S17 CDT Ciencia, Docencia y Tecnología

Suplemento CDyT
VOL. 14 | N° 17 | 2024 ISSN 2250-4559

PID 6187

Implementación en FPGA de una red neuronal pulsante para clasificación de habla y estados emotivos

Iván Peralta, Eduardo Filomena, Juan I. Rufiner, Nahuel E. Ricart, Hugo L. Rufiner, Carlos M. Pais, Nanci Odetti, Iván Gareis, Marcos Formica

Autores: Facultad de Ingeniería, Universidad Nacional de Entre Ríos. Ruta provincial 11, km 10 Oro Verde, Entre Ríos, Argentina
Contacto: ivan.peralta@uner.edu.ar

ARK: https://id.caicyt.gov.ar/ark:/s22504559/0j4wrcn7j

1. Resumen

Las redes neuronales pulsantes (SNN) se destacan por su habilidad para clasificar patrones temporales, gracias a su capacidad de integrar información en el tiempo. Una de las principales dificultades de este tipo de redes es el elevado costo computacional necesario para su simulación en tiempo real. En la última década, se ha intentado aprovechar dispositivos electrónicos programables tipo FPGA para implementar estas redes, ya que comparten con las SNN el tratamiento en paralelo de sus unidades funcionales. Asimismo, el uso de estos dispositivos allana el camino para la eventual creación de sistemas portátiles en el futuro. Es de gran interés lograr un sistema capaz de realizar tareas de clasificación en tiempo real sobre señales de habla. En este proyecto, se abordó la investigación centrada en el diseño e implementación de un sistema de adquisición, procesamiento y clasificación de habla mediante el uso de una red pulsante, junto con su correspondiente algoritmo de entrenamiento. El objetivo principal consistió en lograr la clasificación del habla y los estados emotivos. Para esto se desarrolló el modelo de SNN denominado DELSNN con resultados alentadores para clasificación de habla. La clasificación de estados emotivos no cumplió las expectativas, destacando áreas de ajustes y mejoras futuras.

Palabras clave: Reconocimiento de habla, reconocimiento de emociones, SNN, FPGA

Objetivos propuestos y cumplidos

Durante el transcurso de este proyecto de investigación, se alcanzaron con éxito una serie de objetivos clave que fortalecieron significativamente la labor investigativa en la FI-UNER en el ámbito de la clasificación de patrones biológicos y la aplicación de tecnología electrónica de última generación. Estos logros se desglosan de la siguiente manera:

1. Implementación en FPGA de una red pulsante y desarrollo de un algoritmo de entrenamiento: se realizaron etapas fundamentales, incluyendo la obtención, instalación y el estudio de las herramientas necesarias. Esto proporcionó una base sólida para el desarrollo del proyecto. El objetivo se cumplió completamente.
2. Clasificación exitosa de palabras aisladas: se logró un avance significativo al alcanzar con éxito la clasificación de palabras aisladas, lo que representó un hito importante en la implementación de la red pulsante y el algoritmo de entrenamiento.
3. Clasificación exitosa de frases continuas y acotadas: posteriormente, se continuó con éxito la clasificación de frases continuas y acotadas, consolidando aún más el progreso del proyecto.
4. Investigación en la clasificación de emociones: Se dedicó un año a la investigación en la clasificación de emociones, que incluyó una búsqueda bibliográfica exhaustiva y el estudio de diversas representaciones de características de habla para esta tarea. Se obtuvieron y prepararon meticulosamente los datos, y se probaron diferentes estrategias de codificación.
5. Implementación exitosa de la red en la FPGA: se logró la implementación de la red en la FPGA, lo que representa un logro significativo en términos de tecnología electrónica avanzada.
6. Publicación de dos artículos exitosos sobre clasificación del habla: A pesar de no alcanzar los resultados esperados en la clasificación de emociones, se lograron publicar dos artículos destacados en el campo de la bioingeniería, contribuyendo significativamente al conocimiento en esta área.

En resumen, a lo largo del proyecto, se cumplieron la mayoría de los objetivos planteados, lo que consolidó de manera sólida la investigación y el desarrollo en el ámbito de la clasificación de patrones biológicos y la utilización de tecnología electrónica avanzada en la FI-UNER.

2. Marco teórico y metodológico

A lo largo de las últimas décadas la comunidad científica ha estudiado la señal del habla en un esfuerzo por imitar las capacidades del ser humano para reconocer e interpretar los diferentes sonidos [1]. Actualmente y gracias al advenimiento de las redes neuronales profundas (Deep Neural Networks, DNN), este objetivo se está logrando [2,3]. En la Figura 1 se muestra la evolución en el tiempo del desempeño de los sistemas de reconocimiento automático del habla (Automatic Speech Recognition, ASR) en los últimos años. En la misma se observa que la tasa de error en el reconocimiento de palabras mejoró de manera constante y que en dos de los "benchmarks" más estudiados (LibriSpeech y Switchboard Hub5'00) las tasas de error ya han superado a las obtenidas

por transcriptores humanos. Si bien pareciera que la solución a las tareas de reconocimiento estaría encaminada, aparecen algunas limitaciones de estos modelos que no permitirían su utilización en cualquier lugar ni por cualquier persona. Una de las principales dificultades de las DNN es las grandes demandas computacionales y de energía que estas requieren principalmente durante su entrenamiento. Los actuales sistemas de reconocimiento de voz son tan complejos que para ejecutar sus versiones más avanzadas es necesario recurrir al uso de GPU, asegurando así un rendimiento en un tiempo razonablemente práctico. No obstante, incluso al utilizar GPU, la transcripción en tiempo real no está garantizada al emplear las versiones más completas de los modelos debido a la gran cantidad de parámetros que estos contienen. A modo de ejemplo, el modelo Whisper de OpenAI, uno de los más exitosos en la actualidad, requiere más de 400 segundos para transcribir 29 segundos de audio con la versión "Medium", ejecutándose en una PC independiente moderna, y 50 segundos con la versión más completa ("Large") cuando se ejecuta en una GPU. Otra opción sería ejecutarlos en servidores específicos, pero esto implica que para entrenar y utilizar dichos modelos es necesario conectarse al servidor a través de internet, lo que compromete la privacidad de los datos, el cual es un requisito que en ciertas áreas de aplicación es indispensable. Esto supone que, en adelante, los esfuerzos de los investigadores estarán orientados en lograr sistemas independientes de una conexión a internet, para lograr que la tarea de reconocimiento suceda en el dispositivo y de esta forma garantizar el requisito de privacidad. Otra consecuencia de utilizar internet para realizar estas tareas es la latencia que ocurre en las conexiones. Lograr reconocedores con latencias imperceptibles hace que la interacción con el dispositivo se sienta mucho más receptiva y, por lo tanto, más atractiva. Otra razón para preferir la inferencia en el dispositivo es la disponibilidad del 100%. Que el reconocedor funcione incluso sin conexión a Internet o en áreas con servicio intermitente significa que funcionará todo el tiempo. Desde el punto de vista de la interacción del usuario, hay una gran diferencia entre un producto que funciona la mayor parte del tiempo y uno que funciona de vez en cuando. Por último, otra de las motivaciones que incentivan el direccionamiento de las investigaciones hacia el reconocimiento en el dispositivo es la personalización. Una de las principales diferencias entre el reconocimiento automático del habla y la interpretación humana del habla radica en el uso del contexto. Los humanos dependen en gran medida del contexto al hablar entre ellos. Este contexto incluye el tema de la conversación, lo que se dijo en el pasado, el ruido de fondo y señales visuales como el movimiento de los labios y las expresiones faciales. Para seguir mejorando la comprensión de la máquina del habla humana, será necesario aprovechar el contexto como parte más profunda del proceso de reconocimiento. Una forma de lograr esto es mediante la personalización. La personalización ya se utiliza para mejorar el reconocimiento de enunciados del tipo "llamar a <NOMBRE>". Sim et al. [4] encontraron que personalizar un modelo con la lista de contactos de un usuario mejora la recuperación de entidades nombradas del 2.4% al 73.5% lo cual es una mejora muy significativa. También se ha demostrado que personalizar modelos para usuarios individuales con trastornos del habla mejora las tasas de error de palabra en un 64% relativo [5]. La personalización puede marcar una gran diferencia en la calidad del reconocimiento, especialmente para grupos o dominios que están subrepresentados en los datos de entrenamiento. Para poder llevar a cabo la personalización en el dispositivo requiere entrenamiento en el dispositivo. Es decir que se requieren de modelos que puedan ser fácilmente adaptados a un usuario

o contexto específico, porque si el modelo necesita aprender a partir de los datos de un usuario, entonces el entrenamiento debería ocurrir en el dispositivo. Se cree que para finales de esta década, los modelos de reconocimiento de voz serán profundamente personalizados [2]. El dilema radica en que los dispositivos de uso generalizado en la actualidad no incluyen de forma predeterminada una GPU, y en caso de contar con ella, su consumo energético sería excesivamente elevado, lo que conllevaría a una incompatibilidad para su implementación en dispositivos portátiles independientes de la red eléctrica.

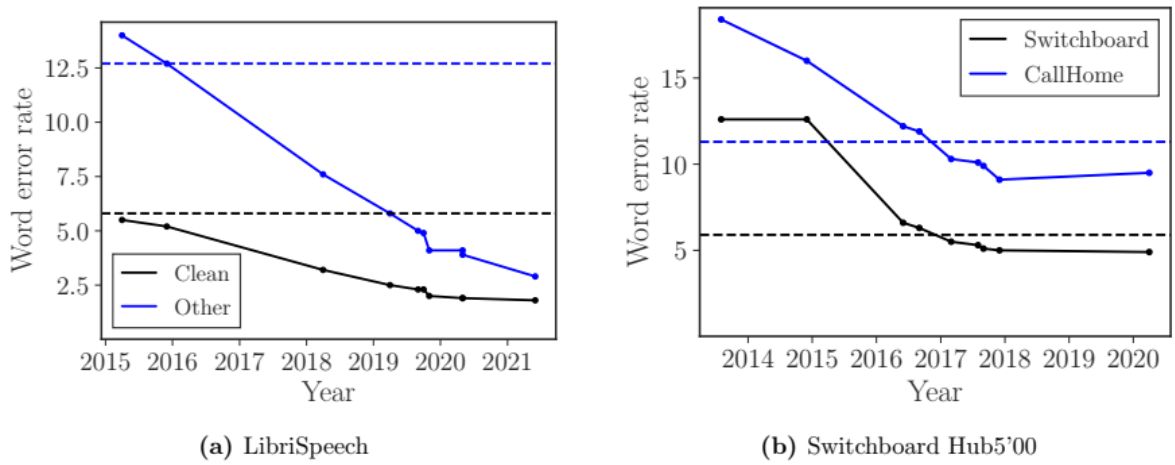


Figura 1: Curvas de evolución del desempeño de los ASR en los últimos años para dos de los benchmarks más estudiados. Las líneas discontinuas indican el rendimiento a nivel humano. Extraído de [2].

