Modelos ocultos de Markov para el desarrollo de un sistema de ayuda al habla para personas que sufren de disartria

Clara Mercedes Guevara Vélez
meche18_1295@hotmail.com / Universidad Nacional Pedro Ruiz Gallo. Lambayeque, Perú

Recepción: 25-5-2018 / Aceptación: 21-8-2018

Resumen. El presente trabajo describe el diseño y desarrollo de un reconocedor automático de voz disártrica en español y una interfaz gráfica para su uso, el sistema en su totalidad será denominado Sistema de Ayuda al Habla (SAH). Esto para realizar funciones de apoyo en el nivel de comunicación de personas con el trastorno de disartria. Para la función de reconocimiento del sistema se utilizó la biblioteca HTK Toolkit, siendo la técnica de modelado acústico los modelos ocultos de Markov. En las pruebas con un vocabulario de cinco palabras, se tuvo un aumento del nivel de comunicación de un 46,67 %, y una tasa de reconocimiento de voz de 65 %. Estos resultados fueron comparando la precisión del SAH con la precisión del reconocimiento humano.

Palabras clave: disartria, reconocimiento automático de voz, HTK Toolkit, modelos ocultos de Markov (HMM), nivel de comunicación

Hidden Markov models for the development of a speech-aid system for people with dysar-thria

Abstract. The present paper describes the design and development of an automatic recognizer of speech in Spanish and a graphical interface for its use; the system as a whole will be called Speech Help System (SAH). It is intended to perform supporting functions in the communication level of people with dysarthria disorder. For the System Recognition function, the HTK Toolkit library was used; the acoustic modeling technique being the Hidden Markov Models (HMM). In tests with a vocabulary of five words, there was an increase in communication level of 46.67 %, and a voice recognition rate of 65 %. These results compared the accuracy of the SAH to the accuracy of human recognition.

Keywords: dysarthria, automatic speech recognition, HTK Toolkit, Hidden Markov Models (HMM), communication level

1. Introducción

Dentro del campo de la inteligencia artificial en ingeniería de sistemas, se busca desarrollar una aplicación directa para personas con discapacidades, en especial del habla. De acuerdo con las estadísticas mostradas por el Ministerio de Salud (MINSA), en el 2016 aproximadamente el total de personas con enfermedades y trastornos que ocasionaron deficiencias era de 117 593 personas, de las cuales un 6 % tiene trastornos de la comunicación. En números esto se traduce en 7 062 peruanos registrados en el MINSA con problemas de comunicación.

Dentro de las discapacidades del habla y comunicación se considera a la disartria, que se puede definir como el trastorno de la expresión verbal causado por una alteración del control muscular del mecanismo del habla, siendo este un problema del habla y no un problema del lenguaje.

Se obtuvo toda la información (bibliográfica y en línea) para conocer acerca de las patologías de lenguaje y motoras en general de las personas con disartria. Se reconoció la utilidad práctica de sistemas computacionales basados en inteligencia artificial que ayudan a los pacientes a comunicarse por medio de la voz en otros países.

Al tener dicha información se realizó investigación en el campo de aplicaciones de reconocimiento automática del habla (RAH), encontrándose proyecto de otros países enfocados en el desarrollo de sistemas para mejorar la comunicación de personas con disartria. Esto llevó a identificar los siguientes problemas relacionados con el desarrollo de un sistema RAH para voz disártrica:

- Tasas variables de precisión en el reconocimiento de voz (25 a 95 %) para usuarios disártricos.
- El rango de anormalidades en la voz disártrica es muy amplio, variando entre personas afectadas.
- Conforme el tamaño del vocabulario del sistema aumenta, el nivel de precisión de reconocimiento disminuye.
- Los síntomas asociados a la disartria dificultan la recopilación de muestras de voz (corpus) para un entrenamiento supervisado robusto del sistema.
- No existe un corpus de voz disártrica en español latino para la realización de análisis o modelado acústico para la construcción de un sistema RAH.
- No existen proyectos de sistemas RAH para el idioma español similares que sirvan como base de comparación, la mayoría están desarrollados para el idioma inglés.

La presente investigación pretende desarrollar un sistema de reconocimiento automático de voz disártrica denominado Sistema de Ayuda al Habla (SAH), superando los problemas

mencionados anteriormente para lograr el aumento del nivel de la comunicación por medio de voz de las personas que sufren del trastorno de disartria.

