

SISTEMA INTELIGENTE DE TRADUCCIÓN DE VOZ A LENGUAJE DE SEÑAS BOLIVIANO

Carlos Walter Pacheco Lora Ph.D. ¹, Francisco Javier Camacho Dorado ²,
Juan Carlos Labrandero Cervantes ²

¹ Ingeniero Universidad San Francisco Xavier de Chuquisaca

² Universitario Universidad San Francisco Xavier de Chuquisaca

Dirección: Regimiento Campos 180

RESUMEN

El presente trabajo de investigación científica, emplea un modelo de red neuronal Contrapropagación (CPN) como mecanismo, que permite lograr la transformación de un fonema resultado de la emisión sonora de una letra, sílaba o palabra en una representación gráfica simbólica correspondiente al alfabeto de señas boliviano que emplean para comunicarse las personas sordas en Bolivia. Para materializar este trabajo de investigación se desarrollaron componentes de software orientados a la captura y procesamiento de los fonemas basados en principios matemáticos, físicos, lógicos y algorítmicos correspondientes al estudio de la voz, el tratamiento del sonido y las redes neuronales, los cuales fueron implementados en un componente de software que permite en principio memorizar un conjunto de letras, sílabas y palabras, para posteriormente, ante la emisión sonora de las mismas por cualquier persona de manera arbitraria, representarlas gráficamente a través de un símbolo del lenguaje de señas boliviano. El trabajo de investigación permitió detectar los patrones básicos que forman los fonemas, además se empleó el análisis de Fourier para señales, permitiendo extraer características en dominio de la frecuencia de la señal de audio original, para que sean usados conjuntamente con los patrones originales en un modelo de red neuronal más avanzado en la detección de voz y en la separación de sílabas de la señal de audio. Un estudio y valoración de varios modelos de redes neuronales permitió decidir utilizar una red de contrapropagación, complementando la medida con experiencias recogidas de otras investigaciones que mostraban que la CPN presentaba la mejor relación entre eficiencia de reconocimiento frente a la velocidad de entrenamiento, además de la capacidad de reconocer patrones de mayor tamaño.

PALABRAS CLAVE

Redes Neuronales, Reconocimiento de Voz, Traducción de voz, Lenguaje de Señas para

Sordos, Personas Sordas, Red Neuronal de Contrapropagación.

INTRODUCCIÓN

En la publicación del artículo titulado "Personas sordas exigen trabajo y más interpretes" desarrollado por Roxana Escobar N., por el periódico "El Deber" de fecha 27 de septiembre de 2010, se publican aspectos abordados, conclusiones y las decisiones asumidas por la "FEDERACIÓN BOLIVIANA DE SORDOS – FEBOS" en su semana mundial que se llevó a cabo en esas fechas, de donde se rescatan los siguientes aspectos:

- La FEBOS, considera que la puesta en marcha de la ley contra el racismo y la discriminación, nunca más permitiría que se los llame sordomudos, sino personas sordas, considerando como desafío superar la mayor barrera que significa comunicarse sin perder su identidad, (hacerlo a través del lenguaje de señas), igualdad de condiciones para acceder a la educación escolar, superior y al trabajo.
- En ese evento el presidente de la Federación Boliviana de Sordos, consideró que la población oyente asumía un trato de discriminación sobre la comunidad de personas sordas y convocó a la sensibilización. "Nos llaman sordomudos, pero nosotros tenemos identidad y cultura propia; nos comunicamos a través de la vista y de señas; nuestros ojos son nuestros oídos", sostuvo.
- Se reconoció también que a esa fecha y hasta hoy no hay cifras actuales de cuántas personas sordas existen a escala nacional, el único dato que se cuenta es del censo del 2001 que sobre la base de solo 3 discapacidades arrojó los siguientes datos: 61.145 hogares declararon tener una o más personas con discapacidad de este Total 26.016 pertenecen al área urbana y 35.129 al el área rural.
- Se estableció la necesidad de contar con un número mayor de intérpretes profesionales y reconocidos, ya que la mayor barrera es que no pueden comunicarse con la población oyente,

ttampoco informarse, porque en la mayoría de los canales de televisión no emiten las noticias en lenguaje de señas y el celular lo usan sólo para enviar mensajes de textos. La limitante de la comunicación es un obstáculo a la hora de acceder a un empleo, pues solo pueden acceder a trabajos inferiores y no jerárquicos.

