Artículo **Revista Ciencia, Tecnología e Innovación**
Junio 2016 Volumen 12, Número 13 755-762

Redes neuronales convolucionales aplicadas a la traducción del lenguaje verbal español al lenguaje de señas boliviano

Convolutionary Neuronal Networks Applied to the Translation of the Verbal Spanish Language to the Bolivian Sign Language

CAMACHO - Francisco*[1] & LABRANDERO – Juan[2]

[1, 2] *Universidad Mayor, Real y Pontificia de San Francisco Xavier de Chuquisaca, Facultad de Tecnología, Carrera de Ingenieria de Sistemas y Telecomunicaciones, Calle Regimiento Campos 180, Teléfono 591-4-6455328, Sucre – Bolivia.*

Recibido Marzo 06, 2016; Aceptado Mayo 06, 2016

Resumen

Se ha utilizado redes neuronales convolucionales para interpretar sonidos emitidos por personas, para posteriormente ser traducidos al lenguaje de señas boliviano, se recurrió a la transformada de Fourier y las escalas de Mel para la creación de patrones de entrenamiento, con diferentes tamaños considerando palabras sueltas y frases del español y una red neuronal convolucional para el reconocimiento. El entrenamiento de la red neuronal considero todos los tamaños de patrones con la finalidad de mejorar el filtrado de la voz capturada antes de aplicar el proceso de reconocimiento. La interpretación de la traducción utilizó el lenguaje dactilológico y el lenguaje de señas boliviano y su representación visual se la realizó a través de animaciones en tercera dimensión. La efectividad de la traducción fue validada a través de un experimento con la participación voluntaria y autorizada de internos del instituto audiológico en la ciudad de Sucre.

Palabras Clave

Redes Neuronales Convolucionales, Reconocimiento de Voz. Traducción de voz, Lenguaje de Señas Boliviano.

Abstract

It has been used convolutional neural networks to interpret sounds emitted by people, later translated into Bolivian sign language, the Fourier transform and the Mel scales were used to create training patterns, with different sizes considering single words and Spanish phrases and a convolutional neural network for recognition. The training of the neural network considered all the sizes of patterns in order to improve the filtering of the captured voice before applying the recognition process. The interpretation of the translation used the sign language and the Bolivian sign language and its visual representation was realized through animations in third dimension. The effectiveness of the translation was validated through an experiment with the voluntary and authorized participation of inmates of the audiological institute in the city of Sucre.

Keywords

Convolutional neural networks, Voice Recognition. Voice translation, Bolivian Signal language.

Citación: Camacho F & Labrandero J. Redes neuronales convolucionales aplicadas a la traducción del lenguaje verbal español al lenguaje de señas boliviano. Revista Ciencia, Tecnología e Innovación 2016, 12-13: 755-762

ISSN-Impreso: 2225-8787

Corresponde al Autor (Correo electrónico: javicamdo@gmail.com)* Investigador primer autor.

Introducción

Los sistemas de cómputo actualmente no pueden interpretar los sonidos generados por una persona y transformarlos en sus correspondientes ideas o expresiones en un determinado lenguaje.

Las redes neuronales profundas han provocado avances notables en la eficiencia de los motores de reconocimiento de voz. Las redes neuronales convolucionales son introducidas con mayor frecuencia en varias etapas de las estrategias de clasificación de la señal de la voz humana o predicción de pertenencia de una muestra de voz a un determinado hablante, generando importantes resultados en estudios relacionados con la interpretación de la voz y su respectiva traducción a otro lenguaje. La imposibilidad de comunicación de personas sordas con otras que no poseen esta limitación y desconocen el lenguaje de señas boliviano ha sido foco de atención y preocupación por varias entidades, asociaciones, instituciones, organismos y fundaciones a nivel mundial, nacional y local por el marginamiento y desigualdad que ocasiona esta limitación en las aspiraciones de integración e introducción a la sociedad productiva de estas personas.

Los japoneses presentaron la idea de Codificación Predictiva Lineal (LPC-Linear Predictive Coding) como modelo de reconocimiento del habla, que es utilizado actualmente de manera efectiva en la codificación y compresión de la voz, a través del uso de medidas de distancias sobre el conjunto de parámetros LPC (B. S., 1974). La Agencia de Proyectos de Investigaciones Avanzadas para Defensa (DARPA-USA) desarrolla un proyecto de procesamiento del lenguaje natural aplicando técnicas de Inteligencia Artificial, logrando reconocer con precisión 1000 palabras en comunicación continua para el control y comando de misiles. (Yuqing Gao, s.f.).

