acústico se realizó mediante el *software* SphinxTrain. La adaptación permite mejorar el modelo que se está utilizando, y este proceso tiene la característica de ser más sólido que el entrenamiento. Además, se puede obtener mejores resultados hasta con pequeños datos. Esto se debe a que la adaptación funciona cuando se desea emplear un nuevo acento o idioma, para nuestro caso se entrena un propio modelo acústico orientado a acento español.

Para dar inicio al proceso se utilizó el corpus al que se desea adaptar. Este corpus debe estar compuesto por una lista de oraciones, un diccionario con la pronunciación de cada palabra y los audios donde están grabadas las oraciones de dicha lista. A continuación, en la figura 3 se presenta el proceso de adaptación.

Figura 3. Proceso de adaptación de un modelo acústico predeterminado



En la creación de un corpus de adaptación se debe considerar dos procesos, los cuales son: 1. los archivos requeridos que se basan en un conjunto grande de oraciones; estos deben tener una buena cobertura de las palabras o fonemas que se usan con mayor frecuencia, dependiendo del tipo de texto que se desea reconocer. 2. Los registros de datos de adaptación; estos son archivos de tipo *transcription* y *fileids* en donde se encuentran registrado los audios de cada oración con las características que están nombrados; es decir, con su numeración secuencialmente (CMUSphinx, n.d.).

Para la adaptación del modelo acústico se inició con la generación de archivos de características acústicas a partir de los audios, para esto se utiliza *sphinx_*fe de la herramienta SphinxBase, generando así archivos *.mf* de cada fichero de audio *.wav*. Después, se procedió a la conversión de archivos *sendump*, tratándose de mezclas comprimidas y cuantificadas. Por otro lado, *mdef* es la definición de mapeo entre los contextos trifónicos a los identificadores de modelo de mezcla gaussianas (GMM), y luego convertir el archivo *.mdef* a formato texto. Posteriormente, se recopiló los datos de la adaptación con la ayuda del fichero *bw* de la Sphinxtrain. Los parámetros deben coincidir con el archivo *feat.params* del modelo acústico, en la configuración se determinó que se trata de un modelo continuo. Como siguiente paso se ejecuta de *mllr_solve* y *map_adapt*, dichas herramientas permiten que se actualice el modelo acústico original, la regresión lineal de máxima verosimilitud (MLLR) es una adaptación económica para datos limitados, mejorando el rendimiento al momento de trabajar con modelos continuos (Singh, 2018). Entonces, se usa MAP para la actualización de los archivos del modelo acústico debido a que permite actualizar cada parámetro del modelo. Si se desea ahorrar espacio para el modelo, una buena opción es utilizar el archivo *sendump*. Finalmente, después de todo el proceso se obtuvo un modelo acústico adaptado (Dhanka, 2017).

B. Entrenamiento del modelo acústico

Según el modelo propuesto por Celis et al. (2017), el entrenamiento del modelo acústico consta de cuatro tareas (verificación, extracción y generación, decodificación y resultado), este modelo se aprecia en la figura 4.:

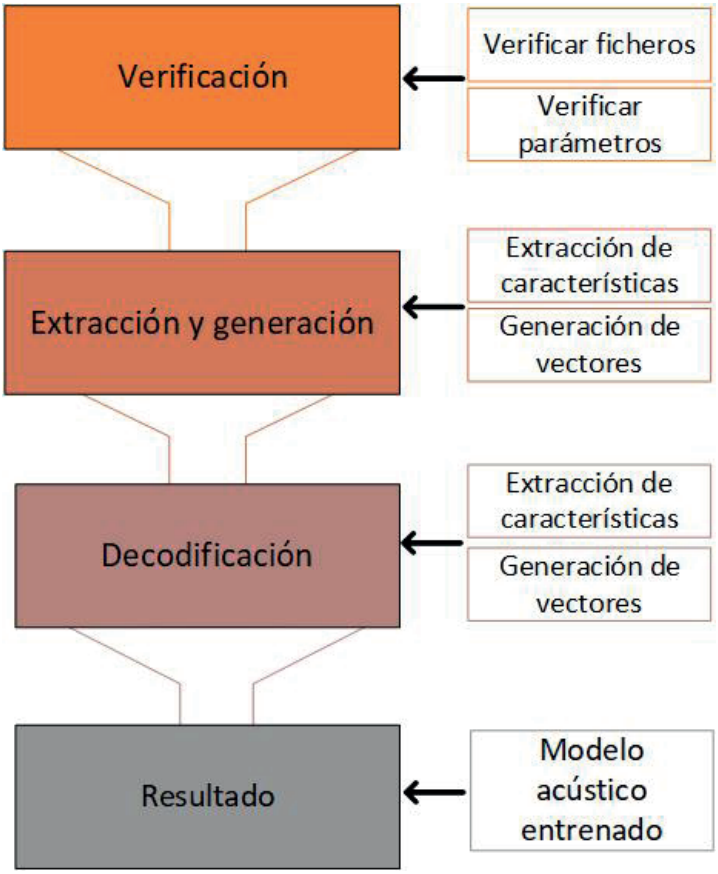
En la primera tarea del entrenamiento del modelo se compone de las actividades de a) verificación de ficheros y b) verificación de parámetros, los mismos que se describen a continuación:

a. Verificación de ficheros: dentro de la carpeta de trabajo deben estar los ficheros con las extensiones. phone, *.filler*, *.fileids*, *.transcription* algunos de estos ya los fueron usados en la etapa de adaptación del modelo.
b. Verificación de parámetros: estos deben ser iguales al archivo feat.params, pueden ser configurados según las características y necesidades.

Luego, en la tarea de extracción y generación se deben extraer las características de los audios y generar vectores que contengan los resultados obtenidos de la tarea anterior, las actividades por realizar son:

a. Extracción de características de audios: la señal es dividida en segmentos para ser procesada.
b. Generación de vectores: este vector se genera como un resultado del análisis que se realiza con la ayuda de la herramienta SphinxTrain.

Figura 4. Proceso de entrenamiento de un modelo acústico



En la tarea de decodificación, se analizaron los resultados de la tarea anterior y se decodificaron los audios, las actividades de esta tarea se describen a continuación:

a. Creación de probabilidad de palabras: con base en el conjunto de vectores se crea un modelo probabilístico, donde se asocian las palabras obtenidas en la señal entrante del audio con las palabras ya existentes en el modelo de lenguaje.
b. Proceso de decodificación de los audios y prueba de entrenamiento: el éxito del paso anterior da como resultado el proceso de decodificación, entonces para verificar el funcionamiento del sistema se realiza a través de una prueba de entrenamiento con los archivos de prueba.

Finalmente, en la tarea de resultados se obtuvo el modelo acústico entrenado, para lo cual fue necesario tener en cuenta que puede variar el tiempo de duración, esto depende de la cantidad de audios que se esté utilizando para el entrenamiento.

Para la evaluación de un sistema RAV se tuvo que tomar en cuenta la métrica WER, anteriormente mencionada. Para este caso se calcula entre la frase generada por el sistema y una frase de referencia correcta. La ecuación 1 permite calcular la métrica WER:

$$WER = \left(\frac{S+D+I}{N}\right) \cdot 100 \quad (1)$$

donde S es el número de palabras sustituidas en la transcripción, D representa el número de palabras borradas u omitidas en el reconocimiento, I es el número de palabras insertadas que no pertenecen a la transcripción real y N representa la cantidad de palabras en la transcripción de referencia.

Otra métrica que fue utilizada para la correcta evaluación de la precisión de un sistema RAV es la exactitud de palabras ($WAcc$), que se define como la proporción de palabras que son reconocidas correctamente frente al total de palabras ingresadas y está estrechamente relacionada esta métrica con WER. La ecuación 2 nos permite obtener el cálculo de la métrica $Wacc$:

$$WAcc = \left(\frac{N\text{-}S\text{-}D\text{-}I}{N}\right) \cdot 100 \tag{2}$$

donde N es el número total de palabras ingresadas, S representa al número total de palabras que fueron reconocidas como otras palabras, D es el número de palabras que se omitieron en el reconocimiento, y la variable I representa al número de palabras que fueron erróneamente agregadas a las palabras reconocidas (Errattahi et. al, 2018).

Estas métricas se emplearon durante el proceso de entrenamiento, para obtener resultados antes y después de realizar el proceso de adaptación y entrenamiento del modelo acústico.