- Se informó de la apertura de la Oficina Central de Investigación de la Lengua de Señas en Santa Cruz, con objetivos centrados en la educación de los niños sordos del país.
- Se reconoció la importancia del decreto presidencial que se reconoce la lengua de señas boliviana y que ésta se suma a las 36 lenguas que hay en Bolivia.
- En ese evento se hizo conocer la dotación de 40 millones de bolivianos que recibirá este sector de parte del Gobierno, para invertir en programas de capacitación de las personas sordas.

De acuerdo con la información obtenida de los resultados del encuentro nacional de las personas sordas del país y la información desplegada en el sitio oficial de FEBOS (<http://www.discapacidadbolivia.org/index.php/extensions56/umadis/organizaciones-civiles14/128-federacion-boliviana-de-sordos-febos>), entre las necesidades más apremiantes se consideraron que las personas sordas puedan acceder a la educación escolar y superior con igualdad de condiciones que la población oyente, siendo pocas las personas sordas que han logrado profesionalizarse, gracias al apoyo de la familia que ha buscado ayuda en forma temprana; la mayoría sólo alcanza la primaria. La inclusión social y la rehabilitación de estas personas es importante para vivir en una sociedad sin discriminación, en la que todas y todos tengan las mismas condiciones de acceso a las diferentes posibilidades de formación y trabajo.

Actualmente, existen institutos especializados para personas con discapacidades auditivas y de comunicación verbal; sin embargo, sus capacidades se ven superadas por sus limitadas infraestructuras y el número reducido de personal que disponen en relación al número de personas que acuden a estos, situación que definitivamente condena, a una gran parte de la población sorda boliviana, a que se vea limitada en su desarrollo personal, humano, educativo y profesional.

Las limitaciones y situaciones adversas que afectan a este grupo de personas, que en su calidad de seres humanos con todo el derecho que la sociedad les brinda, medios y alternativas para superar sus limitaciones, y habiendo revisado que las diferentes instancias no han trabajado en algo

específico, sino solo se ha buscado soluciones en el los mecanismos tradicionales que de ninguna manera ha mejorado las condiciones oficialmente reconocidas en el 2010 a la fecha, ha motivado a que se conforme un grupo de investigación que considerando las posibilidades que nos brindan actualmente las TICS, busque una solución a esta problemática a partir de la Inteligencia Artificial en el área de las Redes Neuronales, considerando la creación de un modelo neuronal que permita en tiempo real interpretar los sonidos que emiten las personas en relación a letras, palabras y expresiones y transformarlas de manera inmediata en representaciones gráficas del lenguaje de señas boliviano.

El interés en el ámbito científico es fundamental por cuanto la investigación considera varios modelos neuronales además de un conjunto de variables que rigen reconocimiento auditivo en tiempo real y sin distinción de una voz específica y modelos matemático – algorítmicos que generan su respectiva interpretación en el lenguaje de señas boliviano, tomando en cuenta las distintas peculiaridades que este tiene.

Los estudios sobre reconocimiento del habla se iniciaron en 1950-1959 con estudios que buscaban explorar las ideas de la fonética – acústica pero sin resultados satisfactorios. A partir de 1960, se inicia una verdadera explosión en las investigaciones de reconocimiento del habla, con importantes aportes de los japoneses. Sin duda la técnica más expresiva desarrollada en esta época fue el Análisis del Cruzamiento por Cero (Zero-Crossing Analysis) que conseguía distinguir entre varias regiones del estímulo auditivo facilitando, en gran medida, el reconocimiento. Otra técnica, no menos importante, desarrollada en la década del 60, es la Programación Dinámica para alineamiento temporal, desarrollada por el científico ruso Vintsync, a pesar que la base del concepto fue considerada rudimentaria, las versiones más avanzadas de su algoritmo son muy conocidas en occidente que utilizó ampliamente este algoritmo a inicios de la década de los 80. En los laboratorios Bell se inició una serie de experimentos con el objetivo de crear un sistema de reconocimiento del habla con independencia de locución. [Rabiner et ai, 1979].