El uso de GMM (Gaussian Mixture Model) y técnica híbrida GMM-HMM, tecnología basada en modelos generativos de lenguaje hablado entrenados discriminativamente logra importantes avances en el ámbito del reconocimiento del habla (Poonam Bansal, 2008).

Posterior a la introducción de las Redes Neuronales Profundas por Geoffrey Hinton y sus estudiantes de la Universidad de Toronto, Li Deng y sus colegas de Microsoft Research (Geoffrey Hinton, 2012), logran definir los basamentos de las redes neuronales profundas para el reconocimiento del habla. Modelos que hasta la fecha son los más efectivos. Los algoritmos más difundidos y utilizados en el ámbito del reconocimiento de la voz, son: Dynamic Time Warping, Los modelos ocultos de Markov, Algoritmo de Viterbi y las Redes Neuronales Convolucionales. Existen varios trabajos en el área del reconocimiento de voz, sin embargo, no se pudo identificar alguno que considere específicamente la traducción del lenguaje verbal español boliviano al lenguaje de señas boliviano.

El presente estudio presenta un conjunto de criterios, técnicas y algoritmos aplicados a redes neuronales convolucionales, para el reconocimiento de sonidos correspondientes a la voz y su representación en el lenguaje de señas boliviano. Las redes neuronales convolucionales (CNN) utilizan una arquitectura especial que está particularmente adaptada para clasificar imágenes, organiza sus neuronas en tres capas o dimensiones (anchura, altura, profundidad); cada capa transforma el volumen de entrada 3D a un volumen de salida 3D de activaciones neuronales. Se utilizaron coeficientes cepstrales para la representación del sonido de manera gráfica, lo cual permitió utilizar este tipo de redes neuronales para el reconocimiento de la voz, considerando varias redes neuronales dependientes del tamaño del patrón.

La máxima tasa de acierto de clasificación responde a 32 mapas de características en una primera capa de convolución y cada capa sucesiva utiliza el doble número de mapas de características y es doblemente pequeño que el predecesor, mientras más profundo está la capa en la que se encuentra, esto debido a las capas de pooling entre cada capa de convolución.

Material y métodos

En la realización de los experimentos se utilizó una estación de trabajo con la siguiente configuración.

El Hardware usado fue:
- Un micrófono unidireccional estándar. Para la captura de sonido.

- CPU tecnología Intel Core i7 de 2 núcleos con frecuencia básica de 2,40GHz. GPU tecnología AMD GNC de 384 procesadores de flujo con frecuencia básico de 300 MHz. Para el entrenamiento y clasificación

El software usado fue:

- Lenguajes de programación C++ (Estándar) y OpenCL. Para la implementación de los algoritmos.

- DeepCL. Para el desarrollo de la red neuronal de convolución.

- Blender, OpenGL. Para la animación del intérprete virtual

Diseño experimental

La tarea de la experimentación práctica se centró de manera inicial en la elección de la arquitectura de red neuronal a utilizar, sus características y configuración en función a las herramientas de hardware disponibles y al software elegido.

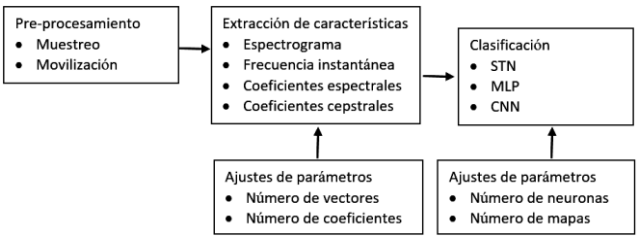

Figura 1. Esquema del Diseño Experimental

Captura y extracción de características Resultados

En el proceso de captura de sonido se ha usado un valor para el muestreo de 22050 Hz de un solo canal con una resolución por muestra de 16 bits. Para hacer el cálculo de la transformada de Fourier se ha usado la ventana de *Hamming* con una longitud de 1024 muestras, esto significa que después de aplicar la transformada nos queda un vector de 512 valores que representan el espectro de frecuencias de la señal en un instante. Para calcular los coeficientes espectrales se ha usado un banco con 16 filtros triangulares en la escala de Mel abarcando un espectro desde 300 Hz hasta 18 kHz. Para Obtener los coeficientes cepstrales (MFCC) se ha aplicado la Transformada discreta del coseno y finalmente se calcularon 8 coeficientes de velocidad (deltas) y 8 coeficientes de aceleración (delta-deltas).

