Eñe'ẽ: Sistema de reconocimiento automático del habla en Guaraní

Diego Manuel Maldonado, Rodrigo Villalba Barrientos, Diego P. Pinto-Roa
Facultad Politécnica, Universidad Nacional de Asunción
San Lorenzo - Paraguay
Email: dimaldon@gmail.com, rodrigovb@yahoo.com, dpinto@pol.una.py

Resumen El guaraní es un idioma hablado por alrededor de ocho millones de personas. En Paraguay, cerca del 25 % de la población habla solamente guaraní. Con el incremento del número de dispositivos que incluyen asistentes personales controlados por voz, la necesidad de desarrollar aplicaciones con interfaces de voz está en crecimiento constante. A pesar de ser un idioma muy hablado, actualmente no existen investigaciones respecto al reconocimiento automático del habla en guaraní. En este contexto, proponemos un sistema de reconocimiento del habla en guaraní denominado Eñe'ẽ.

Construimos el modelo utilizando las herramientas que ofrece el software de código abierto CMU Sphinx. Para ello, se recolectaron muestras de voz en guaraní, mediante el desarrollo de una aplicación web que permite a los interesados enviar lecturas de oraciones en guaraní que se muestran en pantalla.

Este paper muestra los resultados de las pruebas experimentales que validan la eficiencia de la primera versión de Eñe'ẽ, y fomenta el desarrollo de sistemas utilizando este reconocedor de propósito general para futuras aplicaciones específicas.

Automatic Speech Recognition, Hidden Markov Models, Guarani Language.

1. Introducción

En Paraguay, alrededor del 80 por ciento de la población habla guaraní y el nivel del bilingüismo está por encima del 50 por ciento [1]. El idioma es hablado por aproximadamente ocho millones de personas en América del Sur, y no puede ser ignorado por la tecnología para desarrollar nuevas aplicaciones.

La última década ha sido testigo de un progreso sustancial en la tecnología de reconocimiento de voz, que aumentará aún más su presencia en nuestras vidas cotidianas con la multiplicación de los smartphones y las tendencias de Internet of Things. El problema del reconocimiento automático del habla (ASR, por sus siglas en inglés) ha evolucionado a partir de una máquina simple que responde a un pequeño conjunto de sonidos a un complejo sistema que habla con fluidez.

El objetivo de este trabajo es desarrollar un software que permita realizar transcripciones de las elocuciones habladas en guaraní.

El procesamiento y análisis de señales, en especial aplicados al reconocimiento del habla, apoyados en una base matemática funcionan bien en la práctica.

El trabajo se divide en tres fases: recolección de datos, entrenamiento, e implementación. Previamente, se presenta una comparación entre algunas de las técnicas de modelado para reconocedores del habla más utilizadas, y una breve descripción de las herramientas de desarrollo de software para construir nuestro sistema ASR. Para desarrollar el sistema propuesto, utilizamos los modelos ocultos de Markov (HMM, por sus siglas en inglés) mediante la plataforma open source CMU Sphinx. Las propiedades de los HMM las hacen ideales para resolver el problema de reconocimiento. Esta técnica ha proporcionado satisfactorios resultados hasta la fecha tanto para el reconocimiento de habla aislada como continua [2], [3]. Pero los reconocedores del habla basados en HMM siguen siendo difíciles de construir, principalmente porque requieren enormes cantidades de datos de entrenamiento.

El trabajo se organiza de la siguiente forma: en la sección II se definen algunos conceptos generales, en la sección III se describen las ventajas e inconvenientes de la aplicación de algunas de las técnicas más utilizadas en el reconocimiento del habla y en la sección IV se describe la herramienta CMU Sphinx. En la sección V se presenta una descripción del idioma guaraní, en la sección VI se describe la propuesta tecnológica, en la sección VII se exponen los resultados y en la sección VIII se presentan las conclusiones y trabajos futuros.