A su vez, la señal del habla también está siendo utilizada por las DNN para determinar el estado emotivo del hablante o para producir habla con características emotivas [6]. Las personas expresan los afectos a través de una serie de acciones sobre la expresión facial, movimientos corporales, diversos gestos, comportamiento de la voz y otras señales fisiológicas, como ritmo cardíaco, sudor, y otros. Además ya hace mucho tiempo que es ampliamente conocido que la señal de la voz es fuertemente influenciada por el estado emotivo del hablante [7]. Algunos investigadores han estudiado características influenciables, tales como la variación del tono (rango y nivel de la F0) y rapidez de conversación entre otras [8,9]. Todas estas características ya han sido utilizadas por los investigadores de reconocimiento de emociones mediante métodos de reconocimiento clásicos. En los últimos años, el reconocimiento del estado afectivo de las personas ha ganado considerable atención debido a su utilidad en diversas aplicaciones. Algunos ejemplos notables incluyen la evaluación de la depresión y el riesgo de suicidio mediante el análisis del habla [10], la implementación de sistemas para detectar emociones en situaciones de emergencia médica en tiempo real [11], la predicción automática de la frustración [12], la detección del miedo en escenarios anómalos para aplicaciones de seguridad [13], el apoyo semiautomático en el diagnóstico de enfermedades psiquiátricas [14], la recomendación musical [15] y la identificación de las actitudes emocionales de un niño en interacciones de diálogo con una computadora [16]. Todos estos intentos de trabajar con los aspectos emotivos de las personas y su interacción con los sistemas artificiales se enmarcan dentro de una disciplina de muy reciente aparición conocida como computación afectiva o “Affective Computing” [16].

Hoy en día existen distintas estrategias multimodales que combinan distintas señales biológicas con resultados aceptables. A pesar esto, los métodos para registrar y usar estas señales son invasivos, complejos e imposibles en ciertas aplicaciones reales. Por lo tanto, el uso de señales de voz es claramente la opción más recomendada si está disponible. Por otro lado, dentro del contexto que se mencionó arriba sobre el que los futuros ASR deben incorporar para mejorar su desempeño, están las emociones de los hablantes. Los reconocedores de habla pueden hacer uso de las emociones para lograr una mejor comprensión del contexto de la conversación y mejorar su reconocimiento. A su vez lograr determinar el estado emotivo de la persona no solo puede potenciar el desempeño del reconocedor de habla, sino que también puede lograr que los sistemas de inteligencia artificial que interactúan con el usuario, lo hagan de una manera acorde con el estado emotivo del mismo.

Por ende, es crucial continuar explorando nuevas estrategias con el objetivo de desarrollar reconocedores capaces de garantizar una tasa de reconocimiento aceptable, cumpliendo simultáneamente con los requisitos de energía, privacidad, latencia y personalización. Se presta especial atención a aquellas estrategias inspiradas en sistemas biológicos, dada su eficiencia demostrada en términos de consumo energético [17]. En esta línea, el presente Proyecto de Investigación se dirigió hacia la resolución de las dos problemáticas asociadas con la tarea de reconocimiento previamente mencionadas: a) el reconocimiento de habla en sí mismo y b) el reconocimiento de emociones a través de la señal de habla. La investigación se enfocó en el estudio, diseño e implementación de nuevas estrategias que buscan abordar los requisitos de los futuros reconocedores que empleen la señal de habla mencionados anteriormente. A continuación, se detallan las bases conceptuales fundamentales que sustentaron la investigación.

El Reconocimiento Automático del Habla es un campo multidisciplinario con especial vinculación al reconocimiento de formas y a la inteligencia artificial. Su objetivo es la concepción e implementación de sistemas automáticos capaces de interpretar la señal de voz humana en términos de categorías lingüísticas de un universo dado. En la Figura 2 se observa un esquema conceptual de un sistema de reconocimiento automático del habla. En la actualidad dichos sistemas son implementados en el campo discreto, por lo que el primer bloque siempre resulta un conversor analógico digital, que digitaliza la señal de voz. La estructura general de uno de estos sistemas tiene esencialmente dos componentes o etapas:

1. Análisis del habla: la idea central de este bloque es tratar de representar la señal en otro dominio mediante alguna transformación para hacer más evidentes las características necesarias para la etapa de clasificación. A veces el objetivo es reducir la dimensión de los patrones, pero en otros casos, se intenta proyectar la información en una dimensión superior para separar ciertas características, que de otra forma no se podrían identificar.
2. Reconocimiento o clasificación: esta etapa clasifica o identifica los segmentos de voz ya procesados con símbolos fonéticos o categorías lingüísticas, tales como fonemas, difonos, trifonos, sílabas, palabras u oraciones. Para ello utiliza distintas fuentes de conocimiento.

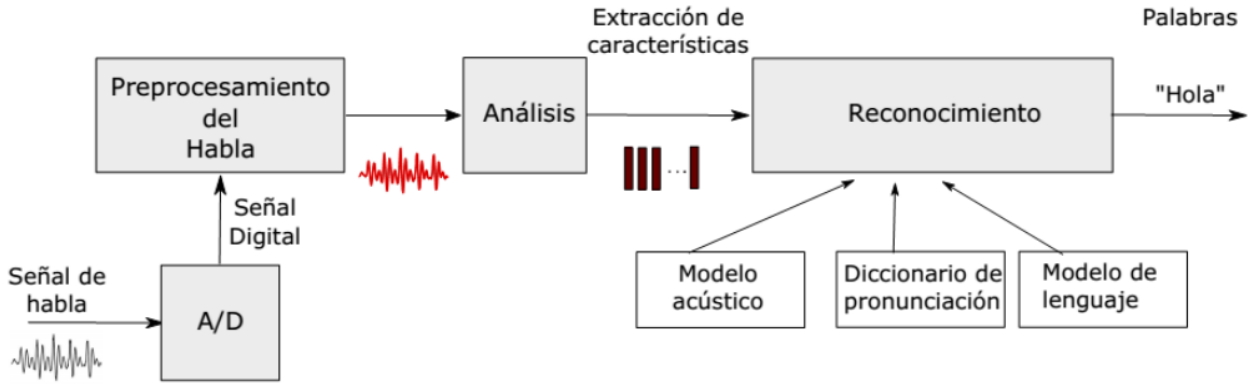


Figura 2: Esquema general de un Sistema de Reconocimiento Automático del Habla.

Los mecanismos o algoritmos utilizados para la tarea de clasificación de habla y emociones son muy diversos, pero como se mencionó antes, actualmente se destacan las DNN, que representan las versiones más recientes de las redes neuronales artificiales (Artificial Neural Network, ANN). Estas redes son los modelos que actualmente logran el mejor rendimiento en la mayoría de las tareas relacionadas con la clasificación. Las DNN también tienen la capacidad de aprender características directamente a partir de muestras de habla en bruto. Pero como se mencionó antes, a pesar de su éxito, las DNN tienen algunas desventajas dentro de las cuales se encuentran principalmente los altos requisitos computacionales y altos consumos de energía, lo que alienta a la comunidad científica a seguir buscando alternativas.

En las últimas dos décadas, aparecieron las *Redes Neuronales Pulsantes* (Spiking Neural Networks, SNN). Éstas son un tipo especial de redes neuronales, las cuales han suscitado un creciente interés por poseer propiedades adicionales a las clásicas redes neuronales artificiales [18]. Entre las propiedades más importantes se destacan la posibilidad de comunicarse entre neuronas mediante “spikes” o pulsos imitando los *potenciales de acción* (PA) de las neuronas biológicas. Dichos potenciales se pueden modelar eficientemente mediante códigos binarios, lo cual facilita la implementación de las SNN en dispositivos digitales. Además, la comunicación a través de pulsos neuronales (spikes) confiere a las SNN una eficiencia energética notable, similar a la del cerebro humano. Esta característica las posiciona como una de las opciones más prometedoras para abordar las limitaciones de las DNN en cuanto a eficiencia energética. Otra de las características relevantes de las SNN es que imitan el potencial de membrana interno siguiendo una dinámica temporal al igual que las neuronas biológicas, lo que les permite integrar información en el tiempo. Esta característica es de suma importancia para tareas que hacen uso de señales temporales como la señal de voz.

Un aspecto crucial a considerar es la capacidad de procesamiento en tiempo real del sistema. Un sistema se clasifica como de tiempo real si puede proporcionar una solución en el mismo intervalo temporal en el que recibe los datos de entrada. La velocidad de procesamiento de un sistema está directamente vinculada a la cantidad de patrones que puede reconocer. Hace tres décadas, ya existían sistemas capaces de reconocer palabras en tiempo real para extensos vocabularios. En 1991, en los primeros esfuerzos por alcanzar este objetivo, se presentó un sistema basado en modelos ocultos de Markov capaz de procesar información en tiempo real para un vocabulario

extenso [19]. Durante varios años, los ASR han logrado procesar información a velocidades hasta 10 veces superiores al tiempo real [20, 21]. Sin embargo, estos sistemas eran considerablemente deficientes en comparación con las capacidades humanas.

Posteriormente, surgieron las DNN, que permiten reconocimientos similares a los humanos, pero como se demostró anteriormente, no pueden satisfacer el requisito de procesamiento en tiempo real con los modelos más avanzados sin el uso de hardware específico, como las GPU. La idea de incorporar nuevas estrategias, como las SNN, para superar estas dificultades de los sistemas actuales, especialmente para lograr reconocedores robustos independientes de la conexión a internet que puedan entrenarse y utilizarse en el mismo dispositivo, plantea el desafío de que los futuros sistemas portátiles requieran una elevada potencia de cómputo. Esto incluso podría resultar en una posible incapacidad para realizar clasificaciones en tiempo real si los modelos son demasiado complejos. En la Figura 3 se presenta el costo computacional de implementación para diversos modelos de SNN de vanguardia. Se observa que a medida que se busca mayor similitud con los sistemas biológicos, los costos computacionales de simulación aumentan notablemente.