2. Metodología

2.1 Datos usados en el estudio

Adams (2017) nos describe los tipos de sistemas de reconocimiento de voz, entre ellos, el sistema de reconocimiento de voz dependiente del hablante. Nos dice que es un tipo de reconocimiento de voz que depende de la persona que habla. Requiere entrenamiento para ser más preciso en la conversión de voz a texto. El entrenamiento se realiza a menudo a través de una serie de muestras de conversión que luego son corregidas por el hablante.

Debido a que este proyecto se desarrollará como un sistema de reconocimiento de voz de tipo dependiente del hablante, se usarán cinco personas para entrenar y testear el sistema. En la fase de testeo, cada persona repetirá cinco palabras del vocabulario, cuatro veces cada una. En total se evaluarán 100 palabras emitidas.

Tomaremos como variable el puntaje obtenido en las evaluaciones con la escala analítica de los exámenes DELE, la cual será medida en dos momentos distintos, que en general llamaremos "antes del SAH" y "después el SAH". Estas variables serán comparadas entre sí, se comprobará la hipótesis siempre y cuando la variable "después del SAH" sea mayor a la variable "antes del SAH".

2.2 Vocabulario del Sistema

Maggiolo (2017) nos presenta el test de articulación a la repetición (TAR), de la profesora fonoaudióloga Edith Schwalm, el cual es comúnmente usado para medir el nivel fonético de los pacientes y así detectar posibles trastornos del lenguaje.

En este proyecto nos hemos basado en la clasificación de las palabras usadas en el TAR para crear un vocabulario que se adecue a las palabras más usadas en la vida diaria del usuario elegido. A continuación se presenta el vocabulario desarrollado:

- Cabeza
- Columba
- Elefante
- Gonzales
- Micrófono

2.3 Escala analítica en expresión e interacciones orales de los exámenes DELE

Se necesitó medir la calidad de la comunicación entre un paciente disártrico y una persona sin discapacidad. Para ello se decidió tomar como referente la escala analítica en expresión e interacciones orales de los exámenes DELE. Para esta medición se adaptó la escala DELE en contexto de conversación bilateral. En la tabla 1 se detallan los criterios de evaluación que se escogieron.

Tabla 1

Criterios de evaluación de conversación bilateral

Fluidez	
4	Se comunica con relativa fluidez, es capaz de mantener el ritmo eficazmente.
3	Se comunica con relativa facilidad aunque no lleva el ritmo de la conversación.
2	Las pausas son claras ya que tiene problemas para formular un discurso fácilmente entendible.
1	Le es imposible expresarse.
Pronunciación	
4	Su pronunciación es clara e inteligible.
3	Su pronunciación está distorsionada pero el receptor sí llega a entender.
2	Su pronunciación está demasiado distorsionada haciendo que el receptor se esfuerce para intentar comprenderlo.
1	Su pronunciación es totalmente ininteligible.
Interacción	
4	Conversa con relativa facilidad y eficacia, colabora con su receptor.
3	Mantiene la conversación en forma adecuada aunque en ocasiones el receptor le pide repetir lo dicho.
2	El receptor pide aclaraciones o repetir lo dicho constantemente.
1	No hay ningún tipo de entendimiento del mensaje dicho por parte del receptor.

Elaboración propia

2.4 Modelos ocultos de Markov

Un modelo oculto de Markov (MOM) es un proceso estocástico que consta de un proceso de Markov no observado (oculto) y un proceso observado O, cuyos estados son dependientes estocásticamente de los estados ocultos (figura 1). La tarea fundamental consiste en determinar los parámetros ocultos a partir de los parámetros observados. Una de las aplicaciones más utilizadas de estos modelos es el reconocimiento del habla, técnica que ha permitido modelar adecuadamente la gran variabilidad en el tiempo de la señal de voz.

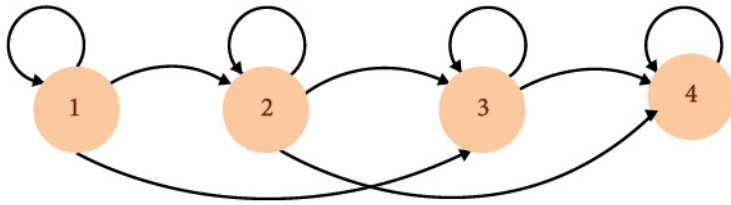

Figura 1. Modelo oculto de Markov
Elaboración propia

Esta arquitectura se da de acuerdo con el número de estados que la componen y las transiciones permitidas entre dichos estados. Se usará el modelo de Bakis, también conocido como modelo izquierda-derecha, pues la secuencia de estados oculta tiene la propiedad de que, conforme el tiempo se incrementa, el estado se incrementa (o permanece constante). Se utiliza en el modelado del habla porque se ajusta a señales cuyas propiedades cambian a lo largo del tiempo, como la voz. Los MOM hacen uso de dos algoritmos de análisis secuencial.