2. Conceptos básicos

El reconocimiento automático del habla es un proceso por el cual una máquina identifica una frase. La máquina toma una expresión hablada como entrada y devuelve una cadena de palabras, frases en forma de texto como salida [4]. Es importante para desarrollar un buen modelo, disponer de una suficiente cantidad de datos, de lo contrario, las aproximaciones no producirán los resultados esperados. Para establecer una relación entre un grupo de fonemas y una oración hablada se utilizan los algoritmos de entrenamiento supervisado; que consisten en un grupo de instrucciones que utilizan un conjunto de datos, y se tiene cómo deberían ser las salidas correctas para cada entrada. La técnica consiste en generar la relación que aproxime con la mayor precisión posible la salida de una entrada futura, por ello, los datos demandan cantidades sustanciales de transcripciones manuales y supervisión para el proceso de entrenamiento.

En aprendizaje automático, las funciones generadas por el entrenamiento (“funciones discriminantes” o “funciones hipótesis”) pueden ser lineales o no lineales, y las salidas pueden ser binarias, números enteros o valores reales. Los algoritmos de entrenamiento actualizan los parámetros de la función discriminante a medida que llegan nuevos datos.

Los modelos de algoritmos de aprendizaje supervisado pueden tener las siguientes funciones:

- Regresión o aproximación: salida continua (ej: regresión lineal, regresión polinomial).

- Clasificación: salida binaria, discreta (ej: regresión logística, árbol de decisiones).

Los reconocedores del habla se comportan como clasificadores, que calculan una probabilidad de correspondencia de un conjunto de fonemas a un vector de características espectrales. Se asocia un segmento del vector de características con un fonema que representa ese segmento del habla.

3. Modelos ocultos de Markov

El reconocimiento del habla es un problema que ha sido abordado utilizando técnicas de regresión lineal y codificación predictiva lineal [5], redes neuronales [4] y HMM [6], [2]. Cada modelo tiene tanto sus ventajas como inconvenientes en función de los objetivos del modelado, muchos sistemas combinan varias técnicas para reconocer palabras. A continuación se presenta una descripción de tres de los algoritmos de entrenamiento supervisado más utilizados en el reconocimiento del habla.

Las técnicas de reconocimiento basadas en HMM fueron un gran avance en grandes vocabularios y habla continua. La técnica de HMM ha sido ampliamente aceptada en los modernos sistemas ASR, principalmente por dos razones: su capacidad para modelar las dependencias no lineales de cada unidad, y además se tiene un enorme conjunto de enfoques analíticos para la estimación de los parámetros de modelos. Este trabajo aplica técnicas de modelado de HMM con enfoques basados en las probabilidades y n-gramas, utilizando CMU Sphinx. CMU Sphinx permite entrenar modelos acústicos tomando como entrada archivos de audio con su correspondiente transcripción en texto. Un sistema ASR, como CMU Sphinx 4, usa tres tipos de modelos que dependen del idioma:

- Modelo acústico: Contiene una representación estadística de un sonido o fonema, creado usando muchos datos acústicos.
- Modelo de Lenguaje: Representa estadísticamente la probabilidad de ocurrencia de las palabras. Cada palabra en el modelo de lenguaje debe estar en el diccionario.
- Diccionario: Determina cómo se pronuncia una palabra en términos de los fonemas base [7].

En modelos ocultos de Markov, el problema se resume en la búsqueda de la oración más probable perteneciente a un lenguaje L, dada la entrada acústica X.

El modelo se define según un conjunto de estados ocultos Q, una matriz de transición de probabilidades A, una matriz B de emisión de probabilidades y una distribución de probabilidades de estados iniciales Π .

Así, un modelo oculto de Markov se denota $\lambda = (A, B, \Pi)$.

Para resolver el problema de ASR, se utilizan tres algoritmos en cada fase del modelado.