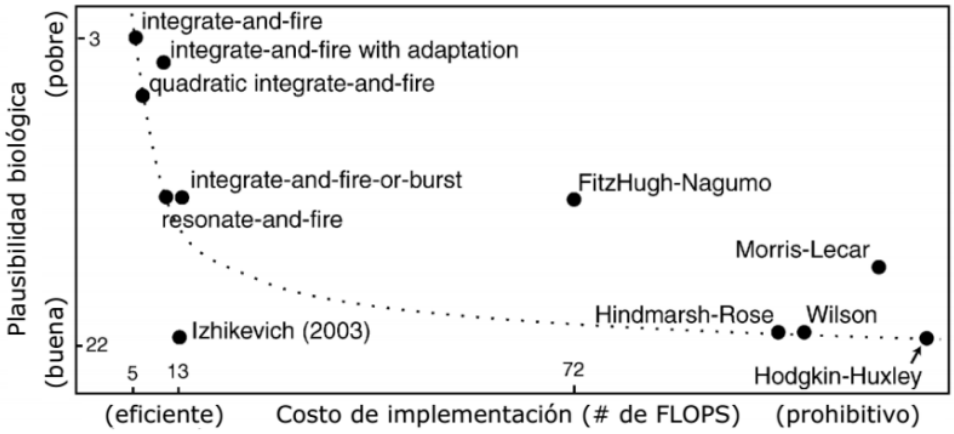


Figura 3: Costo de implementación para diferentes modelos de SNN. El "#of FLOPS" es el número aproximado de operaciones de punto flotante (adición, multiplicación, etc.) necesarios para simular el modelo durante 1 milisegundo. Extraída y adaptada de (Izhikevich, 2004)

Una buena alternativa para lograr satisfacer estos elevados costos computacionales son los dispositivos electrónicos de procesamiento paralelo, los cuales son una de las principales herramientas para alcanzar elevadas potencias de cómputo [22, 23]. Dentro de esta categoría de dispositivos se incluyen las FPGA, que posibilitan la implementación de circuitos digitales reconfigurables de manera notablemente sencilla. Las FPGA se han estado utilizando a lo largo de esta última década en la mayoría de las propuestas que implican un paralelismo inherente en el funcionamiento del sistema, tal como son las ANN o las SNN [24,25]. Los dispositivos FPGA tienen una capacidad lógica de hasta millones de compuertas lógicas, incluyen interfaces programables para varios estándares de interfaz eléctrica y tienen bloques de funciones especiales embebidos entre la lógica programable, tales como memorias, multiplicadores, incluso CPU completas entre otros. Desde un punto de vista comercial y académico, esta gran capacidad y variedad de recursos los hace sumamente útiles a la hora de crear prototipos para el desarrollo rápido de nuevos productos, para los productos que deben ser reconfigurables por naturaleza, o bien productos que se producen en bajos volúmenes

y para los cuales no es económicamente viable crear un circuito integrado a medida. La permanente evolución de las FPGA a lo largo del tiempo tanto en capacidad como en velocidad, así como su reducción en costo y consumo de energía, hacen que estos dispositivos prometen ser una de las principales plataformas de implementación de complejos clasificadores portátiles en el futuro.

Metodología

A lo largo del desarrollo de este proyecto se llevó a cabo un profundo proceso de análisis y procesamiento de la señal de voz, partiendo de las grabaciones originales hasta obtener una codificación en trenes de pulsos o spikes que capturan las características distintivas de las pronunciaciones. Utilizando las modernas herramientas que brinda el procesamiento digital de señales, se construyeron espectrogramas mediante la conocida transformada de Fourier de tiempo corto, empleando ventanas temporales solapadas y una escala logarítmica de frecuencias conocida como escala mel, que es especialmente adecuada para las señales de voz. Para eliminar distorsiones no informativas asociadas al proceso de ventaneo, se exploró el filtrado digital de cada banda espectral, comparando filtros FIR e IIR que suavizan notablemente las componentes frecuenciales. Sobre la base de esta representación tiempo-frecuencia optimizada, se ensayaron diversos métodos de codificación hacia secuencias binarias de pulsos o "spikes", que constituyen el "lenguaje" natural de comunicación neuronal. Tras comparar técnicas como la detección de máximos locales o el uso de umbrales adaptativos, resultó seleccionada una estrategia de umbral fijo sin inhibición lateral entre bandas, por brindar la mejor relación entre densidad de pulsos codificados y capacidad discriminatoria de patrones.

Con el propósito de abordar la tarea de clasificación, se ha propuesto el diseño de una nueva red neuronal supervisada denominada DELSNN (Digital Extreme Learning Spiking Neural Network). Esta red se inspira en un trabajo clásico [26], al cual se le han incorporado aspectos novedosos, tales como la digitalización y otros elementos vinculados al aprendizaje máquinal, entre los que se destaca la inclusión de una capa de aprendizaje extremo (ELL). Estas incorporaciones devinieron en la creación de una nueva arquitectura de SNN que incorpora proyecciones aleatorias para la expansión de características, combinadas con el entrenamiento directo de la capa de salida mediante la minimización de la entropía. En la Figura 4 se muestra un esquema de la red diseñada. La DELSNN sigue una arquitectura de tipo "feedforward" con dos capas de neuronas, etiquetadas como **I** y **J**. Los spikes de entrada son tratados como una sucesión temporal de pulsos, representados por un vector **V** que opera como una ventana deslizante sobre dichos spikes. Cada neurona en la capa **I** se conecta con todos los elementos de **V**, mientras que cada neurona en la capa **J** recibe conexiones de todas las neuronas en la capa **I**. La capa I opera como una capa de proyección que consta de **M** neuronas, y los pesos entre el vector **V** y esta capa se asignan de forma aleatoria, implementando así la estructura ELL de la red. Por otro lado, la capa **J** consta de **D** neuronas, donde cada una está asociada a una de las clases a reconocer. Los pesos entre las capas **I** y **J** son ajustados mediante un algoritmo de entrenamiento. Es importante aclarar que cada conexión entre el vector de entrada y la capa **I** se compone de múltiples líneas de retardos de propagación, donde cada una se encuentra caracterizada por un retardo temporal $\mathbf{d}^{\mathbf{k}}$ y un peso particular $\mathbf{W}^{\mathbf{k}}$, lo que da una idea de la complejidad del sistema.

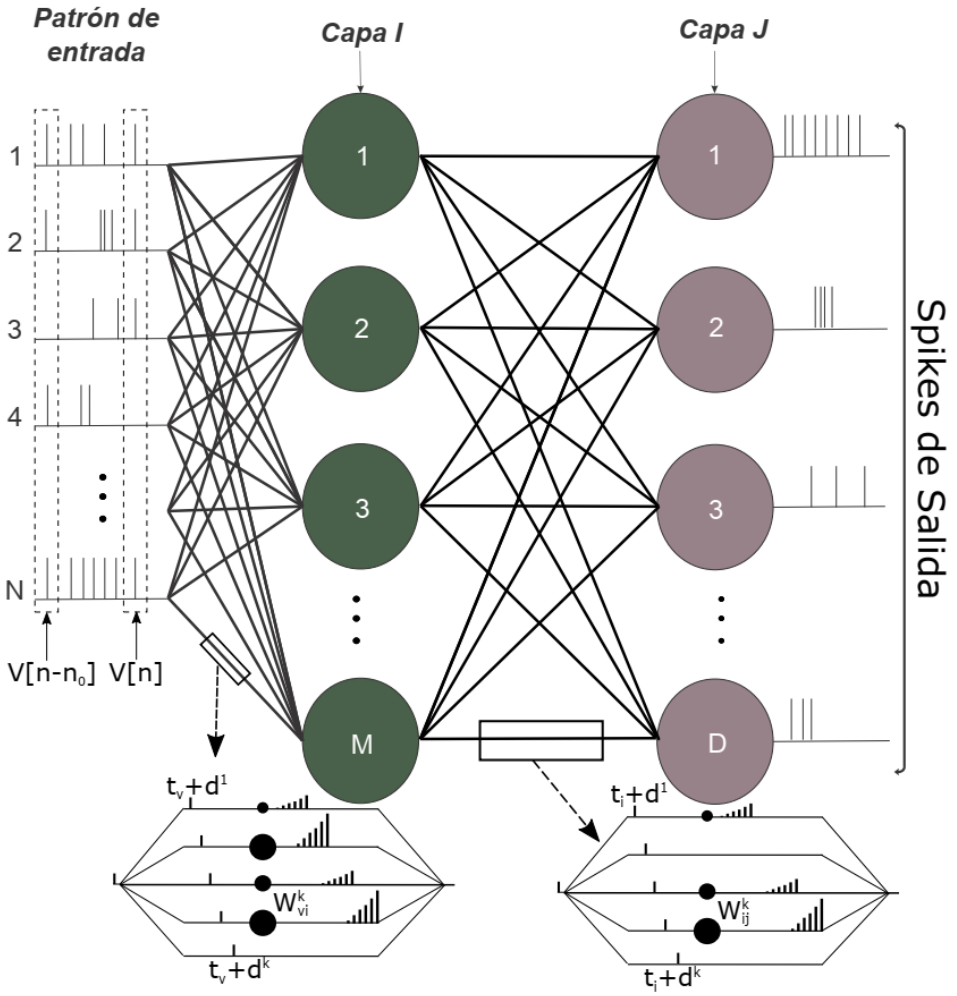


Figura 4: Estructura de la DELSNN. Cada conexión entre neuronas se compone de múltiples líneas de retardo que poseen diferentes retardos y pesos.

La estructura detallada ofrece una representación clara de cómo la DELSNN procesa secuencias temporales de pulsos. La utilización de conexiones con retardos y pesos variables entre capas refleja una aproximación más realista al procesamiento neuronal y destaca el enfoque innovador de la red propuesta.

Uno de los principales logros de este proyecto consiste en la exitosa implementación del modelo neuronal y del clasificador completo sobre una plataforma FPGA de bajo costo. Mediante meticulosas técnicas de modelado y cuantización fue posible mapear la red DELSNN con total funcionalidad sobre la lógica programable. Si bien la versión actual dista de explotar al máximo los recursos disponibles del dispositivo, la misma demuestra la factibilidad de este enfoque para aplicaciones de clasificación de voz en tiempo real. En el artículo [27] publicado en el marco de este proyecto, se describe al detalle el diseño y mapeo de la DELSNN en una FPGA.

En la Figura 5 se muestra un diagrama en bloques donde se puede apreciar la estructura jerárquica de los elementos involucrados. Dentro de la FPGA se implementa la DELSNN cuyo funcionamiento es controlado mediante la UNDAD DE CONTROL GE-

NERAL. Dicha unidad se encarga de controlar e intercambiar los spikes entrantes y salientes de la red con la PC a través de un bloque UART (Universal Asynchronous Receiver-Transmitter). Para cada paso de simulación, la unidad de control coloca en su señal de salida ***vector_in*** un vector binario con los spikes de entrada a la SNN. Posteriormente activa la señal ***clk_spk***. La SNN procesa la entrada y coloca en su salida ***spikes_out*** otro vector binario con las respuestas de cada una de las neuronas de salida para ese paso de simulación. Posteriormente, la SNN activa su salida ***end_processing_SNN*** para indicarle a la unidad de control que ya terminó de procesar las entradas. La unidad de control se encarga, a través de la UART, de enviar a la PC dichas salidas y de recibir las próximas entradas para el próximo paso de simulación. Se utiliza un ***botón pulsador*** para generar una señal de reinicio del sistema completo. También se utilizan los **LED** de la placa para informar posibles problemas o errores en la transmisión de los datos.