El primero es el algoritmo de Viterbi (figura 2), que nos permite encontrar la secuencia de estados más probable en un modelo MOM a partir de una observación; este obtiene la secuencia óptima que mejor explica la secuencia de observaciones. En este algoritmo se puede visualizar la mejor ruta a través de una matriz donde la dimensión vertical representa los estados de los MOM y la dimensión horizontal representa los marcos del habla (es decir, el tiempo).

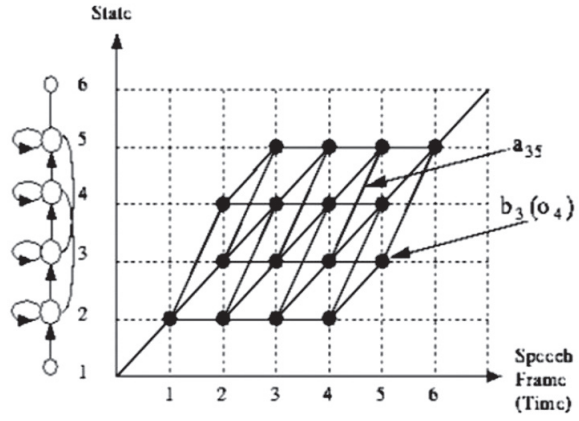

Figura 2. Algoritmo de Viterbi
Elaboración propia

El segundo es el algoritmo Baum Welch (figura 3), que se define como la reestimación de los parámetros de un MOM sobre la base de otro MOM. Uno de los problemas relacionados con los MOM es el de encontrar un modelo que maximice la probabilidad de una secuencia de observaciones, es decir, determinar el modelo que mejor explica tal secuencia. El problema es que no es posible encontrar tal modelo analíticamente y por ello es necesario un algoritmo iterativo como el de Baum y Welch, que permite estimar los parámetros de un modelo que hacen máxima la probabilidad de una secuencia de observables.

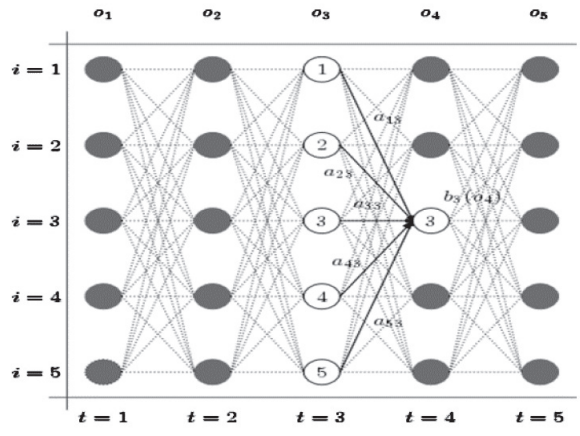

Figura 3. Algoritmo Baum Welch
Elaboración propia

2.5 Sistema de Ayuda al Habla

La construcción del SAH se realizó en dos partes; primero, se desarrolló el reconocedor de voz, con la construcción de diferentes modelos acústicos por cada palabra del vocabulario presentado anteriormente, haciendo uso de la biblioteca HTK Toolkit y del diagrama de flujo (Young et al., 2002).