En 1980 se aplican las Redes Neuronales Artificiales (RNA) [Morgan, 1995], [Jelinek, 1976] para el reconocimiento de patrones del habla. Las RNA, inicialmente, fueron introducidas durante la década del 50; sin embargo, sus resultados no fueron considerados satisfactorios debido a un conjunto de problemas operacionales que fueron

subsanados a medida que el reconocimiento avanzó. La década del 80 es, reconocidamente, la década donde el avance en la pesquisa en reconocimiento del habla presenta sus mejores resultados. Como ejemplo podemos mencionar el sistema desarrollado por la Agencia de Proyectos de Investigaciones Avanzadas para Defensa (DARPA-USA) que reconoce con precisión mil palabras en comunicación continua. Se trata de palabras claves utilizadas para el control y comando de misiles. El programa continua en la década del 90 enfatizando técnicas de Inteligencia Artificial, como es el caso del área de estudio en procesamiento del lenguaje natural [Bahl et al, 1989].

Existen varios trabajos a nivel internacional en relación al reconocimiento de voz y su respectiva transformación a un lenguaje de señas, que asumen diferentes estrategias y modelos neuronales, con sus respectivas limitaciones y peculiaridades en relación al contexto y cultura al que pertenecen y los elementos tecnológicos empleados. Cabe resaltar que los sistemas que ya existen son capaz de reconocer solo ciertas palabras, mientras que el presente proyecto pretende llegar a reconocer cualquier palabra, pues la palabra no será nuestra unidad de sonido sino más bien el fonema.

En Bolivia no se conoce de trabajos que hayan abordado esta problemática, por cuanto en la actualidad se sigue empleando a seres humanos como traductores y comunicadores entre personas sordas y personas oyentes ya sea de manera unidireccional o bidireccional.

En relación a lo desarrollado se establece la formalización del problema científico como:

"La escasa existencia de traductores humanos de lenguaje castellano a lenguaje de señas boliviano o viceversa, constituye la principal causa de las limitaciones y estancamiento en el desarrollo profesional y humano de las personas sordas"

El presente trabajo asume la formulación de la siguiente hipótesis:

"Un Sistema Inteligente de interpretación de fonemas correspondientes a letras, sílabas y palabras en imágenes representativas del lenguaje de señas boliviano, permitirá lograr comunicar información de manera efectiva a personas sordas"

Se asume como objetivo del presente estudio:

"Diseñar un sistema inteligente de interpretación de fonemas correspondientes a letras, sílabas y palabras en imágenes representativas del lenguaje de señas boliviano".

CUERPO DEL TRABAJO

Se establece la presente investigación dentro de los niveles que corresponden al tipo de investigación tecnológica que responde a una instancia de "Aplicación", considerando la necesidad de involucrarse activamente en la atención de sectores sociales olvidados, la importancia del ser humano más allá de las limitaciones que este pueda tener, asumiendo medidas preventivas y correctivas oportunas, que permitan proyectar un futuro prometedor para la zona en estudio, buscando lograr niveles elevados de personas sordas que puedan acceder a información sin la necesidad de un traductor humano, integrándose a la sociedad y reavivando sus posibilidades de desarrollo y progreso personal.

Se construyeron artefactos y procesos considerando el uso de herramientas e instrumentos tecnológicos, lingüísticos, matemáticos, lógicos y fundamentalmente asumiendo como medios involucrados la energía y la información.

Se asumió el diseño como un elemento importante y obligatorio en la construcción de artefactos y procesos, usando el saber formalizado de los diversos campos del conocimiento necesarios para la correcta modelación del objeto de estudio, campo de acción, la problemática y la respectiva propuesta de solución.

Por la esencia experimental del presente trabajo se asumió el método del ensayo y error como aspecto fundamental para la validación de los diferentes resultados.

El trabajo desarrollado corresponde a una Investigación Propositiva, por cuanto se ha desarrollado una actuación crítica y creativa, caracterizada por plantear una opción y alternativa al problema formulado.

También se aplicó el método de la modelación para el diseño del software, el cual está implementado en C++ bajo el paradigma de la programación orientada a objetos.

En relación a las técnicas empleadas se consideraron las siguientes:

- Entrevistas, que permitieron conocer información peculiar y detallada de circunstancias y aspectos

difíciles de consultar a través de una encuesta.

- Encuestas, orientadas a identificar información y datos que responden a tendencias representativas de grupos de personas involucradas con el objeto de estudio a diferentes niveles.