Fase 3: Evaluación del sistema RAV

En esta fase, se emplearon nuevamente las ecuaciones 1 y 2 con el propósito de evaluar los modelos acústicos y de lenguaje, inicialmente se analizaron los audios que fueron proporcionados por el sistema del Servicio Integrado de Seguridad ECU 911. Luego, con los resultados obtenidos de CMU-Sphinx, se utilizaron las ecuaciones con el fin de obtener los resultados de la tasa de error y precisión de las palabras que, a su vez, servirían para crear una tabla de resultados con la cual se evaluaría la eficacia del sistema de reconocimiento de voz para el dialecto español.

4. Resultados

Después del proceso para el desarrollo de sistema RAV y la transcripción de los audios con el script *word_align.pl*, se calculó la métrica WER, debido a que el modelo se basa en un reconocimiento del locutor dependiente de texto, y facilita la obtención de esta tasa.

Tabla 2. Resultados del modelo no entrenado.

Audios	WER (%)	WAcc (%)
3	19.97	80.20
15	82.93	17.07
20	97.71	2.29

En la tabla 2, se visualizan los resultados del WER en un modelo no entrenado de unos grupos de audios. En la primera fila el resultado de exactitud tiene un porcentaje alto debido a que el tamaño de su audio no sobrepasa los 4 segundos. Luego, se incorporó 15 archivos de

audio para transcribir. El tamaño de los audios oscila entre 90 segundos a 180 segundos, por tal motivo afecta a la transcripción y dando así un error del 97.71 %.

En la tabla 3, se representan los valores obtenidos utilizando un modelo entrenado, es decir, para que los resultados sean más precisos es necesario que el volumen de datos entrenados sea alto. El proceso de entrenamiento es un proceso largo que puede tardar horas, así como también pueden tardar semanas. Luego de realizar el entrenamiento con un grupo de audios recreados para la experimentación, así como los audios proporcionados por el Servicio Integrado de Seguridad ECU 911 se ha podido observar que a mayor cantidad de audios entrenados mejora la precisión de los resultados. De manera que, los datos con mayor precisión que se obtuvo fueron de 49.8 %. Indudablemente es posible mejorar el resultado de WER, pero se necesita de un tiempo considerable para realizar el entrenamiento.

Tabla 3. Resultados del modelo entrenado.

Audios	WER (%)	WAcc (%)
50	85.70	13
100	80.10	19.80
200	49.80	50.10

La tasa de error fue tabulada de acuerdo con la metodología planteada. Luego de analizar los resultados del modelo entrenado como el no entrenado, se observan valores notorios en cuanto a la exactitud de dichas transcripciones. Al ser un sistema sin conexión a internet resulta limitado al momento de la transcripción, y es así como se vuelve necesario realizar adaptaciones y entrenamientos a los modelos acústicos.

5. Conclusiones y recomendaciones

Se comprobó que los audios analizados con la herramienta de reconocimiento de voz CMUSphinx proporcionan una tasa de acierto variable puesto que esta depende primordialmente del entrenamiento del modelo acústico. Por tal motivo, un modelo acústico sin entrenamiento o con escaso entrenamiento implica que la tasa de error a calcularse sea elevada. Así pues, mientras se mantenga constantemente actualizado el vocabulario en el modelo, los resultados que se puedan obtener deben ser más aceptables, debido al cambio constante de la jerga popular y el constante cambio del nativo hablante. Cabe mencionar que los resultados se obtuvieron al imponer ciertas restricciones a los audios, por lo que se desconoce el nivel de precisión y de la tasa de error que obtendría esta herramienta al basarse en audios de larga duración, con más de un canal y en diferentes frecuencias.

Debido a que la herramienta de reconocimiento de voz CMUSphinx no posee conexión a internet, hay varias librerías en internet que pueden ayudar al desarrollo del entrenamiento del modelo acústico al basarse en grandes cantidades de datos proporcionadas por usuarios. De manera que, se muestra un tanto limitada en función de su capacidad de transcripción de texto y el nivel de eficiencia del sistema RAV. Sin duda alguna tiene la posibilidad de mejorar este sistema RAV si logra implementarse dentro de un entorno en línea.

Como trabajo futuro se planteará el análisis de nuevas herramientas para mejorar el sistema RAV en cuanto a su eficiencia y eficacia, además se incluirá la creación de un caso de estudio con las nuevas herramientas.