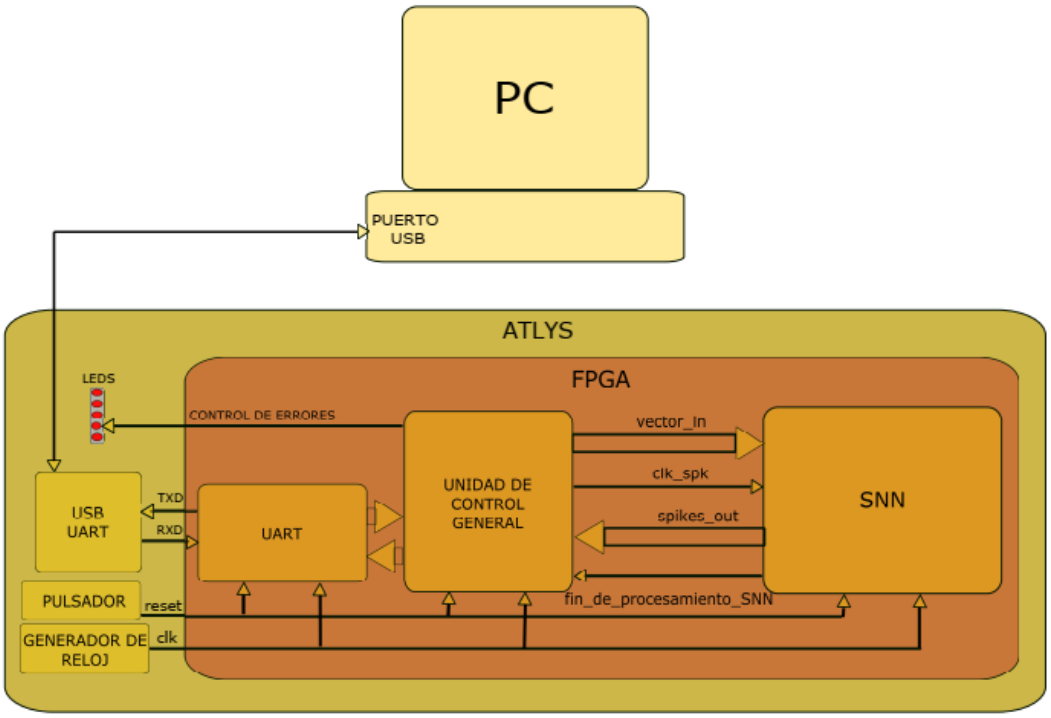


Figura 5: Diagrama en bloques del sistema DELSNN implementado en la FPGA

El diseño del modelo neuronal se orienta a replicar de manera exacta el funcionamiento de las neuronas en el modelo computacional. Se optó por la modelización de las neuronas a través de un enfoque denominado Spike Response Model (SRM), el cual ofrece una representación de la evolución del potencial de membrana equilibrada en términos de plausibilidad biológica y eficiencia en el consumo de recursos. En la Figura 6, se presenta un diagrama que ilustra la interfaz y el diseño interno de cada una de las neuronas en el modelo. La propuesta de diseño reduce significativamente la complejidad de la implementación en FPGA mediante la utilización de acumuladores y registros de desplazamiento para representar la evolución del potencial de membrana. Además, el modelo permite la aplicación de spikes de manera consecutiva en el tiempo, una capacidad que hasta el momento no ha sido evidenciada en ninguno de los modelos de

neuronas SRM implementados en FPGA. La Figura 6 a) detalla los puertos de entrada y salida del modelo neuronal, mientras que la Figura 6 b) describe la arquitectura interna de la neurona y el modelo de sinapsis diseñado.

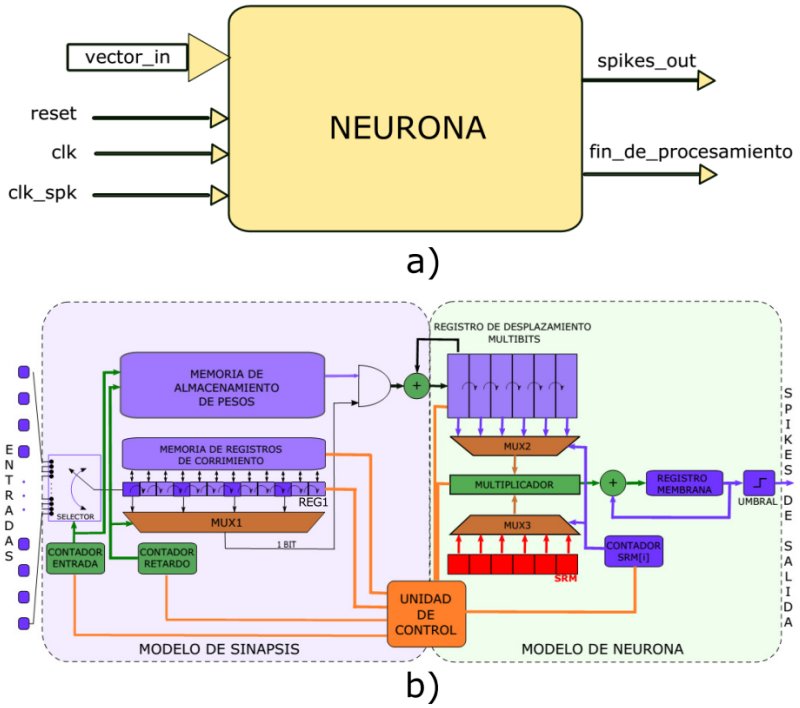


Figura 6: Interfaz y diseño interno de cada una de las neuronas en el modelo diseñado.

Para más detalles sobre el funcionamiento de la red, como su entrenamiento e implementación en FPGA, puede remitirse a dos publicaciones en congresos internacionales que se realizaron en el marco de este proyecto, las cuales describen en detalle dichos aspectos [27, 28].

En la Figura 7 se representa un diagrama que ilustra un resumen de la metodología para el diseño de la DELSNN utilizada en este proyecto. El proceso se inicia con la adquisición de datos de voz provenientes de bases de datos estándar. En la tarea de reconocimiento de habla, se utilizó la base de datos TI-DIGITS [29], mientras que para el análisis de emociones, se optó por la base de datos RAVDESS [30]. Aunque la mayoría de los experimentos siguieron este enfoque, es importante destacar que a través de la tesis de grado realizada en el marco de este proyecto por Nahuel Ricart se llevó a cabo un proceso alternativo de obtención y limpieza de la señal de voz [31]. En esta tesis, se diseñó de un sistema de procesamiento digital de señales que facilita la adquisición y limpieza de la señal de voz mediante su implementación en FPGA. El proceso abarcó la captura de la señal de voz a través de un micrófono estándar, seguida de su digitalización y preprocesamiento, todo ello realizado internamente en la FPGA.

Una vez obtenida la señal de voz, ya sea mediante la adquisición por micrófono o desde una base de datos, se realiza la extracción de características espectrales y su codificación correspondiente mediante diversas estrategias de generación de spikes. Inicialmente, esta fase se realizó en computadora, de manera independiente a la FPGA. No obstante, de manera simultánea, esta etapa fue implementada en la FPGA gracias a otra tesis de grado desarrollada en el marco de este proyecto por Juan Rufiner [32].

La siguiente etapa abordó el diseño de la red, denominada DELSNN, que consta de capas de entrada, proyección aleatoria y salida. Simultáneamente al diseño de la red, se desarrolló el método de entrenamiento, que implica exclusivamente el ajuste de los pesos de la capa de salida utilizando el método de minimización de entropía relativa. Una vez completado el entrenamiento de la red, se llevó a cabo la implementación en FPGA para replicar fielmente el funcionamiento del modelo computacional. La comunicación entre el diseño computacional y el diseño en FPGA se estableció mediante un archivo de configuración que define los parámetros del modelo. La implementación en FPGA se realizó utilizando el lenguaje VHDL. Para verificar la correcta operación del sistema implementado en la FPGA, se diseñó un bloque de verificación que permite comparar las salidas de este con el modelo computacional, garantizando la identidad entre ambas salidas.

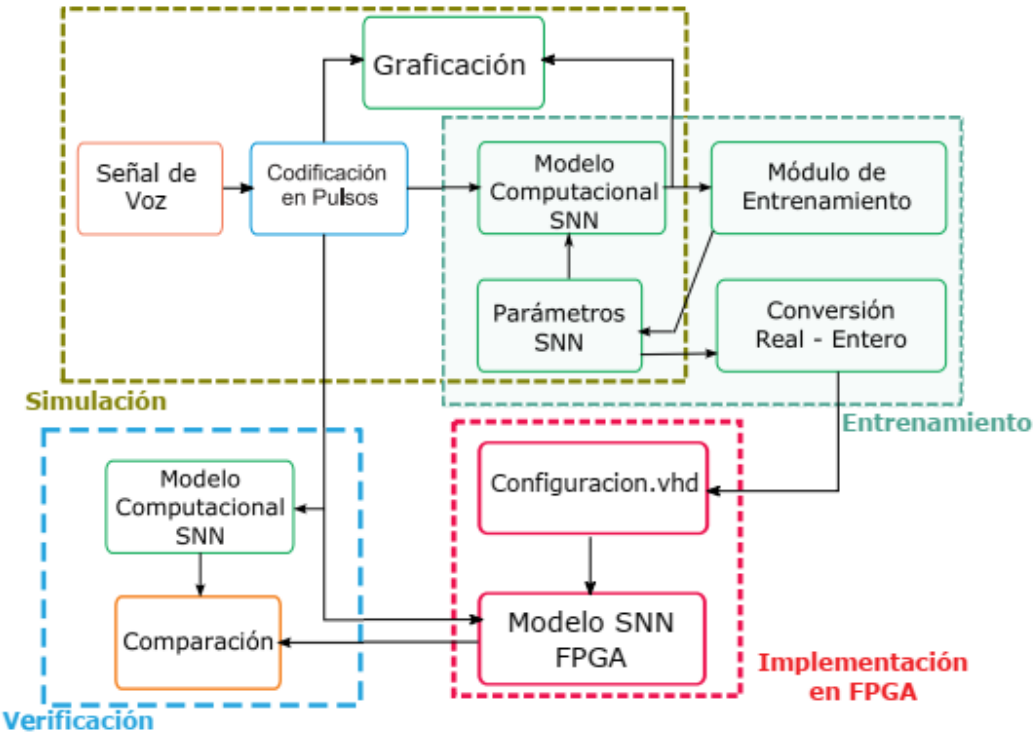


Figura 7: Esquema conceptual que representa de la metodología utilizada para el diseño, simulación entrenamiento, implementación y verificación de la DELSNN.

Síntesis de resultados y conclusiones

Los resultados obtenidos en el reconocimiento tanto de habla como de emociones podrían considerarse que están en una etapa preliminar, ya que aún no se han podido explorar estrategias de codificación más sofisticadas, que han demostrado mejoras significativas en tareas similares. A pesar de esto, el sistema propuesto demostró su capacidad de aprendizaje y logró tasas de reconocimiento aceptables para la clasificación del habla. En la tarea de reconocimiento de emociones, los resultados fueron bastante pobres, lo que refuerza la necesidad de continuar trabajando los aspectos relacionados con la codificación de las señales de voz a spikes.