- Evaluación: Dada una secuencia de observaciones acústicas O y un modelo oculto λ, elegir cómo evaluar eficientemente $P(O|\lambda)$.
- Decodificación: Dados O y λ, elegir la secuencia de estados ocultos Q de manera tal a obtener $P(O|\lambda)$.
- Entrenamiento: Ajustar los parámetros de $\lambda = (A, B, \Pi)$ para maximizar $P(O|\lambda_w)$.

En la siguiente sección se profundizará más en los algoritmos que utiliza CMU Sphinx para resolver los problemas asociados a HMM.

La secuencia de observaciones $O = o_1, o_2, ..., o_T$ se obtiene a partir de la entrada acústica X, donde la señal fue dividida en T muestras o_i de igual duración.

La oración de salida w compuesta por M palabras, puede representarse como $w = w_1, w_2, ..., w_M$. La relación entre O y $w*$ puede ser aproximada mediante los modelos ocultos de Markov, donde $w*$ es la aproximación de w.

Se tiene un conjunto de palabras y un conjunto de entrenamiento para cada palabra, y se construye un modelo oculto de Markov para cada palabra utilizando el conjunto de entrenamiento asociado. Si λ_w es el conjunto de parámetros del modelo acústico asociado con una palabra w, entonces cuando se tiene una secuencia de observaciones O se tiene una salida según la probabilidad.

$$w^* = \arg\max P(O|\lambda_w) \quad (1)$$

El sistema debe preguntarse dado un modelo acústico, cuál es la probabilidad de que cierto segmento de la señal de la elocución representada corresponda a cada fonema. El argumento que obtenga la probabilidad máxima definirá el mismo, y dada una secuencia de fonemas, según el diccionario y el modelo de lenguaje, se obtendrán la palabras más probables que conforman $w*$ a la salida en forma de texto.

Si bien los HMM funcionan bien en la práctica, son difíciles de optimizar de manera eficiente para tareas posteriores, como el Spoken Language Understanding (SLU) [6].

4. Reconocimiento utilizando CMU Sphinx

La idea de implementar la teoría de los HMM en el reconocimiento del habla ya data de los años 1970, pero los primeros resultados satisfactorios en reconocimiento de habla continua recién se dieron en la década de los 90. Precisamente, la cantidad de años de desarrollo de esta técnica aplicada al reconocimiento del habla continua es una de las ventajas más importantes respecto a las demás técnicas. La idea general del ASR basado en HMM se ilustra en la Figura 4.

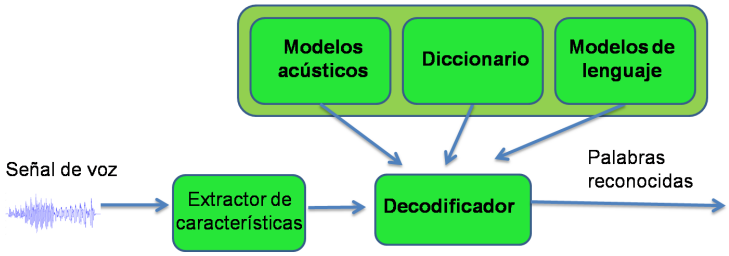

Figura 4. Esquema de un sistema reconocedor del habla.

Para desarrollar un reconocedor de voz en un idioma cualquiera, es necesario construir un modelo acústico que establezca una relación entre un archivo de sonido y una cadena que represente a un fonema. Se puede decir que el modelo acústico se trata del componente central de cualquier sistema de reconocimiento del habla [3].

Primeramente por un proceso de cuantización vectorial, la señal acústica se convierte en un vector numérico y se realiza una extracción de características.

El extractor de características debe pasar una porción de la información de entrada que represente correctamente los datos de una muestra al predictor. Esto es necesario puesto que se necesita procesar la entrada de modo a mejorar el algoritmo en términos de tiempo de cómputo.

El vector de características es la entrada de la siguiente fase, la decodificación, que depende de un modelo de lenguaje y un modelo acústico.