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 Ciencias de la Computación, y el Laboratorio de Investigación y Desarrollo en Informática - LIDI. De la misma manera, este trabajo se enmarca en el proyecto de investigación "Diseño de arquitecturas y modelos de interacción para entornos AAL dirigidos a adultos mayores: entornos lúdicos y sociales", por lo que agradecemos a la Universidad de Cuenca por su aporte.

Bibliografía

Aguiar de Lima, T., y Da Costa-Abreu, M. (2020). A Survey on Automatic Speech Recognition Systems for Portuguese Language and its Variations. *Computer Speech and Language*, *62*. https://doi.org/10.1016/j.csl.2019.101055

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., y Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Accedido 9:* 131858–131876. https://doi.org/10.1109/ACCESS.2021.3112535

Ali, A., y Renals, S. (2018). Word error rate estimation for speech recognition: E-wer. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, *2*(2014), 20–24. https://doi.org/10.18653/v1/p18-2004

Ankit, A., Mishra, S. K., Shaikh, R., Gupta, C. K., Mathur, P., Pawar, S., y Cherukuri, A. (2016). Acoustic Speech Recognition for Marathi Language Using Sphinx. *ICTACT Journal on Communication Technology*, *7*(3), 1361–1365. https://doi.org/10.21917/ijct.2016.0201

Celis, J., Llanos, R., Medina, B., Sepúlveda, S., y Castro, S. (2017). Acoustic and Language Modeling for Speech Recognition of a Spanish Dialect from the Cucuta Colombian Region. *Ingeniería*, 22(3): 362–376. https://doi.org/10.14483/23448393.11616

Belinkov, Y., y Glass, J. (2019). Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7, 49-72.

Dhankar, A. (2017). Study of deep learning and CMU sphinx in automatic speech recognition. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (2296-2301). IEEE.

Singh, R., Raj, B., y Stern, R. M. (2018). Model Compensation and Matched Condition Methods for Robust Speech Recognition. En *Noise Reduction in Speech Applications* (pp. 245-275)*. CRC press.*

Errattahi, R., El Hannani, A., y Ouahmane, H. (2018). Automatic Speech Recognition Errors Detection and Correction: A review. *Procedia Computer Science*,128: 32-37.

Peinl, R., Rizk, B., y Szabad, R. (2020). Open-source Speech Recognition on Edge Devices. En 2020 10th International Conference on Advanced Computer Information Technologies (ACIT) (pp. 441-445). IEEE.

Kim, D., Oh, J., Im, H., Yoon, M., Park, J., y Lee, J. (2021). Automatic Classification of the Korean Triage Acuity Scale in Simulated Emergency Rooms Using Speech Recognition and Natural Language Processing: A Proof of Concept Study. *Journal of Korean Medical Science*, *36*(27): 1-13. https://doi.org/10.3346/JKMS.2021.36.E175

Lakdawala, B., Khan, F., Khan, A., Tomar, Y., Gupta, R., & Shaikh, A. (2018). Voice to Text transcription using CMU Sphinx A mobile application for healthcare organization. *Proceedings of the International Conference on Inventive Communication and Computational Technologies, ICICCT 2018*, *Icicct*, 749–753. https://doi.org/10.1109/ICICCT.2018.8473305

Medina, F., Piña, N., Mercado, I., y Rusu, C. (2014). Reconocimiento de palabras en español con Julius. *ACM International Conference Proceeding Series*, *2241*. https://doi.org/10.1145/2590651.2590660

Peralta Vásconez, J. J., Narváez Ortiz, C. A., Orellana Cordero, M. P., Patiño León, P. A., y Cedillo Orellana, P. (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. https://doi.org/10.37815/rte.v33n2.840

Tavi, L., Alumäe, T., y Werner, S. (2019). Recognition of Creaky Voice from Emergency calls. *Proceedings of the Annual Conference of the International Speech Communication Association. INTERSPEECH*, *2019-Septe*: 1990-1994. https://doi.org/10.21437/Interspeech.2019-1253

Zhao, L., Alhoshan, W., Ferrari, A., Letsholo, K. J., Ajagbe, M. A., Chioasca, E.-V., y Batista-Navarro, R. T. (2020). Natural Language Processing (NLP) for Requirements Engineering: A Systematic Mapping Study. *Computing Surveys 54*(3): 1-41.