La tarea de clasificación se basa en la premisa de que, tras entrenar la red, al proporcionarle una pronunciación como entrada, la red debería generar disparos en la

neurona de salida que coincida con la clase de la pronunciación proporcionada. En la práctica se observa que puede ocurrir que más de una neurona de salida produzca disparos, por lo que se definieron algunos criterios para determinar cuál es la neurona ganadora en estos casos. En la Figura 8se muestra la evolución de las curvas de error de reconocimiento global durante el proceso de entrenamiento. Las curvas VAL están asociadas con pronunciaciones que no fueron empleadas durante el entrenamiento, con el propósito de verificar la capacidad de generalización de la red. En la gráfica se describe una curva VAL para cada uno de los criterios utilizados para determinar la neurona ganadora. De todos los criterios empleados, el más efectivo resultó ser CMATS, el cual establece que la neurona ganadora es aquella que logra la máxima duración del tren de pulsos en su salida. Este criterio se utilizó para todas las tareas de clasificación realizadas. La curva ENT DURO está relacionada con las pronunciaciones usadas para entrenar la red, evidenciando que la totalidad de las mismas ha sido correctamente reconocida.

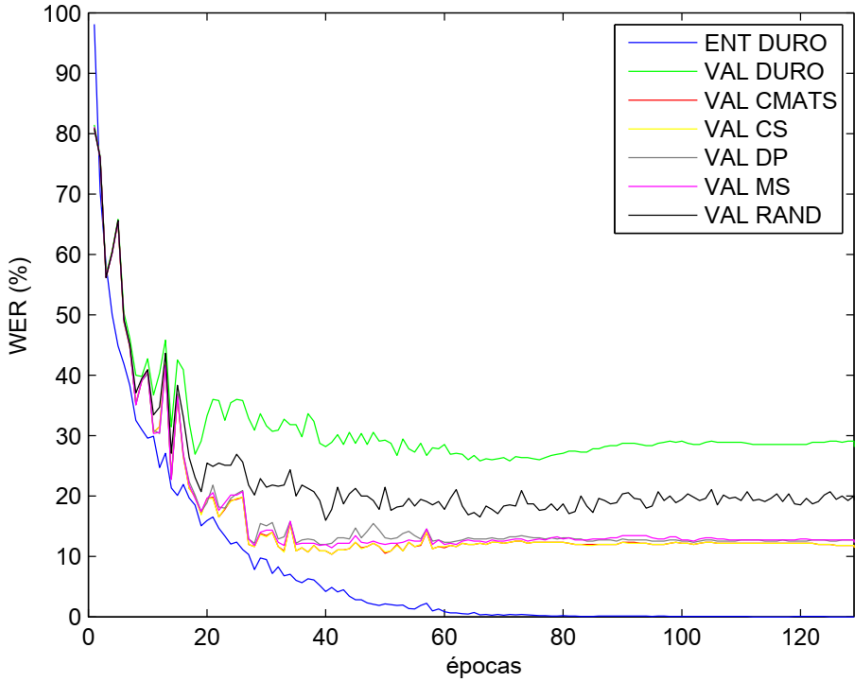


Figura 8: Error de reconocimiento WER (%). Se muestra la curva de entrenamiento (ENT) con el criterio DURO y las de validación (VAL) con todos los criterios de selección de neurona ganadora.

En la Figura 9, se puede observar la respuesta de la primera capa de neuronas al presentar un patrón de pulsos a la SNN. Se evidencia la ampliación de la representación del ejemplo en el dominio temporal, lo que incrementa las posibilidades de destacar características que podrían estar ocultas en la representación inicial de los datos de entrada. No obstante, expansiones demasiado exageradas conllevan un mayor costo computacional y retraso en la clasificación del ejemplo. Por lo tanto, es crucial seleccionar los parámetros de manera que equilibren estas necesidades.

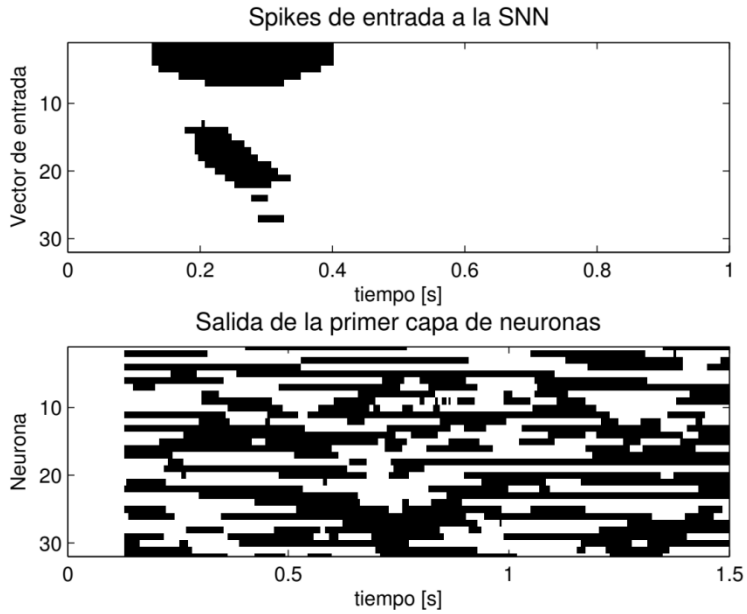


Figura 8: Respuesta de la primera capa de neuronas para una pronunciación del dígito <<three>>. Arriba: Patrón de entrada a la SNN. Abajo: respuesta de la primera capa de neuronas ante la presentación de dicho patrón.

Reconocimiento de dígitos aislados

El sistema diseñado ha demostrado su habilidad para reconocer las clases objetivo, mostrando un rendimiento aceptable en comparación con las arquitecturas clásicas del estado del arte, según el conjunto de evaluación de la base de datos TIDIGITS. Además, ha alcanzado un desempeño robusto en la clasificación en condiciones de habla ruidosa. Los resultados más destacados para diversos niveles de relación señal-ruido (SNR) y habla limpia se presentan de manera resumida en la Tabla 1. Según la configuración específica de cierto parámetro (Ψ), que representa el umbral para la obtención de los pulsos a partir del espectrograma en escala de mel, se observa que la precisión se mantiene relativamente constante hasta los 10 dB de SNR. Esto señala la capacidad de tolerancia al ruido proporcionada por la codificación y la arquitectura de la SNN. Sin embargo, el rendimiento se degrada más bruscamente por debajo de 10 dB de SNR, ya que el ruido enmascara completamente las características discriminativas del habla. No obstante, la DELSNN puede operar en niveles de SNR donde los sistemas convencionales basados en MFCC y HMM tienen tasas de error de casi el 70%. Para valores mayores de Ψ, se mantiene la estabilidad de la precisión para niveles más altos de ruido, pero el reconocimiento del habla en señales limpias se deteriora ligeramente.

La mayor tolerancia al ruido de la red DELSNN propuesta en comparación con los sistemas convencionales de reconocimiento del habla destaca el potencial de los enfoques basados en SNN. Como se indicó anteriormente, consideramos que se pueden obtener mejoras adicionales al explorar estrategias alternativas de codificación de pulsos que sean robustas al ruido, como la Codificación Auditiva Biológicamente Plausible (BAE) [33] que se presenta en la Tabla 1, evidenciando un rendimiento superior.

Tabla 1. Comparación de diferentes reconocedores de habla para esta tarea bajo varios niveles de ruido blanco. Los valores faltantes corresponden a datos no informados en la bibliografía.

Strategy	Threshold Ψ	Clean	20 dB	10 dB	5 dB	0 dB
Lyon ear + LSM [34]	-	97.50%	84.00%	79.50%	-	-
LAM + HMM [35]	-	**98.80%**	**95.75%**	72.79%	-	-
MFCC + HMM [35]	-	**98.80%**	27.50%	12.20%	-	-
BAE + SNN [33]	W/masking	97.40%	91.90%	87.50%	-	**78.20**
MSE+DELSNN	0.785	79.65%	79.44%	79.81%	**80.37%**	58.57%
	0.576	82.58%	82.54%	83.15%	78.76%	18.02%
	0.370	89.14%	89.10%	**88.2%**	43.08%	1.81%

Reconocimiento en habla continua

Aunque tanto el entrenamiento como la evaluación del reconocimiento se llevaron a cabo utilizando dígitos aislados, el sistema implementado también es aplicable a la detección de dígitos en habla continua. Es decir, después de entrenar la red con pronunciaciones de dígitos aislados, es posible evaluarla con pronunciaciones de dígitos conectados. En la Figura 9, se presentan las variaciones de las variables de estado pertenecientes a las neuronas de la última capa durante la pronunciación de una secuencia particular de dígitos. El reconocimiento de cada pronunciación se logra mediante la emisión de pulsos en la neurona correspondiente durante un periodo cercano al final del dígito pronunciado. Es importante considerar que, debido a la resolución temporal de la gráfica, se producen varios pulsos de salida para cada detección en lugar de un único pulso, como podría apreciarse. Asimismo, se llevó a cabo una interpolación en el dominio del tiempo para una mejor comprensión de la evolución temporal de las variables de estado.

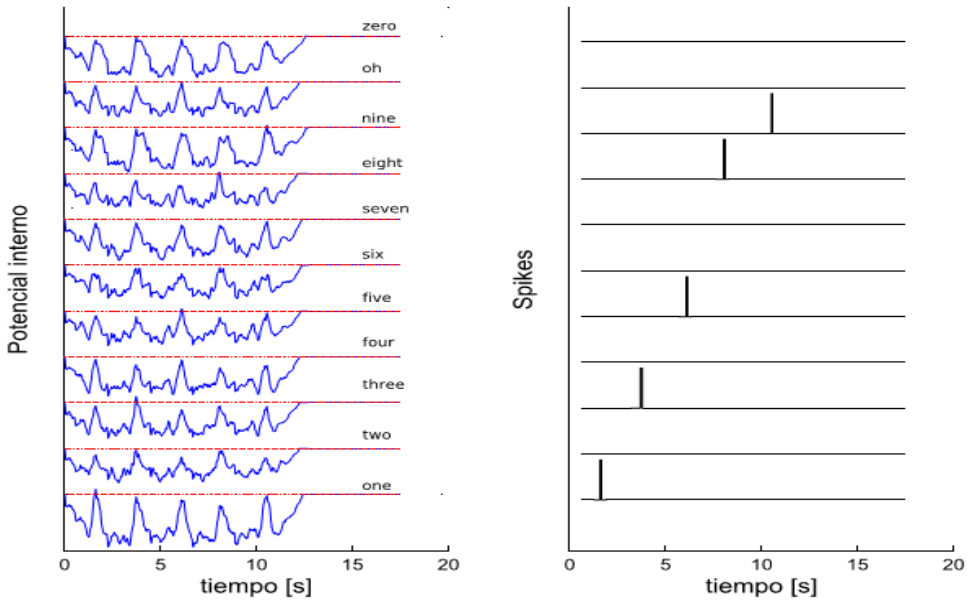


Figura 9: Salidas de las neuronas de la última capa para la presentación de la secuencia de dígitos de evaluación <1 3 5 8 9>. La gráfica de la derecha muestra la variación del potencial interno o variable de estado como función del tiempo. A la derecha se muestran los spikes de salida de cada neurona. La línea de trazos indica el umbral.