Un fonema corresponde a un estado del modelo oculto de Markov del modelo acústico. Al obtener la secuencia de estados Q más probable dada una secuencia de observaciones, se obtiene una secuencia de fonemas, pero la secuencia de fonemas no es la salida deseada. Luego, mediante el modelo de lenguaje se realizan las correspondencias entre palabras y secuencias de fonemas.

En la Figura 5 se muestra cómo se forma el modelo oculto de Markov al calcularse la secuencia de fonemas más probables dada una entrada acústica “Hola”. Cada fonema es representado mediante un estado del modelo oculto de Markov, en este caso la secuencia de estados forman los sonidos de “ola”; pero “ola”, que es una secuencia de fonemas no es la salida, sino que se muestra la palabra “Hola”para un reconocedor del habla con modelo de lenguaje en español.

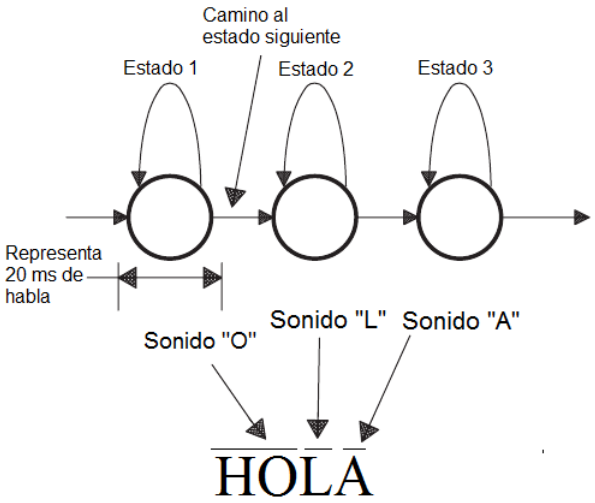

Figura 5. Relación entre secuencia de fonemas y una palabra.

Al finalizar la fase de decodificación se obtiene la secuencia de palabras más probable dados los vectores de características espectrales, los cuales se computaron según la entrada acústica durante la fase de extracción de características. Con esto llega a su culminación el proceso básico del reconocimiento del habla.

CMU Sphinx es una herramienta de reconocimiento de voz que tiene como objetivo ayudar en la creación de aplicaciones con habla incorporado. La herramienta posee varias aplicaciones orientadas a diferentes tipos de tareas:

- Pocketsphinx: Librería ligera para el reconocimiento de la voz escrita en el lenguaje C.
- Sphinxbase: Librería base requerida por las demás aplicaciones.
- Sphinx4: Reconocedor ajustable y modificable escrito en Java.
- Sphinxtrain: Herramienta para el entrenamiento del modelo acústico.

Para el problema de entrenamiento, CMU Sphinx utiliza el algoritmo de Baum-Welch. Con suficientes datos y potencia de computación, los algoritmos de Baum-Welch y el modelo oculto de Markov pueden calcular las probabilidades para construir el reconocedor [8]. Para la decodificación, utiliza algoritmos de Viterbi [9], y para la evaluación, aproxima mediante el algoritmo forward-backward [10].

Para construir el modelo de lenguaje, CMU Sphinx utiliza n-gramas. La utilización de n-gramas para construir modelos de lenguaje se basa en la propiedad de Markov en el que cada estado depende sólo del estado inmediatamente anterior, en este caso, cada palabra depende sólo de las n-1 palabras precedentes. Un n-grama es una secuencia de n elementos de un texto. Para su aplicación a los modelos de lenguaje, cada elemento corresponde a una palabra perteneciente al diccionario. La técnica basada en n-gramas es la predominante para construir modelos de lenguaje, debido a su simplicidad y efectividad [3].

5. El idioma guaraní

De [1] se puede obtener la Tabla 1, que ilustra la realidad lingüística paraguaya, en el que se muestran los porcentajes de la población paraguaya con su correspondiente idioma con el que más se comunican.

Tabla 1. Las cifras muestran la cantidad de personas que hablan guaraní en Paraguay.