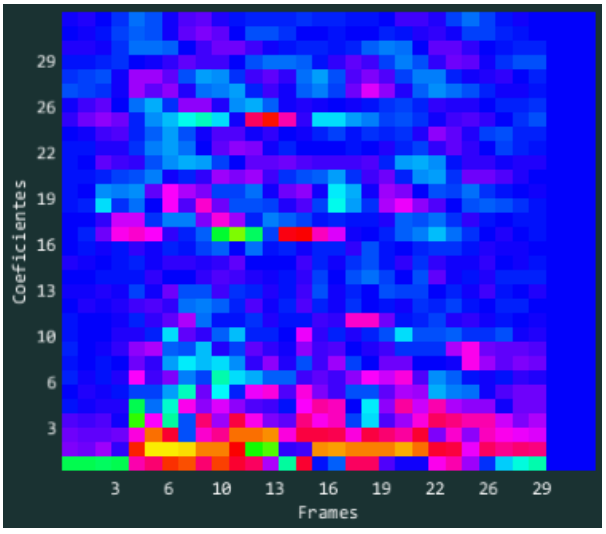

Figura 2. Matriz de 32 x 32 (16 MFCCs, 8 deltas y 8 deltas-deltas).

El vocabulario quedó compuesto de 10 palabras sueltas y 10 frases compuestas pronunciados en lengua española por dos vocalistas de sexo masculino con registros vocales falsetto y modal, con una cantidad mínima de 10 muestras de cada una para el entrenamiento y 10 muestras para la validación. Con lo que la base de datos de vocalizaciones quedó con 400 muestras por vocalista en total.

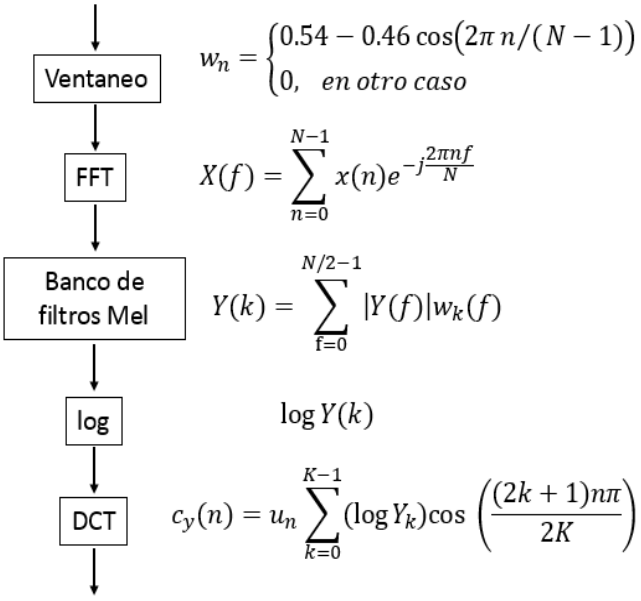

Figura 3. Proceso de Extracción de características.

Clasificación

En la etapa de clasificación se toman las características extraídas de la señal de voz y se convierten en vectores de características puntuales. En éste procedimiento se tienen matrices de $N \times P$ donde N es el número de instantes de tiempo y P es el número de características o variables dinámicas, que se organizan para obtener un supervector de tamaño $N \times P \times 1$ por cada muestra. Luego si la cantidad total de muestras del conjunto de entrenamiento es M, entonces se construye la matriz de tamaño $N \times P \times M$.