Reconocimiento de emociones

Para entrenar y evaluar la red se utilizó la base de datos RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [30]. La base de datos incluye a 24 actores profesionales (12 mujeres y 12 hombres), vocalizando dos declaraciones lexicalmente emparejadas en un acento neutro de América del Norte. El discurso incluye expresiones de calma, felicidad, tristeza, enojo, miedo, sorpresa y disgusto. Cada expresión se produce en dos niveles de intensidad emocional (normal, fuerte), con una expresión neutral adicional.

Se llevaron a cabo varios experimentos donde se garantizó que el proceso de reconocimiento sea independiente del hablante. Los resultados alcanzados para esta tarea de reconocimiento de emociones no han logrado las expectativas deseadas y están muy por debajo del estado del arte actual, no obstante, se sigue trabajando en la mejora de su desempeño con cierta expectativa de éxito dado que todavía existen múltiples aspectos que aún no han sido explorados. En ese sentido se experimentó con diferentes estrategias de codificación basadas en las características espectrales, pero ninguna logro superar una tasa de reconocimiento superior al 40%. En la Figura 10 se muestra el desempeño para uno de los experimentos donde se observa que la red es capaz de aprender a la perfección los ejemplos de entrenamiento, pero le es muy difícil generalizar dicho aprendizaje a los ejemplos de evaluación.

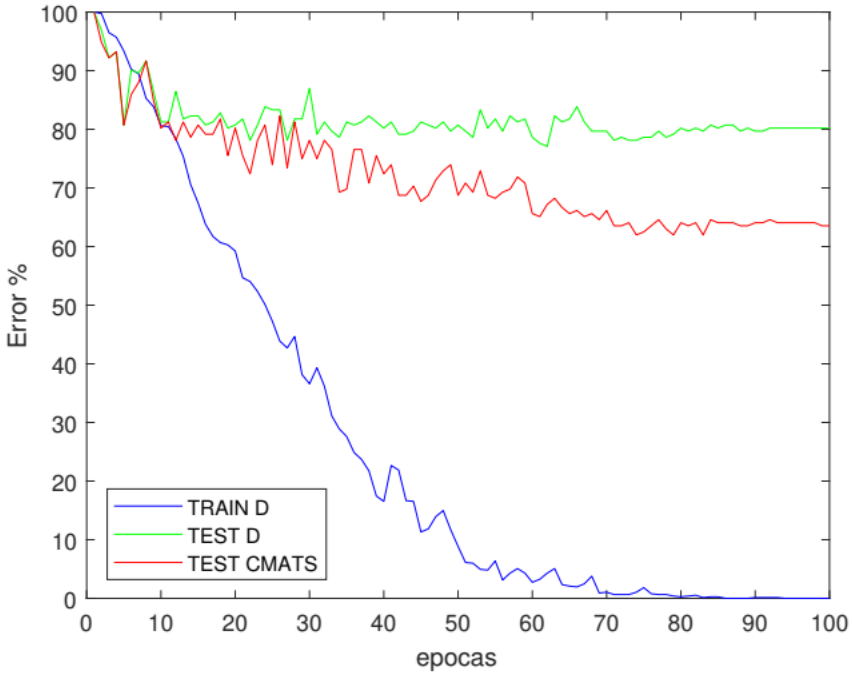


Figura 10: Error de reconocimiento para los conjuntos de entrenamiento y prueba durante el entrenamiento para la estrategia “Espectrograma”.

Implementación en FPGA

La implementación en VHDL de la DELSNN es completamente sintetizable en la FPGA. Como ejemplo en la Tabla 2 se muestran los recursos utilizados en diferentes

estructuras de la red para el dispositivo Spartan 6 XC6SLX45 con una representación de pesos de 8 bits. Debido a la cuantización de 8 bits, la mayoría de las BRAM se utilizaron en la memoria de registros de desplazamiento, que se emplean para los modelos de sinapsis entre neuronas. En el caso de 32x4x16 y 1920 sinapsis, solo se utilizaron 30 de las 116 BRAM disponibles. El gran número de slices utilizados se debe principalmente al hecho de que todos los registros internos del diseño son de tipo entero (32 bits). Cabe destacar que esta excesiva resolución de bits no es necesaria para la mayoría de los registros internos de las neuronas. Todos los casos en la Tabla se probaron con una frecuencia de reloj de 100 MHz, y el peor de los casos limita la frecuencia máxima posible a 117 MHz.

Tabla 2: Comparación de la utilización de recursos lógicos en diferentes estructuras de redes con representación de pesos de 8 bits para K=3 (en la parte superior de cada fila) y K=10 (en la parte inferior de cada fila) retrasos/pesos por conexión, respectivamente.

DELSNN Structure	Synapses	Slice Registers	Slice LUTs	Block RAM
2x2x2	24 80	1599 2271	3152 3641	4 6
9x9x4	351 1170	5112 7296	10028 11673	13 20
32x4x16	576 1920	7145 10569	11312 14188	20 30

Como se mencionó previamente, en el contexto de este proyecto se llevaron a cabo dos tesis de grado. La tesis de Nahuel Ricart se centró en desarrollar un sistema con la capacidad de adquirir y mejorar la calidad e inteligibilidad de señales de voz afectadas por ruido. Para alcanzar este objetivo, se implementó un algoritmo de sustracción espectral que fue evaluado en una FPGA Artix-7. Los resultados obtenidos demostraron la eficacia del sistema para reducir el ruido presente en las señales de audio, preservando al mismo tiempo las características espectrales fundamentales del habla, como las formantes. Se realizaron pruebas utilizando tanto señales sintéticas como muestras de voz reales adquiridas con un micrófono. En ambos casos, el sistema implementado logró una reducción significativa del ruido de fondo. Además, se llevó a cabo una evaluación objetiva de la inteligibilidad de las señales procesadas, mediante el cálculo del índice de transmisión del habla. Los resultados indicaron que el método de sustracción espectral no produce una pérdida grave de inteligibilidad, por lo que podría ser incorporado previo a un sistema automático de reconocimiento del habla. El sistema desarrollado también se evaluó en conjunto con una red neuronal artificial previamente entrenada para la clasificación de dígitos. Se observó una mejora significativa en la tasa de reconocimiento de dígitos para señales con bajas relaciones señal-ruido después de aplicar el algoritmo de sustracción espectral. La Figura 11 ilustra cómo mejoran las tasas de reconocimiento de la red tras la limpieza de la señal mediante sustracción espectral. Asimismo, en la Figura 12 se presenta la aplicación de software diseñada para interactuar con la FPGA.

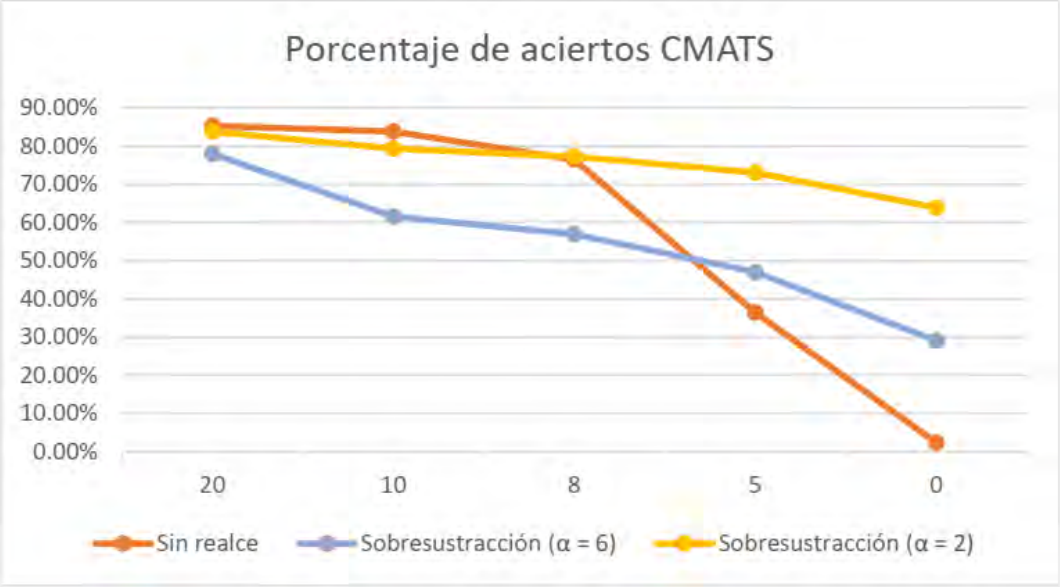


Figura 11: Porcentaje de aciertos de la SNN para diferentes niveles de ruido aplicando la sustracción espectral previa.

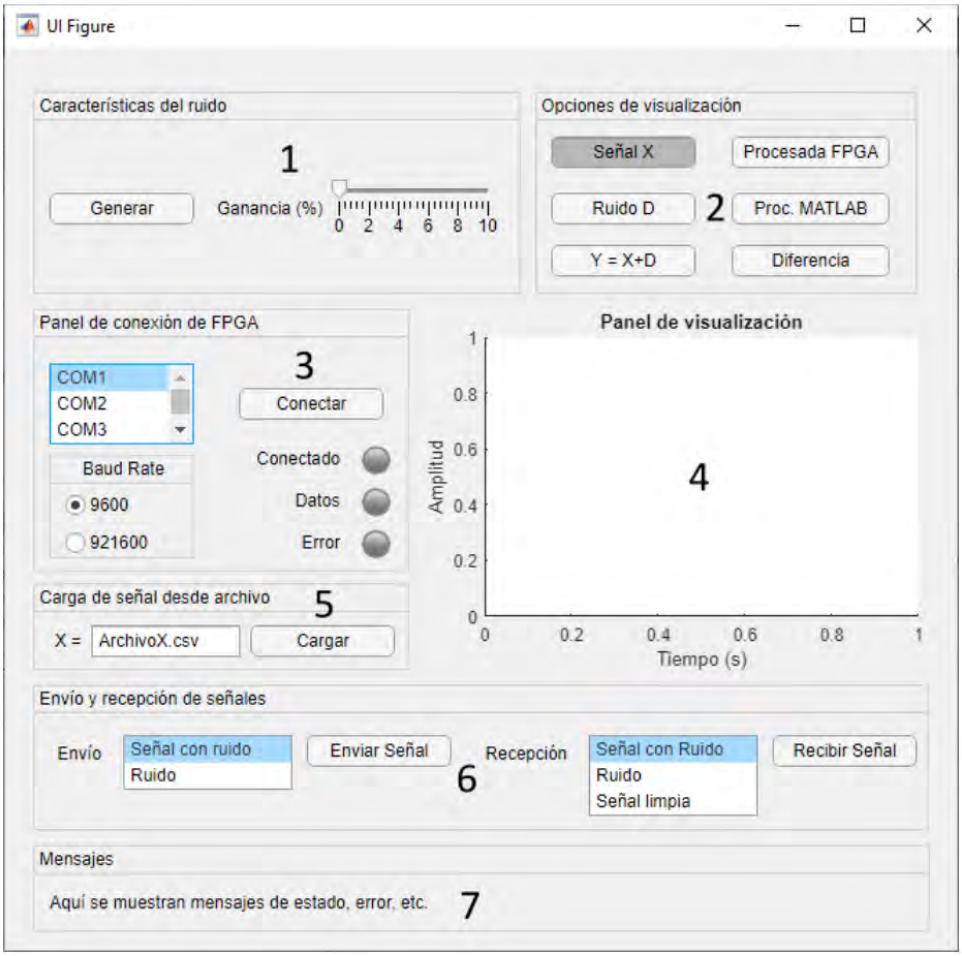


Figura 12: Interfaz gráfica del software utilizado para comunicarse con la FPGA en el sistema de adquisición y limpieza de ruido desarrollada en la Tesis de grado de Nahuel Ricart.