Guaraní	27%
Bilingüe (guaraní-castellano)	59%
Bilingüe (castellano-guaraní)	26%
Castellano	8%

Los números muestran el potencial impacto social que tendrá el sistema Eñe'ẽ, teniendo en cuenta que es el primer reconocedor del habla en guaraní en territorio paraguayo.

El idioma guarani consta de 33 fonemas, 12 son vocales y 21 consonantes. A cada fonema le corresponde un único grafema, lo que facilita el trabajo del reconocimiento del habla respecto a otros idiomas como el español o inglés.

En Paraguay, la mayoría de las personas hablan guaraní mezclado con el castellano [11]. A la resultante de esta fusión morfosintáctica, gramatical y semántica del idioma guaraní con el castellano se lo denomina popularmente como "jopara", que en guaraní significa "mezcla". El reconocedor propuesto funciona solamente con el idioma guaraní, pero podría ampliarse a un reconocedor en jopara.

Un ejemplo de oración en jopara es: "ajogua che telerã", que significa "compré para mi tele". Un ejemplo de oración en guaraní es: "ñane ñe'ẽ iporã", que significa "nuestro idioma es lindo".

Generalmente, el jopara es el guaraní que recurre a palabras del castellano ante alguna necesidad para expresarse. Aunque que el jopara sea un fenómeno tan ampliamente difundido no se debe a la supuesta incapacidad estructural del guaraní de adaptarse a la vida moderna [11]. Por ejemplo, "ta'anga mýi"significa "cine", existen otros ejemplos que demuestran que el guaraní puede adaptarse a palabras modernas.

6. Propuesta tecnológica

El trabajo se centra principalmente en aproximar un modelo acústico óptimo mediante la ejecución eficiente de algoritmos de entrenamiento.

Las incorrectas suposiciones de simplificación en el proceso de construcción del algoritmo, daría lugar a un sistema de pobre precisión y sensibilidad inaceptable a los cambios en el entorno operativo o ruido [2].

Cuando la gente pronuncia una misma palabra, las señales acústicas no son idénticas, y de hecho incluso pueden ser notablemente diferentes. Es por ello que, para desarrollar un modelo acústico con el que se puedan dictar oraciones en un idioma, pruebas experimentales demuestran que es necesario entrenar el modelo con audios de por lo menos cincuenta horas de grabaciones de doscientas personas, para luego ser convertidas al formato adecuado, con una frecuencia de muestreo de 16 kHz de un sólo canal [1].

Uno de los mayores inconvenientes para la construcción del modelo acústico es la obtención de datos. Para recolectar la mayor cantidad de datos de entrenamiento posibles, se desarrolló una aplicación web https://www.eñee.xyz, mediante la cual las personas interesadas pueden enviar muestras de voz leyendo las oraciones en guaraní que se les muestra en pantalla.

La aplicación web fue desarrollada sobre el framework de código abierto Django y con la librería Pocketsphinx. La misma provee de una interfaz sencilla que cumple dos propósitos: la recolección de archivos de voz para el entrentamiento y la puesta en línea de la primera versión del reconocedor. La página web hace uso de la API WebRTC (actualmente soportada por Google Chrome, Mozilla Firefox, Opera y Microsoft Edge) para realizar la grabación de las muestras de voz en el lado del cliente. Una vez finalizada la grabación, la misma es enviada al servidor para su posterior verificación y aceptación.

Para la primera versión del sistema, se logró recolectar 1000 oraciones de 114 hablantes durante un período de 40 días, supervisadas por inteligencia humana, contabilizando 54 minutos de entrenamiento. Estas oraciones fueron clasificadas basándose en la correspondencia entre la oración a ser leída y su grabación, y en la calidad de las mismas, descartando las que hayan sido grabadas en un ambiente ruidoso.

Una vez finalizada la verificación, se procedió a convertir las muestras de voz al formato adecuado. Finalmente, con todos los datos preparados, se procedió a realizar el entrenamiento del modelo acústico. En base a los modelos creados, se desarrolló el sistema de reconocimiento del habla en guaraní.