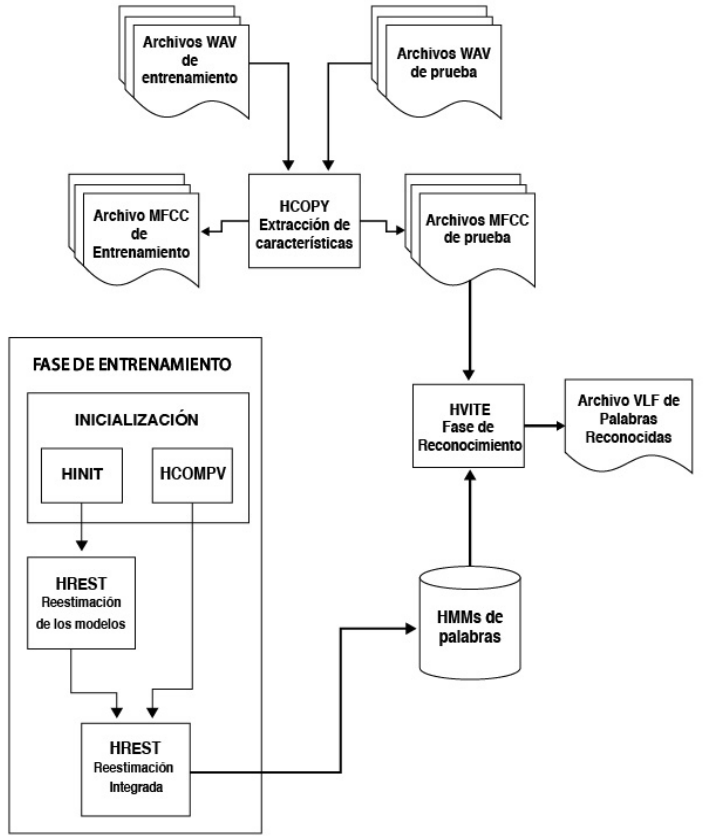

Figura 4. Diagrama de flujo
Elaboración propia

El diseño de esta herramienta tiene el propósito principal de construir modelos ocultos de Markov para el procesamiento de voz, específicamente para el diseño de reconocedores de voz. Como primer paso, las herramientas del HTK se utilizan para estimar los parámetros del conjunto de HMM, utilizando archivos de entrenamiento que contienen la pronunciación y su transcripción asociada. Posteriormente, las pronunciaciones desconocidas se transcriben a través de las herramientas del HTK para su reconocimiento.

Una vez que se han construido los modelos acústicos se realiza la segunda parte de la construcción, el desarrollo de la interfaz gráfica, la cual presenta tres etapas: grabación, reconocimiento y síntesis.

2.5.1 Etapa de grabación

La etapa de grabación le permitirá al usuario realizar la captura de la señal de voz y archivarla en un archivo de formato WAV. Está compuesta del interruptor ON/OFF y el cronómetro que se encuentra a su lado izquierdo. El usuario deberá activar el interruptor, posicionándolo en el estado ON, automáticamente se dará inicio al cronómetro y se empezará a grabar la señal de voz. Cuando el usuario haya terminado de hablar, deberá desactivar el interruptor, posicionándolo en el estado OFF. Automáticamente, se dará termino a la grabación y se creará el archivo WAV que contendrá la señal de audio capturada.

Figura 5. Interfaz gráfica Sistema SAH

Elaboración propia

2.5.2 Etapa de reconocimiento

La etapa de reconocimiento, como lo dice su nombre, realiza el proceso de reconocimiento, ejecuta las herramientas HCopy y HVite de la biblioteca HTK Toolkit para obtener la transcripción de la palabra reconocida en el audio previamente grabado. Consta del botón

RECONOCER, ubicado en la parte inferior de la interfaz, el cual, al ser pulsado, muestra en el área de texto que se encuentra en la parte superior de la venta, la palabra reconocida.

2.5.3 Etapa de síntesis

La etapa de síntesis se encarga de realizar la síntesis de texto a voz de la palabra mostrada en el área de texto. Si el usuario es incapaz de leer la palabra impresa o desea que sea reproducida, deberá pulsar el botón REPRODUCIR, el cual se encuentra ubicado debajo del área de texto donde se imprime la palabra reconocida.

Esta GUI es creada en el lenguaje de programación Python, y presenta el flujo de información que muestra la figura 6.