Otra de las tesis de grado desarrolladas en el marco de este proyecto de investigación fue llevada a cabo por Juan Rufiner. En su tesis, se abordó el diseño y desarrollo del sistema de codificación de la señal de habla en la FPGA. La implementación se realizó utilizando una placa de desarrollo Arty de la empresa Digilent, que incorpora el chip de FPGA Artix-35T de Xilinx, considerado un dispositivo con recursos limitados. Se exploraron diversas alternativas para la codificación en spikes basadas en el cálculo del espectrograma a partir de la FFT, con el objetivo de optimizar el rendimiento general del sistema. Las diferentes propuestas fueron evaluadas mediante pruebas del sistema de codificación, alimentando la red DELSNNN en el problema de reconocimiento de dígitos aislados. A partir de los resultados obtenidos, se determinó que la mejor opción era el codificador implementado con FFT de 128 muestras. Este sistema logró la mejor tasa de reconocimiento global, ocupó la menor cantidad de recursos en la placa y presentó el consumo más eficiente. Los resultados obtenidos validan la viabilidad de las propuestas formuladas en este estudio, logrando un óptimo aprovechamiento de los recursos disponibles en el sistema. Esto posibilita el uso del mismo chip de FPGA para un sistema integral de reconocimiento. Además, dado el reducido tiempo de respuesta evaluado mediante simulación, se puede afirmar que el dispositivo es apto para aplicaciones en tiempo real. La Figura 13 exhibe la interfaz de usuario desarrollada en esta tesis, la cual permite realizar un análisis comparativo de las distintas codificaciones implementadas dentro y fuera de la FPGA. En futuras investigaciones, se anticipa la integración de todas las etapas (eliminación de ruido, codificación y sistema de clasificación) en un solo dispositivo, seguido de las pruebas pertinentes para comparar la implementación final con los resultados obtenidos de manera preliminar.

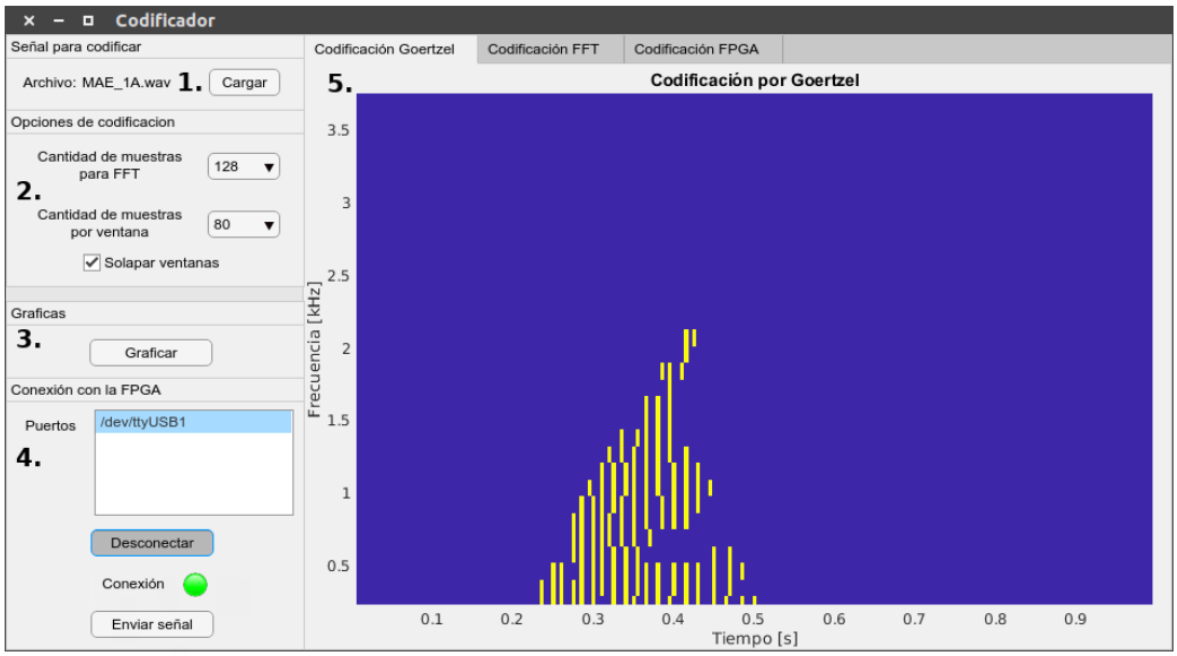
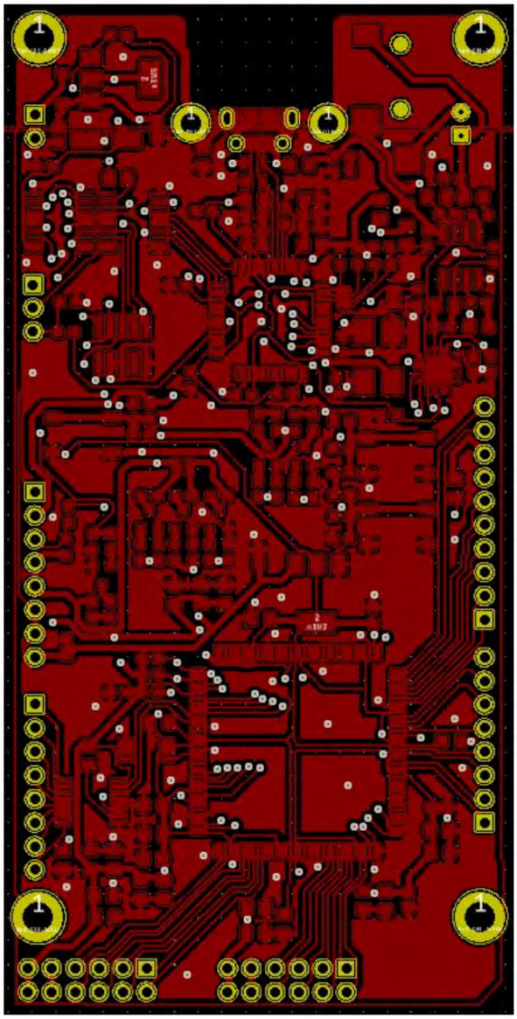


Figura 13: Interfaz Gráfica de usuario diseñada en la que se muestra un espectrograma para una pronunciación del dígito <<one>>por un hablante masculino.

Aunque se implementaron los sistemas propuestos en FPGAs que formaban parte de kits de desarrollo comerciales, además de los trabajos mencionados anteriormente, se

llevó a cabo el diseño e implementación de una plataforma basada en FPGA para la adquisición de señales de voz. Este logro fue posible gracias a la participación de un becario de iniciación en la investigación, una figura contemplada en los Proyectos de Investigación y Desarrollo (PID). Lamentablemente, debido a las significativas devaluaciones del peso argentino y a la necesidad de adquirir todos los componentes electrónicos en el extranjero, no fue posible concluir el desarrollo final del sistema. A pesar de ello, se logró avanzar hasta la etapa de desarrollo, dejando los diagramas esquemáticos y los diagramas PCB disponibles para futuros desarrollos. La Figura 14 muestra el PCB logrado en una placa de dos capas. Además, con la intención de fortalecer el conocimiento en sistemas embebidos y en el diseño y fabricación de circuitos electrónicos, el director de este PID participó como director de la tesis de posgrado titulada "Monitor de ECG/movimiento personal para sistema telemétrico basado en telefonía móvil", realizada por el Bioing. Julián Botello. La tesis fue aprobada con calificación sobresaliente y proporcionó valiosos conocimientos al equipo de trabajo del PID en cuanto al diseño electrónico, los cuales han sido de gran utilidad en el desarrollo del kit FPGA mencionado anteriormente.



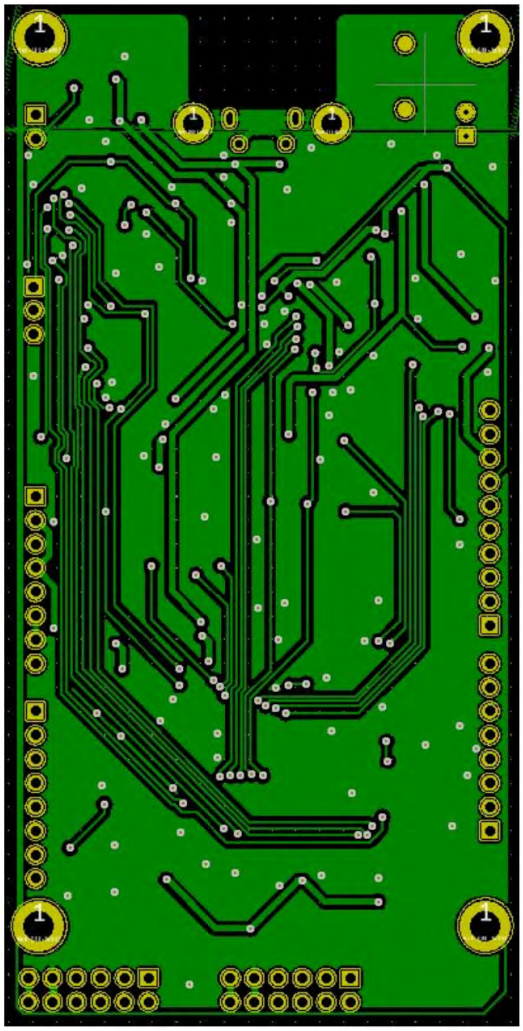


Figura 14: Diseño en PCB del kit diseñado: cara frontal (arriba), cara posterior (abajo)

Conclusiones

El presente proyecto permitió sentar bases sólidas tanto en el procesamiento e interpretación de señales de voz, como en el desarrollo de modelos neuronales pulsantes y su implementación eficiente en hardware digital reconfigurable. La propuesta de red neuronal DELSNN resultó prometedora para la clasificación de palabras aisladas y secuencias simples de dígitos conectados. Si bien en estados emotivos no se obtuvieron los resultados esperados, existen aún múltiples estrategias de codificación y arquitecturas de red por explorar en el futuro. Se logró con éxito la implementación de la red neuronal propuesta en una FPGA de bajo costo utilizando VHDL, habilitando aplicaciones de clasificación de voz en tiempo real. Las tesis de grado permitieron integrar etapas previas de adquisición y limpieza de la señal, así como alternativas de codificación en spikes, sentando bases para un sistema completo implementado internamente en el hardware reconfigurable. La incorporación de la capa de aprendizaje extremo introdujo mejoras significativas tanto en la tasa de aprendizaje como en la capacidad de generalización de la red neuronal diseñada. Este proyecto contó también con la participación de un becario de iniciación en la investigación y gracias

a su contribución, se llevó a cabo el diseño e implementación de una plataforma electrónica basada en FPGA destinada específicamente a la adquisición de señales de voz. Si bien por limitaciones presupuestarias no pudo concluirse la construcción y prueba final del dispositivo, el becario desarrolló completamente los diagramas esquemáticos y de circuito impreso, sentando así las bases para el desarrollo de un sistema de adquisición de voz propio en futuros trabajos. En síntesis, el proyecto cumplió satisfactoriamente los objetivos planteados inicialmente, afianzando una sólida plataforma para continuar esta prolífica línea de investigación sobre el procesamiento de voz con redes neuronales pulsantes y su implementación en FPGA.