La arquitectura de red neuronal de convolución usada fué LeNet-5 (Y. LeCun, 1998), con una matriz de entrada de 32x32 características MFCC, 3 capas convolucionales de 16, 32 y 64 mapas de salida respectivamente y con filtros de 5x5 cada una, 2 capas de pooling con factor de 1/2 entre cada una de las anteriores capas y finalmente una capa de neuronas totalmente conectadas. En el entrenamiento de la red neuronal se ha usado la técnica del Descenso del Gradiente Estocástico Asíncrono (Stochastic Gradient Descent - SGD) con parámetros de momentum de 0.02, y factor de aprendizaje fijo de 0.01, reduciendo el factor de aprendizaje un 2% cada 10 épocas.

Intérprete virtual de la lengua de señas boliviana

La secuencia de palabras en cada expresión o frase y la pronunciación de cada palabra están relacionados con una o más señas que siguen las normas y la estructura de la lengua de señas boliviana. La base de datos de señas está compuesta por 30 señas que corresponden a cada una de las letras en el alfabeto dactilológico y 20 señas específicas relacionadas con una palabra o frase. Para la interpretación de la lengua de señas se ha utilizado animaciones 3D aplicando la técnica de animación de esqueletos.

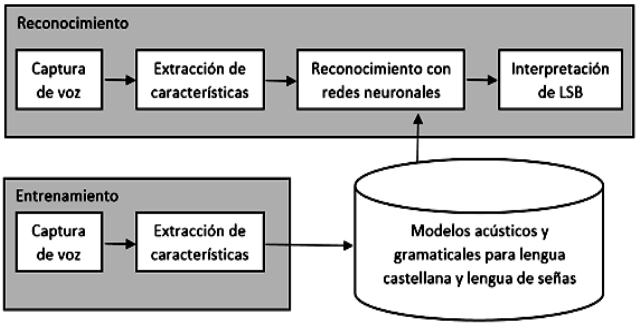

Figura 4. Proceso de Extracción de características.

Resultados

Empleando métodos y técnicas de contrastación de la efectividad del rendimiento de las redes neuronales convolucionales desarrolladas para el proceso de aprendizaje y traducción del lenguaje verbal, considerando la auto-organización de los patrones aprendidos, y los resultados obtenidos en la recuperación o etapa de aplicación de la red neuronal convolucional, en el proceso de experimentación se asumieron los siguientes procesos: pre-procesamiento de la señal de voz, cálculo de espectrogramas, cálculo de características, clasificación, ajuste de parámetros, evaluación del rendimiento de cada conjunto de características y evaluación general de rendimiento con las mejores características y pruebas de sensibilidad y especificidad (transmisión de información en LSB).

Obteniéndose los siguientes resultados:

- El Pre-procesamiento de la señal de voz asumió una señal de muestreo a 22050 Hz en un solo canal con una resolución por muestra de 16 bits, multiplicándose posteriormente por una ventana tipo Hamming.

- El cálculo de espectrogramas utilizó el cálculo de la transformada de Fourier considerando la utilización de la ventana de Hamming con una longitud de 1024 muestras, lo que significa que después de aplicar la transformada se obtiene un vector de 512 valores que representa el espectro de frecuencias de la señal en un instante dado.

- La primera fase de extracción de características necesitó adecuar la señal a la metodología, asumiendo 3 grupos para las características y cada grupo con un número diferente de variables, dependiente de los diferentes métodos aplicados para la estimación de las características o de la naturaleza de cada conjunto.

- Se consideraron aproximaciones similares para los coeficientes espectrales y cepstrales, por la función de división del espectro de frecuencias en un número determinado de bandas. Estableciendo que el parámetro principal de las aproximaciones asumidas corresponde con el número de filtros del banco que se aplicó a la Transformada de Fourier para separar cada una de las bandas de interés. Determinando adecuadamente los filtros y su solapamiento en el dominio de la frecuencia. Las características extraídas a partir de los coeficientes espectrales corresponden a la ubicación del coeficiente y la energía concentrada alrededor de dicho coeficiente. Para el cálculo de éstos, se establecieron filtros de Hamming con un solapamiento de 30%, distribuidos linealmente y con aplicación de la Transformada de Fourier, estableciéndose el banco de filtros en 16. El cálculo de los coeficientes cepstrales utilizo filtros distribuidos según la escala de Mel , considerando que el espectro de la señal analizada se encontraba dentro del rango auditivo humano, finalmente se utilizaron filtros triangulares, con un solapamiento de 50% y una cantidad constante de 16 filtros.