El sistema permitirá el desarrollo de aplicaciones con fines prácticos que ayudarán a acortar la brecha tecnológica existente entre zonas rurales y urbanas, puesto que en las zonas rurales la mayoría de las personas hablan solamente guaraní. Además, será posible desarrollar aplicaciones útiles para personas con discapacidad o personas que hablan pero no escriben guaraní.

Podría por ejemplo ser utilizado para desarrollar tecnologías en los hospitales, donde los pacientes puedan regular la temperatura de la habitación, la inclinación de la cama u otros servicios que requieren asistencia humana. Además en ASR se tienen aplicaciones relacionadas a telefonía, que utilizan el reconocimiento del habla para manejar tareas tradicionalmente delegadas a un operador, como consultas de facturación y asistencia de llamadas. Las aplicaciones de re-

[1] http://cmusphinx.sourceforge.net/wiki/tutorialam

conocimiento de voz permiten el enrutamiento de llamadas, servicios bancarios, automatización de consulta de directorio y abren una posibilidad de una gran cantidad de otros servicios.

A partir del reconocedor en guaraní se podrá desarrollar aplicaciones de Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés), que analicen las salidas de texto para mejorar la experiencia de interacción con la máquina, o aplicaciones SLU como Microsoft Cortana, Google Now, Apple Siri. En los sistemas SLU y NLP, la transcripción de las elocuciones son sólo una parte de todo el gran proceso que permite interactuar con los dispositivos. Además se podrá desarrollar aplicaciones de traducción automática del habla.

Permitirá además, por ejemplo, integrar el reconocedor a traductores, realizar búsquedas por voz, comandar por voz sistemas inteligentes y desarrollar tecnologías educativas para personas monolingües en guaraní. Este trabajo, facilita y fomenta el desarrollo de los sistemas de reconocimiento del habla en guaraní, mediante la puesta a disposición de una API, que logrará que los desarrolladores puedan acceder al servicio en línea de transcripción de elocuciones en guaraní para fines específicos.

7. Pruebas experimentales

Para probar la primera versión del sistema, se realizó una validación cruzada de múltipes iteraciones.

Se tienen unos pares de entradas-salidas como conjunto de entrenamiento, y normalmente se reserva 5 % o 10 % de los datos, que no se utilizan para entrenar el modelo pero se utilizan como conjunto de prueba. El modelo entrenado se prueba con estos pares de entradas-salidas reservados y el error promedio observado se toma como una aproximación de la tasa de error real del sistema.

Para mejorar el resultado, se puede utilizar una validación cruzada de k iteraciones. El conjunto de datos se divide en k subconjuntos al azar del mismo tamaño. El modelo se entrena k veces, y se prueba con el subconjunto que fue dejado de lado en cada iteración, y se toma el promedio de las mediciones como la tasa de error [12].

Las muestras de voz como archivos de audio se agruparon en $k = 10$ subconjuntos, de los cuales un conjunto se utilizó como conjunto de prueba, y nueve subconjuntos como conjuntos de entrenamiento para cada iteración de la validación cruzada.

Para medir el error con la validación cruzada, tomamos en cuenta como parámetro la tasa de error de palabras (WER, por sus siglas en inglés)[13].

En total se utilizaron mil oraciones en guaraní para realizar la validación experimental.

Los resultados arrojaron una tasa de error de palabras del 9,29 %. En la Figura 6 se muestran los resultados obtenidos al realizar la validación cruzada, teniendo como parámetro la tasa de error de palabras en función a la cantidad de oraciones utilizadas para el entrenamiento.