Indicadores de producción

Se pueden destacar indicadores de producción de distinta índole:

- Publicaciones:
    - Un artículo publicado en la conferencia internacional Speech and Computer (SPECOM) 2023. Título: "Extreme Learning Layer: A Boost for Spoken Digit Recognition with Spiking Neural Networks".Este artículo se encuentra publicado y en la revista Lecture Notes in Computer Science, vol 14338. De la editorial Springer, una de las plataformas de mayor prestigio en publicaciones científicas.
    - Un artículo presentado en la Conferencia Southern Programmable Logic (SPL) 2019. Título: "A new Spiking Neural Network with Extreme Learning for FPGA implementation". Este artículo describe las bases de la
- Formación de Recursos Humanos:
    - 1 tesis de posgrado dirigida (Julián Botello, FIUNER, 2021).
    - 1 becario de iniciación (Osvaldo Marcos Zanet, FIUNER, 2019-2021).
    - 2 tesis dirigidas (Juan Ignacio Rufiner y Nahuel Ricart, FIUNER, 2020).
- Transferencia:
    - Presentación del PID en el Ciclo de Seminarios I+D+i 2019 de la FIUNER.
- Implementación en FPGA:
    - Desarrollo en VHDL y pruebas en placa de desarrollo con FPGA para validar la implementación en hardware.
    - Diagramas esquemáticos y de circuito impreso del diseño e implementación de una plataforma electrónica basada en FPGA destinada específicamente a la adquisición de señales de voz.

En resumen, los principales indicadores son 2 publicaciones internacionales, formación de 5 recursos humanos, 1 actividad de transferencia y la implementación en hardware del sistema propuesto.

Bibliografía

[1] Wang, D., Wang, X., &Lv, S. (2019). An overview of end-to-end automatic speech recognition. Symmetry, 11(8), 1018.

[2] Hannun, A. (2021). The history of speech recognition to the year 2030. arXiv preprint arXiv:2108.00084.

[3] Dua, M., Akanksha, &Dua, S. (2023). Noise robust automatic speech recognition: review and analysis. International Journal of Speech Technology, 1-45.

[4] Khe Chai Sim, Fran¸coiseBeaufays, Arnaud Benard, DhruvGuliani, Andreas Kabel,-

Nikhil Khare, Tamar Lucassen, Petr Zadrazil,Harry Zhang, Leif Johnson, et al. Personalization of end-to-end speech recognition onmobile devices for named entities. In 2019IEEE Automatic Speech Recognition and UnderstandingWorkshop (ASRU), pages 23{30.IEEE, 2019.

[5] Khe Chai Sim, Petr Zadrazil, and Fran¸coiseBeaufays. An investigation into ondevice personalization of end-to-end automatic speech recognition models. arXiv preprint arXiv:1909.06678, 2019.

[6] Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., &Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. IEEE Access, 7, 117327-117345.

[7] Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. Psychological bulletin, 99(2), 143.

[8] Petrushin, V. A. (2000). Emotion recognition in speech signal: experimental study, development, and application. studies, 3(4).

[9] Tato, R., Santos, R., Kompe, R., & Pardo, J. M. (2002). Emotional space improves emotion recognition. In Seventh International Conference on Spoken Language Processing.

[10] Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF (2015) A review of depression and suicide risk assessment using speech analysis. Speech Communication 71:10–49.

[11] Devillers L, Vidrascu L (2007) Speaker Classification II: Selected Projects, Lecture Notes in Computer Science, vol 4441/2007, Springer-Verlag, Berlin, Heidelberg, chap Real-Life Emotion Recognition in Speech, pp 34–42.

[12] Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. International Journal of Human Computer Studies 65(8):724 – 736

[13] Clavel C, Vasilescu I, Devillers L, Richard G, Ehrette T (2008) Fear-type emotion recognition for future audio-based surveillance systems. Speech Communication 50(6):487–503.

[14] Tacconi D, Mayora O, Lukowicz P, Arnrich B, Setz C, Tröster G, Haring C (2008) Activity and emotion recognition to support early diagnosis of psychiatric diseases. In: Proceedings of 2nd International Conference on Pervasive Computing Technologies for Healthcare, Tampere, Finland, pp 100–102.

[15] Chin YH, Lin SH, Lin CH, Siahaan E, Frisky A, Wang JC (2014) Emotion Profile-Based Music Recommendation. Proc 7th International Conference on Ubi-Media Computing and Workshops (UMEDIA) pp 111–114.

[16] Yildirim S, Narayanan S, Potamianos A (2011) Detecting emotional state of a child in a conversational computer game. Computer Speech & Language 25(1):29 – 44. Yu, F., Chang, E., Xu, Y. Q., & Shum, H. Y. (2001). Emotion detection from speech to enrich multimedia content. Advances in multimedia information processing—PCM 2001, 550-557.

[17] Bohnstingl, T., Garg, A., Woźniak, S., Saon, G., Eleftheriou, E., &Pantazi, A. (2022, May). Speech Recognition Using Biologically-Inspired Neural Networks. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6992-6996). IEEE.

[18] Rathi, N., Chakraborty, I., Kosta, A., Sengupta, A., Ankit, A., Panda, P., & Roy, K. (2023). Exploring neuromorphic computing based on spiking neural networks: Algorithms to hardware. ACM Computing Surveys, 55(12), 1-49.

[19] Stolzle, Anton, Narayanaswamy, Shankar, Murveit, Hy, Rabaey, Jan M y Brodersen,

Robert W: Integrated circuits for a real-time large-vocabulary continuous speech recognition system. IEEE Journal of Solid-State Circuits, 26(1):2–11, 1991.
[20] Price, Michael, Glass, James y Chandrakasan, Anantha P: A 6 mW, 5,000-Word Real-Time Speech Recognizer Using WFST Models. IEEE Journal of Solid-State Circuits, 50(1):102–112, 2015.
[21] Lin, Edward C y Rutenbar, Rob A: A multi-FPGA 10x-real-time high-speed search engine for a 5000-word vocabulary speech recognizer. En Proceedings of the ACM/SIGDA international symposium on Field programmable gate arrays, páginas 83–92. ACM,2009.
[22] Price, Michael, Glass, James y Chandrakasan, Anantha P: A 6 mW, 5,000-Word Real-Time Speech Recognizer Using WFST Models. IEEE Journal of Solid-State Circuits, 50(1):102–112, 2015.
[23] Misra, Janardan y Saha, Indranil: Artificial neural networks in hardware: A survey of two decades of progress. Neurocomputing, 74(1):239–255, 2010.
[24] Isik, M. (2023). A survey of spiking neural network accelerator on fpga. arXiv preprint arXiv:2307.03910.
[25] Saravanan, K., &Kouzani, A. Z. (2023). Advancements in On-Device Deep Neural Networks. Information, 14(8), 470.
[26] Unnikrishnan, KP, Hopfield, John J y Tank, David W: Connected-digit speaker dependent speech recognition using a neural network with time-delayed connections. IEEE Transactions on Signal Processing, 39(3):698–713, 1991.
[27] Peralta, I., Odetti, N., Filomena, E., Rufiner, J., Ricart, N., Rufiner, H.L. (2019). A new spiking neural network with extreme learning for FPGA implementation. En Proceedings of the 10th Southern Programmable Logic Conference (pp. 49-54).
[28] Peralta, I., Odetti, N., Rufiner, H.L. (2023). Extreme Learning Layer: A Boost for Spoken Digit Recognition with Spiking Neural Networks. En 25th International Conference on Speech and Computer SPECOM 2023.
[29] Leonard, R Gary y Doddington, George: Tidigits. Linguistic Data Consortium, Philadelphia, 1993.
[30] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.
[31] Ricart, Nahuel Emiliano. (2020). Implementación en FPGA de un Sistema de Limpieza de la Señal de Habla en Ambiente Ruidoso (Tesis de grado). Facultad de Ingeniería, Universidad Nacional de Entre Ríos (FIUNER), Paraná, Argentina.
[32] Rufiner Juan Ignacio (2020). Sistema de procesamiento y codificación de la
señal de voz en FPGA con aplicación en clasificación del habla
[33] Pan, Z., Chua, Y., Wu, J., Zhang, M., Li, H., Ambikairajah, E.: An efficient and perceptually motivated auditory neural encoding and decoding algorithm for spiking neural networks. Frontiers in Neuroscience 13 (2020), https://www.frontiersin.org/articles/10.3389/fnins.2019.01420
[34] Verstraeten, D., Schrauwen, B., Stroobandt, D., Van Campenhout, J.: Isolated word recognition with the liquid state machine: a case study. Information Processing Letters 95(6), 521–528 (2005).
[35] Deng, Y., Chakrabartty, S., Cauwenberghs, G.: Analog auditory perception model for robust speech recognition. In: Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on. vol. 3, pp. 1705–1709. IEEE (2004).

PID 6187 **Denominación del Proyecto**
Implementación en FPGA de una red neuronal pulsante para clasificación de habla y estados emotivos

Director
Iván Rodolfo Peralta

Codirector
Eduardo Filomena

Unidad de Ejecución
Universidad Nacional de Entre Ríos

Dependencia
Facultad de Ingeniería

Contacto
ivan.peralta@uner.edu.ar

Cátedra/s, área o disciplina científica
Electrónica Digital, Señales Sistemas y Modelos Biológicos, Fundamentos de Programación

Instituciones intervinientes públicas o privadas.
Laboratorio de Prototipado Electrónico y 3D (Facultad de Ingeniería UNER)

Integrantes del proyecto
Docentes: Hugo Leonardo Rufiner, Carlos Marcelo Pais, Nanci Odetti, Ivan Gareis, Marcos Formica. Estudiantes de grado: Juan I. Rufiner, Nahuel E. Ricart. Becario: Osvaldo M. Zanet

Fechas de iniciación y de finalización efectivas
02/10/2018 y 24/04/2023
Aprobación del Informe Final por Resolución C.S. N° 052/24 (27-03-2024)

<<< Volver al inicio