- Se generaron los vectores de características estimados y se verificó su efectividad aplicando tres tipos de clasificadores, seleccionando al ganador para la implementación final del prototipo: STN (Redes Neuronales Espacio Temporales), MLP (Perceptrón Multicapa con una capa oculta) y CNN (Redes neuronales convolucionales).

- El clasificador basado en CNN a medida que se varió el número de coeficientes espectrales, se ratificó que el número con el cual se obtiene mejor rendimiento es 16. El número óptimo de coeficientes cepstrales a utilizar es de 32, resultado de experimentos que consideraron 8 coeficientes de velocidad (Deltas) y 8 coeficientes de aceleración (Delta-Deltas). Se verificó que el mejor clasificador con mejor rendimiento es el CNN seguido por MLP y luego STN.

- El clasificador CNN fue el que tuvo el mejor desempeño para la metodología propuesta y las características estudiadas. Los tiempos de entrenamiento de cada clasificador es mayor cuando se usan los coeficientes cepstrales, esto es de esperar ya que éste es un vector mucho mayor en comparación con los coeficientes espectrales, CNN posee los mayores tiempos de entrenamiento, debido a que la arquitectura de su red neuronal posee muchas más capas ocultas (5 en el prototipo implementado). La metodología de entrenamiento de los clasificadores CNN se basó en la técnica del Descenso del Gradiente Estocástico Asíncrono con parámetros de momentum de 0.02, y factor de aprendizaje fijo de 0.01, reduciendo el factor de aprendizaje un 2% cada 10 épocas). Los datos para el entrenamiento se clasifican en 15 categorías en un mapa jerárquico con 30 muestras de cada categoría. Existiendo 150 muestras para el entrenamiento, 150 muestras para la validación y 150 muestras para las pruebas. Cada muestra está asociada de fondo con una categoría, un hablante y el rendimiento se mide en base a la mayor aproximación puntuada en el reconocimiento.

- El sistema devuelve 3 datos: El nivel de precisión al comparar los datos de salida de la red contra los datos de la primera clase reconocida, el nivel de error comparando la salida con las 5 primeras clases reconocidas y finalmente una representación en la LSB asociada a la clase ganadora. La evaluación final sobre el nivel de transmisión de información en LSB se realizó en ambiente controlado con 7 personas sordas con conocimientos de la LSB (I1 – I6) y 1 oyente (I7) también con conocimiento de la LSB, donde el sistema recibió la información verbal capturada mediante micrófono e interpretar la información traduciendo a LSB en tiempo real.

Para la captura de los datos sobre la evaluación se aplicó un cuestionario “Evaluación sobre los niveles de Transmisión de Información” con preguntas organizadas estratégicamente que permitieron obtener información y datos en relación a la opinión sobre el prototipo en funcionamiento y aspectos importantes para la presente investigación. Los datos sobre las pruebas y evaluaciones aplicadas al sistema permiten observar cómo el nivel de comunicación con las personas I1 – I6 asciende de un promedio de 14% a 89% general.

Discusion

Las redes neuronales convolucionales permiten traducir los sonidos emitidos, resultado de una oración, un pensamiento o una expresión verbal auditiva en su correspondiente equivalente expresión física representada gráficamente en tercera dimensión por un software.

Se cumple con el objetivo general establecido demostrando y validando la posibilidad de comunicación de personas sordas con otras que carecen de este impedimento sin la necesidad de conocimiento alguno del lenguaje de señas boliviano o la intervención de un traductor.

La disponibilidad de vocabulario y capacidad de interpretación es directamente proporcional al entrenamiento y ajuste que se logre en la fase de entrenamiento, ratificando que desafíos y avances importantes requieren de equipamiento con alta capacidad de procesamiento.

Se asume un límite de 100 Palabras, 100 Expresiones y 100 Oraciones por las limitantes de los equipos de cómputo utilizados.

El presente trabajo de investigación enmarcado en las propuestas y recomendaciones de otros trabajos anteriores revisados y considerados, ratifica y valida las propuestas y principios establecidos.

Si bien el presente trabajo se limita o está condicionado a la posibilidad tecnológica que se dispuso, las regularidades y principios establecidos ratifican que con la posibilidad de acceder a un HCP (Centro de Procesamiento de alta Capacidad) los resultados revelarán mayores beneficios y ratificarán positivamente la pertinencia y efectividad del presente trabajo de investigación.