Los instrumentos empleados son:

- Guía de entrevista, con la finalidad de organizar y planificar adecuadamente la secuencia y objetivo de las preguntas que se realizaron en las respectivas entrevistas con personas involucradas con el objeto de estudio y se constituyeron en un elemento importante en la obtención de información y datos peculiares y específicos.
- Cuestionario, documento con preguntas organizadas estratégicamente que permitieron obtener información y datos en relación a la opinión o conocimiento de diferentes aspectos importantes para la presente investigación.
- En cuanto a las herramientas que se emplearon en la presente investigación, se consideraron las siguientes:
- Equipo servidor de alta capacidad.
- Equipos computacionales de escritorio y portátiles.
- Sistemas de audio y sonido (Parlantes, micrófonos, reproductores, etc.).
- Software ofimático y de procesamiento estadístico.
- Lenguaje de alto nivel y un entorno interactivo para el cálculo numérico, visualización y programación (C++).
- Motor de Graficación 2D y 3D.
- Motor de captura de audio.

En principio se experimentó con varios modelos de redes neuronales, posterior al análisis de sus resultados se decidió trabajar con el modelo de red neuronal de contrapropagación, además que esta decisión se respaldó con la revisión de otros trabajos que coincidían en la decisión asumida, pues en todos ellos se constató que el CPN dispone de mejor relación entre eficiencia de reconocimiento y velocidad de entrenamiento, además de la capacidad de reconocer patrones de mayor tamaño. Por ejemplo un perceptrón multicapa, principal alternativa para este tipo de reconocimiento, aplicado a compresión de imágenes, tardó alrededor de 15 min en reconocer ocho patrones con 1024 entradas por patrón, mientras que la red de contrapropagación, utilizada para el reconocimiento de caracteres manuscritos tardó tan solo entre 3 y 5 segundos en reconocer 26 patrones con 15000 entradas para cada patrón.

RESULTADOS

En relación al trabajo de investigación desarrollado es importante mencionar los resultados a partir de los componentes considerados en la construcción de la solución tecnológica, en este sentido tenemos:

CAPTURA

En principio se consideró la utilización de un micrófono común existente en la mayoría de los sistemas computacionales; sin embargo, este sumaba mucho ruido a la señal, posterior a realizar pruebas e indagaciones con otro tipo de micrófonos, se identificó que los micrófonos unidireccionales o direccionales son aquellos micrófonos muy sensibles a una única dirección y relativamente sordos a las restantes. Su principal ventaja es que permite una captación localizada del sonido, siendo estos los aspectos tomados en cuenta para seleccionar este tipo de dispositivo para realizar la captura de la señal de voz, garantizando de esta manera la reducción de ruido.

Posterior a varios estudios se estableció que la calidad mínima considerada para el muestreo debía asumir una frecuencia de 11025 muestras/segundo, pues menor calidad conducía a perder las características de la señal y una elevada calidad imposibilitaba la aplicación efectiva de la red neuronal por la amplitud de los parámetros innecesarios que condicionan el rendimiento y pronta respuesta de la red neuronal, para superar esta dificultad se desarrolló un componente de software que normalizaba las señales con elevada calidad.

Se desarrolló un filtro pasa alto, caracterizado por permitir el paso de las frecuencias más altas y atenuar las frecuencias más bajas, para poder reducir los efectos del ruido que acompaña la señal de voz ingresada, posterior al desarrollo del modelo neuronal se pudo establecer que este no era necesario, descartando su uso.

Un problema importante, es la detección de la presencia del habla en un entorno relativamente ruidoso, denominado en literatura como "Localización de principio y final de la palabra", cuya resolución consiste en desarrollar un mecanismo para la detección de los bordes de la palabra, que consiga distinguir entre voz y silencio. El proceso de determinar el principio y el final de una palabra en un entorno ruidoso es importante en muchos campos del proceso del habla. Por ejemplo, en un sistema de reconocimiento de

palabras aisladas es primordial detectar los límites de la palabra pronunciada para seleccionar la región de traza que se debe tratar posteriormente. Para la detección del inicio del habla se aplica el método de energía a la señal de voz $y[n]$ con longitud L , previamente normalizada. Luego, aplicando un umbral determinado y calculando la energía de la señal (para la presente investigación se hizo el cálculo de la energía cada 735 muestras) y comparando si la energía supera al umbral, realizando ese análisis se logró reducir los ruidos de cada agrupación de muestras. Para este propósito se empleó la siguiente ecuación:

$$E = \sum_{n=r1}^{r2} |y(n)|^2$$

Se tomó un intervalo determinado entre $r1$ y $r2$ donde $r1 < r2$ y se desplaza desde $y[1]$ hasta encontrar un valor que supere un umbral preestablecido. En ese instante se captura la posición inicio donde se supera el umbral y se recorta la señal desde $y[1]$ hasta $y[\text{inicio}]$. La señal en ese instante inicia en $y[\text{inicio}]$ hasta $y[L]$. Finalmente se normalizan los datos en el rango de 0 a 1, para su posterior tratamiento en la red.

1. RED NEURONAL DE CONTRAPROPAGACIÓN (CPN)

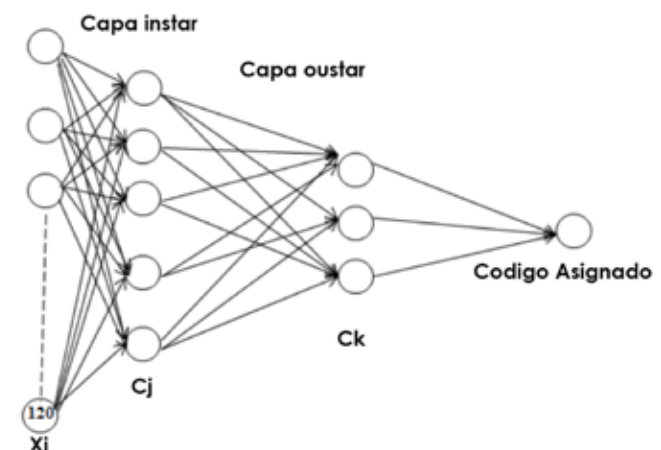
Se utilizó la red de Contrapropagación, siendo esta la que mejor se adecuaba a los propósitos del trabajo de investigación, debido a su rapidez de entrenamiento. El prototipo realiza el reconocimiento de voz en tiempo real, todos tenemos diferentes tonos de voz, necesita ser entrenado para reconocer la voz de un locutor específico. Se determinaron los valores $\alpha=0.7$ y $\beta=0.01$ como los valores óptimos para el entrenamiento de la red, pues como se sabe α puede tomar valores altos, mientras que β debería tomar valores bajos.

Inicialmente, las entradas, una vez filtradas casi sin ruido y llevadas a una escala entre 0 y 1, se normalizan a un tamaño de 120 muestras. Se probaron diferentes tamaños para el patrón de entrada, habiéndose observado que un patrón significativo para un fonema se encontraba en promedio en 60 muestras, por lo que se decidió buscar la muestra con mayor valor del patrón, un pico, a partir del cual se toman las siguientes 120 muestras, 2 veces el tamaño de un patrón, para así tomar una muestra representativa del sonido pronunciado. Se probó con un mayor número de muestras pero no mejoraron sustancialmente los resultados, mientras que un número menor no permitía diferenciar los fonemas.

Entonces, las entradas asumidas son de 120 muestras por patrón de entrada para el reconocimiento de los fonemas vocálicos o sílabas vocálicas, si se toma la sílaba como unidad de sonido tendremos 5 patrones de entrada, que será 1 patrón por cada vocal y así también tendremos 5 clases, una clase para cada patrón de entrada, por otro lado, la capa de outstars cuenta con unidades igual al exponente de 2, necesario para representar las 5 clases de nuestra salida, código binario que será asignado internamente al cual después se le dará un valor decimal, ejemplo: vocal "a" código será 0, vocal "b" código será 1 y así sucesivamente. (vale la pena aclarar que la red se modifica de manera dinámica, es decir, se construye en función a la cantidad de patrones que le mandemos y se asigna la correspondencia con los códigos de acuerdo a este orden).

A continuación se presenta la arquitectura de la red de Contrapropagación asumida para el reconocimiento de voz de las 5 vocales:

Figura 1. XXXXXX



Donde x_i será los diferentes patrones de entrada provenientes de la señal de voz capturada, una vez realizado todos los tratamientos de señal, las instars C_j representan cada una de las clases para cada patrón de entradas, de las cuales se obtendrá un ganador representativo de cada patrón, las outstars C_k representa el valor del dígito binario y el S_k es el código asignado a la salida binaria.