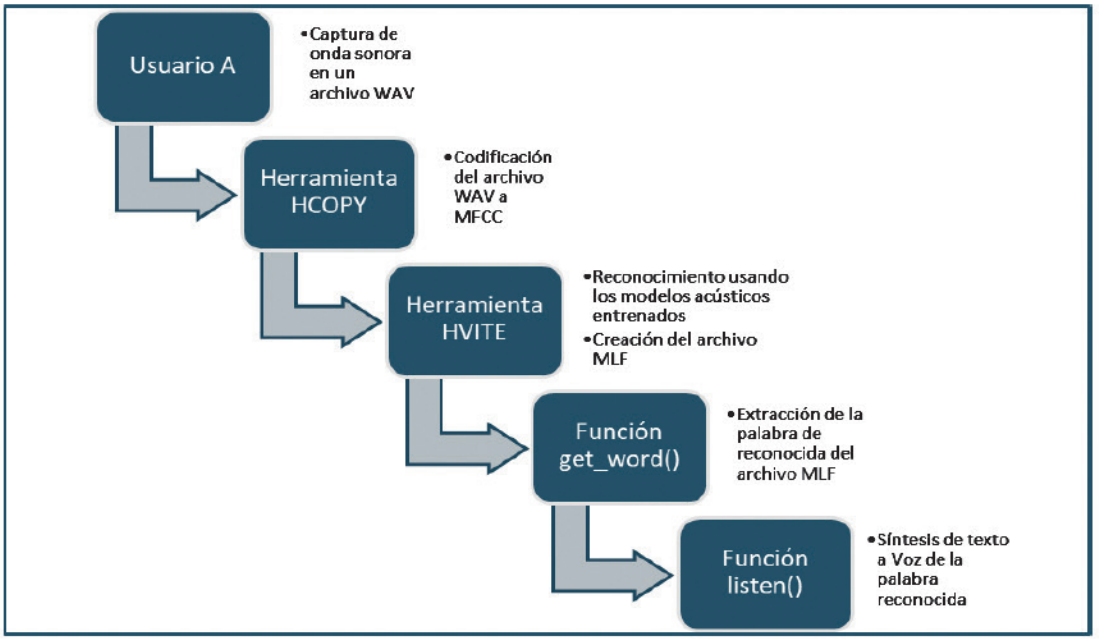

Figura 6. Flujo de información del sistema SAH
Elaboración propia

3. Resultados

El puntaje total de la escala analítica de expresión e interacción orales de los exámenes DELE, es de 12 puntos. En la tabla 2 se muestra la variación de resultados, en las dos fases: antes y después de la implementación del sistema SAH.

Tabla 2
Nivel de comunicación

	Antes de SAH (%)	Después de SAH (%)
Paciente CG	41,67	91,67
[illegible]	48,33	91,67
Paciente MS	41,67	85,00
Paciente BD	55,00	100,00
Paciente MV	43,33	95,00

Elaboración propia

Como se observa en la tabla 2, los porcentajes muestran que todos los pacientes presentaron un mayor nivel de comunicación, después de la implementación del sistema SAH.

El nivel de comunicación promedio en la fase "antes de SAH" es de 46 %, pero con la implementación del sistema SAH, esta cifra en la fase "después de SAH" aumentó a 92,67 %, mostrando una mejora del 46,67 %.

Se utiliza la siguiente fórmula para calcular la tasa porcentual de reconocimiento de un sistema RAH.

$$TR = \frac{número\ palabras\ correctamente\ reconocidas}{número\ total\ de\ palabras} x\ 100$$

Mostramos en la tabla 3 la cantidad de palabras correcta e incorrectamente reconocidas por el sistema con cada paciente.

Tabla 3
Registro de reconocimiento

	Correctos	Incorrectos
Paciente CG	15	5
Paciente ER	15	5
Paciente MS	14	6
Paciente BD	20	0
Paciente MV	17	3
Total	81	19

Elaboración propia

Al aplicar la fórmula anterior, se calcula una tasa de reconocimiento del 81 %, el cual es un porcentaje cercano a las tasas de reconocimiento de otros sistemas de reconocimiento de voz comerciales.

4. Discusión

Antes de la implementación del sistema SAH y durante la realización de las pruebas se fueron obteniendo las siguientes observaciones en la comunicación de los pacientes.

- Los pacientes tenían como un límite de repeticiones por palabra un promedio de tres veces; repetir más veces los llevaba a la frustración por no ser entendidos y se negaban a seguir intentándolo, truncando el proceso de comunicación.
- Muchas veces la oyente tenía que interpretar adivinando o tanteando en palabras similares a las que escuchaba hasta que diera con la palabra correcta.
- Los pacientes se quejaban de dolor al tener que articular con exageración o tener que aumentar el volumen de su voz para lograr ser entendidos.

Después de la implementación de sistema SAH y mientras se realizaban las pruebas se fueron obteniendo las siguientes observaciones en la comunicación de los pacientes.

- Los pacientes mostraron mucha afinidad con el uso del sistema SAH, se sentían más motivados a intentar emitir más palabras.
- La interacción con el oyente fue mayor, ya que tomaba menos tiempo para el oyente entender lo que el paciente le decía.
- No hubo ningún tipo de frustración por parte de los pacientes, por lo que en ningún caso la comunicación fue interrumpida.

5. Conclusiones

El nivel de comunicación tuvo un aumento de 46,67 % puntos, es decir, tuvo un incremento del 46 % al 92,67 % de la escala analítica de los interacción y expresión oral de los exámenes DELE, después de la implementación del SAH.