Con base en los resultados presentados se pudo comprobar y verificar que una red neuronal convolucional puede traducir expresiones verbales del lenguaje español al lenguaje de señas boliviano. Quedando demostrado mediante las pruebas realizadas y los resultados empíricos obtenidos por la experimentación, evaluación y valoraciones a los que fue sometida la aplicación.

Se establece el aporte práctico de la presente investigación en relación a la propuesta de alternativas tecnológicas que contribuyan a fortalecer una de las dimensiones de la competencia comunicativa de las personas sordas con el resto en el trabajo, en reuniones, los medios de comunicación, los espectáculos, medios de transporte, etc. y se pretende que sirva como apertura para la creación y establecimiento de proyectos más avanzados apoyados en la utilización de la Lengua de Señas Boliviana.

El aporte tecnológico de la investigación en asumir la evidente necesidad de fortalecer el componente de la comunicación de las personas sordas con el resto.

El desarrollo y fortalecimiento de las competencias comunicativas de las personas sordas con el resto, debe asumirse como una tarea importante para todos, ya que esto es la base para garantizar el éxito en el proceso de eliminación de toda forma de discriminación y en la búsqueda de canales de participación en la educación, el mercado de trabajo y en la vida social.

Agradecimientos

A la Carrera de Ingeniería de Sistemas y las autoridades facultativas y de carrera por el apoyo humano, técnico y los recursos provistos para este emprendimiento.
Al Ing. Carlos Walter Pacheco Lora Ph.D. por el apoyo y colaboración prestada en el desarrollo de este trabajo.

Referencias

- Alex Graves, A.-r. M. G. H., 2012. *Speech Recognition With Deep Recurrent Neural Networks.* s.l.: Department of Computer Science, University of Toronto.

- B. S., A., 1974. *Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification.* New Jersey: Murray Hill, Bell laboratories.

- Bernhard Schölkopf, J. C. P. T. H., 2007. *Advances in Neural Information Processing Systems 19.* s.l.: MIT Press.

- Chris J., W., s.f. Introduction to Speech Recognition Using Neural Networks. En: *Communications Multimedia.* s.l.: Institut Eurecom.

- Geoffrey Hinton, L. D. D. Y., 2012. *Deep Neural Networks for Acoustic Modeling in Speech Recognition.* s.l.: University of Toronto, Microsoft Research, Google Research, IBM Research.

- Y. LeCun, L. Bottou, Y. Bengio and P. Haffner., 1998: *Gradient-Based Learning Applied to Document Recognition*, Proceedings of the IEEE, 86(11):2278-2324.

- Herv´e Bourlard, N. M., 1994. *Connectionist Speech Recognition, A Hybrid Approach.* s.l.: Kluwer Academic Publisher.

- James A. Freeman, D. M. S., 1991. *Neural Networks: Algorithms, Applications, and Programming Techniques.* Computation and Neural Systems Series Computation and neural systems series ed. s.l.: Addison-Wesley.

- Ossama Abdel-Hamidy, A.-r. M. H. J. G. P., 2012. *Applying Convolutional Neural Networks Concepts to Hybrid NN-HMM Model for Speech Recognition.* s.l.: Department of Computer Science and Engineering, York University, Toronto, Canada.

- Paul Lamere, P. K., s.f. *The CMU SPHINX-4 Speech Recognition System.* s.l.: Sun Microsystems Laboratories, Carnegie Mellon University, Mitsubishi Electric Research Labs.

- Poonam Bansal, A. K., 2008. Improved Hybrid Model of HMM/GMM for Speech Recognition. *Intelligent Information and Engineering Systems INFOS.*

- Shrikanth Narayanan, K. N., s.f. *USC-TIMIT: A database of multimodal speech production data.* California: Signal Analysis and Interpretation Laboratory

- Wu Chou, B.-H. J., 2003. *Pattern Recognition in Speech and Languaje Processing.* Electrical Engineering & Applied Signal Processing Series ed. s.l.: CRC Press.

- Yuqing Gao, H. E. Y. L., s.f. *Recent Advances in Speech Recognition System for IBM DARPA Communicator.* s.l.: IBM Thomas J. Watson Research Center.