2. FUNCIONAMIENTO

I. EJECUCIÓN

Se utiliza un micrófono unidireccional para la captura de la señal de sonido, realizando inmediatamente una corrección de la señal, posteriormente se aplican los filtros y se eliminan los

silencios y se separa por sílabas el mensaje (Buffer de bits que representan cada sílaba). Seguidamente, la red neuronal hace el reconocimiento del sonido, devolviendo el código asignado al mismo, para finalmente realizar una traducción de la voz al Lenguaje de Señas Boliviano, a través de una interfaz que inicia el proceso con el botón "Iniciar Sonido" para capturar la voz, lo cual es visualizado en tiempo real y se detiene el proceso oprimiendo el botón "Detener Sonido".

II. ENTRENAMIENTO

Se seguirá el mismo procedimiento que para la ejecución, una vez que haya grabado los patrones de entrenamiento a través del micrófono apretando "Iniciar Sonido" y "Detener Sonido".

En este punto se almacenan todos los patrones de entrenamiento en una matriz que serán aprendidos por la red neuronal con el botón "Iniciar Entrenamiento". La primera vez que se inicia la aplicación esta en modo entrenamiento.

CONCLUSIONES

Fruto de las experiencias y resultados obtenidos en el presente trabajo de investigación se establecieron las siguientes conclusiones:

- El modelo de red neuronal CPN, es muy útil para el reconocimiento de patrones, incluso en tiempo real por las características explicadas, pero especialmente por la velocidad de entrenamiento de patrones grandes.
- Los patrones de la señal de voz son muy complejos, pues son muy variables, lo que dificulta la elección de características para su reconocimiento.
- Se hicieron pruebas con las voces de varios sujetos con diferentes características en cuanto a su entonación, la velocidad y modismos idiomáticos, entre otras cosas, se observó que el prototipo no podía reconocer otra voz sino con la que fue entrenado, es decir el reconocimiento de voz es dependiente del locutor.
- Se desarrolló un prototipo rápido con el que se logró el objetivo de reconocer las vocales, lo cual es el gran primer paso para un reconocimiento de voz en tiempo real. Creemos que es cuestión de tiempo para poder llegar a reconocer una oración o una frase.

- El reconocimiento de voz en tiempo real, requiere un estudio profundo y a detalle de la señal de voz y sus características a un nivel básico. Cuando se produce cualquier sonido, nosotros podemos oírlo porque las partículas que se encuentran entre nosotros y la fuente de ese sonido se mueven, nos referimos a que generan ondas, que van de un lado para otro. La cantidad de veces que esas partículas se mueven de un lado para otro se denomina frecuencia y seguro que todos hemos oído hablar de la frecuencia de un sonido, la espectrografía, en este caso, se dedica a medir la frecuencia de los sonidos en un determinado espacio de tiempo. Cada sonido tiene una frecuencia diferente en cada momento y eso nos permite diferenciar, sobre un espectrograma, qué sonidos están sonando.
- Se logró demostrar que las redes neuronales pueden ser aplicadas para el reconocimiento de voz en tiempo real. Desde el punto de vista asumido como investigadores consideramos que la red de Contrapropagación es la más indicada para este trabajo, por su velocidad, forma de clasificación competitiva y su arquitectura que combina una red competitiva con la estructura outstars de Grossberg entre otras cosas.

RECONOCIMIENTOS y AGRADECIMIENTOS

Para la ejecución del presente trabajo, es importante agradecer:

- A la Carrera de Ingeniería de Sistemas y al Instituto de Investigación en TICS, que a través de sus autoridades apoyaron decididamente esta iniciativa.
- A todos los docentes que apoyaron en el proceso investigativo, fortaleciendo el conocimiento en diferentes áreas y estando siempre prestos a responder ante las dudas presentadas en el desarrollo del presente trabajo.
- A los compañeros que apoyaron desinteresadamente en la elaboración de pruebas, toma de muestras y otras actividades importantes y que requerían de la participación de un importante grupo de personas.
- A nuestros padres por apoyo incondicional y siempre estar con nosotros para afrontar cualquier reto que nos propongamos.