Para el desarrollo del SAH se abordó el enfoque de dependiente de usuario (DU). En los trabajos citados se ha argumentado que los sistemas RAH dependientes de usuario son mejores para individuos con disartria, aunque requieren más tiempo y trabajo de cercanía con el paciente para ser desarrollados. Esto porque este enfoque implica el desarrollar un corpus de entrenamiento con la voz del usuario que va a usar el sistema, el cual debe estar etiquetado correctamente.

De esta manera se considera que este sistema, con las funciones implementadas, contribuye al campo del desarrollo de tecnologías para personas con discapacidad en el habla. Siendo que este campo no se ha explorado de manera significativa para el español latino.

Referencias

Adams, C. (17 de marzo de 2017). *Explaining the different types of voice recognition.* Recuperado de https://www.thoughtco.com/types-of-voice-recognition-1205856

Maggiolo, M. (2017). Test de articulación a la repetición (TAR): Un legado de la profesora fonoaudióloga Edith Schwalm. *Revista Chilena de Fonoaudiología 16*. https://doi.org/10.5354/0719-4692.2017.47557

Ministerio de Salud-Minsa (2016). Datos estadísticos sobre incapacidad. Recuperado de https:www.inei.gob.pe/estadísticas/índice-tematico/discapacidad-7995

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., ..., Valtchev, V. (2002). The HTK Book. *Cambridge University Engineering Department 3*, p. 175.

Bibliografía

Álvarez, A. (2001). *Apuntes de fundamentos del reconocimiento automático de la voz.* Madrid: Departamento de Arquitectura y Tecnología de Sistemas Informáticos. Facultad de Informática. UPM.

Davis, S., y Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing 28*(4), pp. 357–366. https://doi.org/10.1109/TASSP.1980.1163420

Hatzis, A. (2003). Automatic speech recognition with sparse training data for dysarthric speakers. *Proc. European Conference on Speech Communication Technology*, pp. 1189–1192.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., ..., Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine 29*(6), pp. 82–97. https://doi.org/10.1109/MSP.2012.2205597

Hubert, P. (2017). *PyAudio Documentation - PyAudio 0.2.11 documentation*. Recuperado de https://people.csail.mit.edu/hubert/pyaudio/docs/

Instituto Cervantes. (S. f.) *Qué son los DELE. Exámenes*. Recuperado de https://examenes.cervantes.es/es/dele/que-es

Jurafsky, D., y Martin, J. H. (2008). *Speech and language processing*. Segunda edición. Nueva Jersey: Prentice Hall.

Kadi, K. L., Selouani, S. A., Boudraa, B., y Boudraa, M. (2016). Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. *Biocybernetics and Biomedical Engineering 36*(1), pp. 233–247.

Lambayeque: Más de 100 mil tratamientos en medicina física se realizaron en hospital. (26 octubre de 2015). Diario *Correo*. Recuperado de https://diariocorreo.pe/edicion/lambayeque/lambayeque-mas-de-100-mil-tratamientos-en-medicina-fisica-se-realizaron-en-hospital-628286/

Lizandra Laplaza, R. (S. f.). *Dificultades en el desarrollo del lenguaje oral e intervención*. Recuperado de https://www.yumpu.com/es/document/ view/14250103/dificultades-en- el-desarrollo-del-lenguaje-oral-e-intervencion/14

Moriana, M. J. (2009). La disartria. *Revista Digital Innovación y Experiencias Educativas.* Recuperado de https://fonologizatte.wordpress.com/trastornos-del-habla/

Peña-Casanova, J. (2013). *Manual de logopedia*. Barcelona: Elsevier.

Pérez, A. M. S., Fernández, M. D. V., y Torres, I. H. (2006). La comunicación oral, sus características generales. *Ciencias Holguín 12*(2), pp. 1-6. Cuba.

Samira, H., Fateh, B., Smaine, M. y Mohamed, B. (2013). A novel speech recognition approach based on multiple modeling by hidden Markov models. En: *2013 International Conference on Computer Applications Technology (ICCAT)* (pp. 1-6). https://doi.org/10.1109/ICCAT.2013.6522028

Sánchez, M. G. (2008). *Desórdenes motores del habla y PROMPT (Parte II)*. EspacioLogopedico.com. Recuperado de https://www.espaciologopedico.com/ revista/articulo/1493/desordenes-motores-del-habla-y-prompt-parte-ii.html

Tamayo, M. T. (2004). *El proceso de la investigación científica*. México: Editorial Limusa.