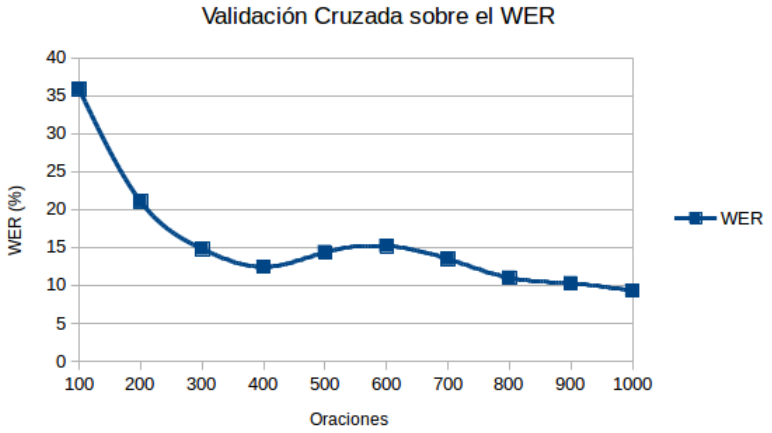

Figura 6. Tasa de error de palabras en función a la cantidad de oraciones.

La validación cruzada nos permitió conocer la calidad del reconocedor entrenado en su primera versión, con sólo una porción del volumen de datos acuśticos de entrenamiento necesarios, que continuaremos recolectando de manera a mejorar la performance del sistema propuesto. El sistema Eñe'ẽ en su primera versión cuenta con un vocabulario de 678 palabras.

8. Conclusiones y Trabajos Futuros

Este trabajo presenta los resultados iniciales del desarrollo de un modelo acústico en guaraní mediante la utilización eficiente de algoritmos de entrenamiento.

A partir de esta herramienta Eñe'ẽ, se podrán desarrollar nuevas aplicaciones utilizando como base el modelo acústico propuesto. Pruebas experimentales validan el potencial del sistema ASR, que tendrá un impacto a nivel tecnológico nacional puesto que el guaraní es un idioma hablado por la amplia mayoría de los paraguayos.

Se plantea como trabajo futuro la puesta a disposición de una API de un software de reconocimiento del habla en guaraní y la aplicación de técnicas de Procesamiento de Lenguaje Natural en guaraní, para mejorar la interacción con las máquinas, como también mejorar el desempeño de Eñe'ẽ y ampliar a un reconocedor del habla jopara. Un aspecto fundamental sería la comparación de los distintos algoritmos en el estado del arte para reconocimiento del habla en guaraní.

Referencias

1. B. Melià, "Paraguay multicultural y plurilingüe," 2011.
2. L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
3. S. Laviosa and A. Benítez, "Proyecto de reconocimiento automático del habla," 2015.
4. N. U. Maheswari, A. P. Kabilan, and R. Venkatesh, "Speaker independent speech recognition system using neural networks," , 2009.
5. C. A. de Luna Ortega, M. G. Mora, J. C. M. Romo, and F. J. L. Rosas, "Reconocedor de palabras con el uso de regresión lineal y coeficiente muestral," *Conciencia Tecnológica*, no. 44, pp. 5–9, 2012.
6. A. L. Maas, Z. Xie, D. Jurafsky, and A. Y. Ng, "Lexicon-free conversational speech recognition with neural networks," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015.
7. H. Satori, M. Harti, and N. Chenfour, "Introduction to arabic speech recognition using cmusphinx system," *arXiv preprint arXiv:0704.2083*, 2007.
8. J. A. Bilmes *et al.*, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.
9. P. Lamere, P. Kwok, W. Walker, E. B. Gouvêa, R. Singh, B. Raj, and P. Wolf, "Design of the cmu sphinx-4 decoder.," in *INTERSPEECH*, Citeseer, 2003.
10. K.-F. Lee, H.-W. Hon, and R. Reddy, "An overview of the sphinx speech recognition system," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 38, no. 1, pp. 35–45, 1990.
11. W. Lustig, "Mba'éichapa oiko la guarani? guaraní y jopara en el paraguay," *en: Ñemity*, vol. 33, no. 2, pp. 12–32, 1996.
12. G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

13. Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pp. 577–582, IEEE, 2003.